Solutions to Exercises

Chapter 2

2.1 Two-oracle variant of the PAC model

- Assume that $\mathcal{C}$ is efficiently PAC-learnable using $\mathcal{H}$ in the standard PAC model using algorithm $\mathcal{A}$. Consider the distribution $\mathcal{D} = \frac{1}{2}(\mathcal{D}_- + \mathcal{D}_+)$. Let $h \in \mathcal{H}$ be the hypothesis output by $\mathcal{A}$. Choose δ such that:
$$\mathbb{P}[R_{\mathcal{D}}(h) \leq \epsilon/2] \geq 1 - \delta.$$
From
$$\begin{aligned} R_{\mathcal{D}}(h) &= \underset{x\sim\mathcal{D}}{\mathbb{P}}[h(x) \neq c(x)] \\ &= \frac{1}{2}(\underset{x\sim\mathcal{D}_-}{\mathbb{P}}[h(x) \neq c(x)] + \underset{x\sim\mathcal{D}_+}{\mathbb{P}}[h(x) \neq c(x)]) \\ &= \frac{1}{2}(R_{\mathcal{D}_-}(h) + R_{\mathcal{D}_+}(h)), \end{aligned}$$
it follows that:
$$\mathbb{P}[R_{\mathcal{D}_-}(h) \leq \epsilon] \geq 1 - \delta \quad \text{and} \quad \mathbb{P}[R_{\mathcal{D}_+}(h) \leq \epsilon] \geq 1 - \delta.$$
This implies two-oracle PAC-learning with the same computational complexity.
- Assume now that $\mathcal{C}$ is efficiently PAC-learnable in the two-oracle PAC model. Thus, there exists a learning algorithm $\mathcal{A}$ such that for $c \in \mathcal{C}$, $\epsilon > 0$, and $\delta > 0$, there exist m_- and m_+ polynomial in $1/\epsilon$, $1/\delta$, and $size(c)$, such that if we draw m_- negative examples or more and m_+ positive examples or more, with confidence $1 - \delta$, the hypothesis h output by $\mathcal{A}$ verifies:
$$\mathbb{P}[R_{\mathcal{D}_-}(h)] \leq \epsilon \quad \text{and} \quad \mathbb{P}[R_{\mathcal{D}_+}(h)] \leq \epsilon.$$
Now, let $\mathcal{D}$ be a probability distribution over negative and positive examples. If we could draw m examples according to $\mathcal{D}$ such that $m \geq \max\{m_-, m_+\}$, m polynomial in $1/\epsilon$, $1/\delta$, and $size(c)$, then two-oracle PAC-learning would imply standard PAC-learning:
$$\begin{aligned} &\mathbb{P}[R_{\mathcal{D}}(h)] \\ &\leq \mathbb{P}[R_{\mathcal{D}}(h)|c(x) = 0]\,\mathbb{P}[c(x) = 0] + \mathbb{P}[R_{\mathcal{D}}(h)|c(x) = 1]\,\mathbb{P}[c(x) = 1] \\ &\leq \epsilon(\mathbb{P}[c(x) = 0] + \mathbb{P}[c(x) = 1]) = \epsilon. \end{aligned}$$
If $\mathcal{D}$ is not too biased, that is, if the probability of drawing a positive example, or that of drawing a negative example is more than ϵ, it is not hard to show, using Chernoff bounds or just Chebyshev's inequality, that drawing a polynomial number of examples in $1/\epsilon$ and $1/\delta$ suffices to guarantee that $m \geq \max\{m_-, m_+\}$ with high confidence.
Otherwise, $\mathcal{D}$ is biased toward negative (or positive examples), in which case returning $h = h_0$ (respectively $h = h_1$) guarantees that $\mathbb{P}[R_{\mathcal{D}}(h)] \leq \epsilon$.

To show the claim about the not-too-biased case, let S_m denote the number of positive examples obtained when drawing m examples when the probability of a positive example is ϵ. By Chernoff bounds,

$$\mathbb{P}[S_m \leq (1-\alpha)m\epsilon] \leq e^{-m\epsilon\alpha^2/2}.$$

We want to ensure that at least m_+ examples are found. With $\alpha = \frac{1}{2}$ and $m = \frac{2m_+}{\epsilon}$,

$$\mathbb{P}[S_m > m_+] \leq e^{-m_+/4}.$$

Setting the bound to be less than or equal to $\delta/2$, leads to the following condition on m:

$$m \geq \min\{\frac{2m_+}{\epsilon}, \frac{8}{\epsilon}\log\frac{2}{\delta}\}$$

A similar analysis can be done in the case of negative examples. Thus, when $\mathcal{D}$ is not too biased, with confidence $1-\delta$, we will find at least m_- negative and m_+ positive examples if we draw m examples, with

$$m \geq \min\{\frac{2m_+}{\epsilon}, \frac{2m_-}{\epsilon}, \frac{8}{\epsilon}\log\frac{2}{\delta}\}.$$

In both solutions, our training data is the set T and our learned concept $L(T)$ is the tightest circle (with minimal radius) which is consistent with the data.

2.2 PAC learning of hyper-rectangles

The proof in the case of hyper-rectangles is similar to the one given presented within the chapter. The algorithm selects the tightest axis-aligned hyper-rectangle containing all the sample points. For $i \in [2n]$, select a region r_i such that $\mathbb{P}_{\mathcal{D}}[r_i] = \epsilon/(2n)$ for each edge of the hyper-rectangle R. Assuming that $\mathbb{P}_{\mathcal{D}}[R - R'] > \epsilon$, argue that R' cannot meet all r_is, so it must miss at least one. The probability that none of the m sample points falls into region r_i is $(1-\epsilon/2n)^m$. By the union bound, this shows that

$$\mathbb{P}[R(R') > \epsilon] \leq 2n(1-\epsilon/2n)^m \leq 2n\exp\Big(-\frac{\epsilon m}{2n}\Big). \tag{E.35}$$

Setting δ to the right-hand side shows that for

$$m \geq \frac{2n}{\epsilon}\log\frac{2n}{\delta}, \tag{E.36}$$

with probability at least $1-\delta$, $R_{\mathcal{D}}(R') \leq \epsilon$.

2.3 Concentric circles

Suppose our target concept c is the circle around the origin with radius r. We will choose a slightly smaller radius s by

$$s := \inf\{s' \colon P(s' \leq ||x|| \leq r) < \epsilon\}.$$

Let A denote the annulus between radii s and r; that is, $A := \{x \colon s \leq ||x|| \leq r\}$. By definition of s,

$$P(A) \geq \epsilon. \tag{E.37}$$

In addition, our generalization error, $P(c \,\Delta\, L(T))$, must be small if T intersects A. We can state this as

$$P(c \,\Delta\, L(T)) > \epsilon \implies T \cap A = \emptyset. \tag{E.38}$$

Using (E.37), we know that any point in T chosen according to P will "miss" region A with probability at most $1-\epsilon$. Defining $error := P(c \,\Delta\, L(T))$, we can combine this with (E.38) to see that

$$P(error > \epsilon) \leq P(T \cap A = \emptyset) \leq (1-\epsilon)^m \leq e^{-m\epsilon}.$$

Setting δ to be greater than or equal to the right-hand side leads to $m \geq \frac{1}{\epsilon}\log(\frac{1}{\delta})$.

2.4 Non-concentric circles

As in the previous example, it is natural to assume the learning algorithm operates by returning the smallest circle which is consistent with the data. Gertrude is relying on the logical implication

$$error > \epsilon \implies T \cap r_i = \emptyset \text{ for some } i, \tag{E.39}$$

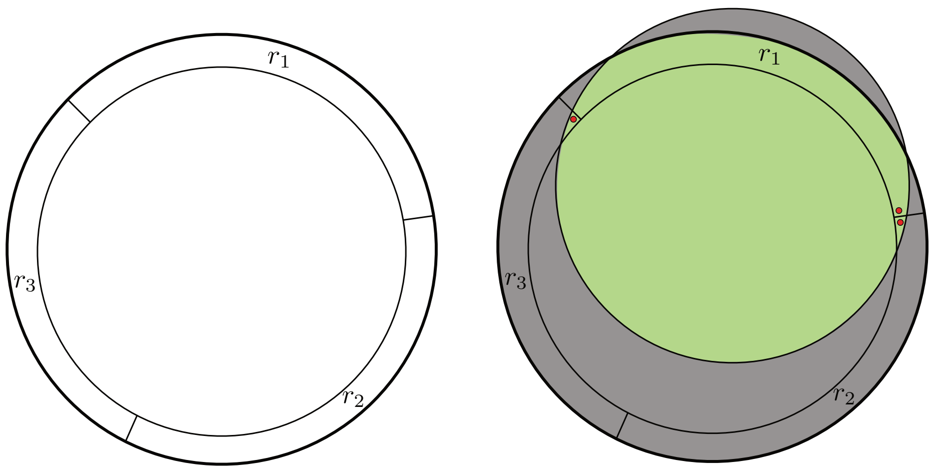

Figure E.5
Counter-example shows error of tightest circle in gray.

which is not necessarily true here. Figure E.5 illustrates a counterexample. In the figure, we have one training point in each region r_i. The points in r_1 and r_2 are very close together, and the point in r_3 is very close to region r_1. On this training data (some other points may be included outside the three regions r_i), our learned circle is the "tightest" circle including these points, and hence one diameter approximately traverses the corners of r_1. In the figure, the gray regions are the error of this learned hypotheses versus the target circle, which has a thick border. Clearly, the error may be greater than ϵ even while $T \cap r_i \neq \emptyset$ for any i; this contradicts (E.39) and invalidates poor Gertrude's proof.

2.5 Triangles

As in the case of axis-aligned rectangles, consider three regions r_1, r_2, r_3, along the sides of the target concept as indicated in figure E.6. Note that the triangle formed by the points $A", B", C"$ is similar to ABC (same angles) since $A"B"$ must be parallel to AB, and similarly for the other sides.

Assume that $\mathbb{P}[ABC] > \epsilon$, otherwise the statement would be trivial. Consider a triangle $A'B'C'$ similar to ABC and consistent with the training sample and such that it meets all three regions r_1, r_2, r_3.

Since it meets r_1, the line $A'B'$ must be below $A"B"$. Since it meets r_2 and r_3, A' must be in r_2 and B' in r_3 (see figure E.6). Now, since the angle $\widehat{A'B'C'}$ is equal to $\widehat{A"B"C"}$, C' must be necessarily above $C"$. This implies that triangle $A'B'C'$ contains $A"B"C"$, and thus $error(A'B'C') \leq \epsilon$.

$$error(A'B'C') > \epsilon \implies \exists i \in \{1,2,3\} \colon A'B'C' \cap r_i = \emptyset.$$

Thus, by the union bound,

$$\mathbb{P}[error(A'B'C') > \epsilon] \leq \sum_{i=1}^{3} \mathbb{P}[A'B'C' \cap r_i = \emptyset] \leq 3(1-\epsilon/3)^m \leq 3e^{-3m\epsilon}.$$

Setting δ to match the right-hand side gives the sample complexity $m \geq \frac{3}{\epsilon} \log \frac{3}{\delta}$.

2.8 Learning intervals

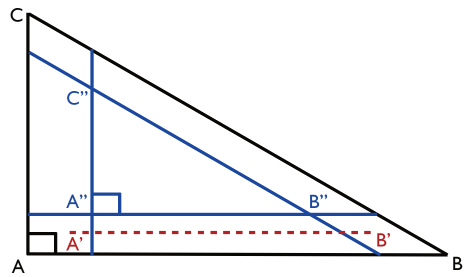

Figure E.6
Rectangle triangles.

Given a sample S, one algorithm consists of returning the tightest closed interval I_S containing positive points. Let $I = [a, b]$ be the target concept. If $\mathbb{P}[I] < \epsilon$, then clearly $R(I_S) < \epsilon$. Assume that $\mathbb{P}[I] \geq \epsilon$. Consider two intervals I_L and I_R defined as follows:

$$I_L = [a, x] \quad \text{with } x = \inf\{x \colon \mathbb{P}[a, x] \geq \epsilon/2\}$$
$$I_R = [x', b] \quad \text{with } x' = \sup\{x' \colon \mathbb{P}[x', b] \geq \epsilon/2\}.$$

By the definition of x, the probability of $[a, x[$ is less than or equal to $\epsilon/2$, similarly the probability of $]x', b]$ is less than or equal to $\epsilon/2$. Thus, if I_S overlaps both with I_L and I_R, then its error region has probability at most ϵ. Thus, $R(I_S) > \epsilon$ implies that I_S does not overlap with either I_L or I_R, that is either none of the training points falls in I_L or none falls in I_R. Thus, by the union bound,

$$\begin{aligned}\mathbb{P}[R(I_S) > \epsilon] &\leq \mathbb{P}[S \cap I_L = \emptyset] + \mathbb{P}[S \cap I_R = \emptyset] \\ &\leq 2(1 - \epsilon/2)^m \leq 2e^{-m\epsilon/2}.\end{aligned}$$

Setting δ to match the right-hand side gives the sample complexity $m = \frac{2}{\epsilon} \log \frac{2}{\delta}$ and proves the PAC-learning of closed intervals. □

2.9 Learning union of intervals

Given a sample S, our algorithm consists of the following steps:

(a) Sort S in ascending order.

(b) Loop through sorted S, marking where intervals of consecutive positively labeled points begin and end.

(c) Return the union of intervals found on the previous step. This union is represented by a list of tuples that indicate start and end points of the intervals.

This algorithms works both for $p = 2$ and for a general p. We will now consider the problem for $\mathcal{C}_2$. To show that this is a PAC-learning algorithm we need to distinguish between two cases.

The first case is when our target concept is a disjoint union of two closed intervals: $I = [a, b] \cup [c, d]$. Note, there are two sources of error: false negatives in $[a, b]$ and $[c, d]$ and also false positives in (b, c). False positives may occur if no sample is drawn from (b, c). By linearity of expectation and since these two error regions are disjoint, we have that $R(h_S) = R_{\text{FP}}(h_S) + R_{\text{FN},1}(h_S) + R_{\text{FN},2}(h_S)$, where

$$\begin{aligned}R_{\text{FP}}(h_S) &= \underset{x \sim \mathcal{D}}{\mathbb{P}}[x \in h_S, x \notin I], \\ R_{\text{FN},1}(h_S) &= \underset{x \sim \mathcal{D}}{\mathbb{P}}[x \notin h_S, x \in [a, b]], \\ R_{\text{FN},2}(h_S) &= \underset{x \sim \mathcal{D}}{\mathbb{P}}[x \notin h_S, x \in [c, d]].\end{aligned}$$

Since we need to have that at least one of $R_{\mathrm{FP}}(h_S)$, $R_{\mathrm{FN},1}(h_S)$, $R_{\mathrm{FN},2}(h_S)$ exceeds $\epsilon/3$ in order for $R(h_S) > \epsilon$, by union bound

$$\begin{aligned}\mathbb{P}(R(h_S) > \epsilon) &\leq \mathbb{P}(R_{\mathrm{FP}}(h_S) > \epsilon/3 \text{ or } R_{\mathrm{FN}(h_S),1} > \epsilon/3 \text{ or } R_{\mathrm{FN}(h_S),2} > \epsilon/3)\\ &\leq \mathbb{P}(R_{\mathrm{FP}}(h_S) > \epsilon/3) + \sum_{i=1}^{2} \mathbb{P}(R_{\mathrm{FN}(h_S),i} > \epsilon/3) \end{aligned} \tag{E.40}$$

We first bound $\mathbb{P}(R_{\mathrm{FP}}(h_S) > \epsilon/3)$. Note that if $R_{\mathrm{FP}}(h_S) > \epsilon/3$, then $\mathbb{P}((b,c)) > \epsilon/3$ and hence

$$\mathbb{P}(R_{\mathrm{FP}}(h_S) > \epsilon/3) \leq (1-\epsilon/3)^m \leq e^{-m\epsilon/3}.$$

Now we can bound $\mathbb{P}(R_{\mathrm{FN}(h_S),i} > \epsilon/3)$ by $2e^{-m\epsilon/6}$ using the same argument as in the previous question. Therefore,

$$\mathbb{P}(R(h_S) > \epsilon) \leq e^{-m\epsilon/3} + 4e^{-m\epsilon/6} \leq 5e^{-m\epsilon/6}.$$

Setting, the right-hand side to δ and solving for m yields that $m \geq \frac{6}{\epsilon}\log\frac{5}{\delta}$.

The second case that we need to consider is when $I = [a,d]$, that is, $[a,b] \cap [c,d] \neq \emptyset$. In that case, our algorithm reduces to the one from exercise 2.8 and it was already shown that only $m \geq \frac{2}{\epsilon}\log\frac{2}{\delta}$ samples is required to learn this concept. Therefore, we conclude that our algorithm is indeed a PAC-learning algorithm.

Extension of this result to the case of $\mathcal{C}_p$ is straightforward. The only difference is that in (E.40), one has two summations for $p-1$ regions of false positives and $2p$ regions of false negatives. In that case sample complexity is $m \geq \frac{2(2p-1)}{\epsilon}\log\frac{3p-1}{\delta}$.

Sorting step of our algorithm takes $O(m\log m$ time and steps (b) and (c) are linear in m, which leads to overall time complexity $O(m\log m)$.

2.10 Consistent hypotheses

Since PAC-learning with L is possible for any distribution, let $\mathcal{D}$ be the uniform distribution over $\mathcal{Z}$. Note that, in that case, the cost of an error of a hypothesis h on any point $z \in \mathcal{Z}$ is $\mathbb{P}_{\mathcal{D}}[z] = 1/m$. Thus, if $R_{\mathcal{D}}(h) < 1/m$, we must have $R_{\mathcal{D}}(h) = 0$ and h is consistent. Thus, choose $\epsilon = 1/(m+1)$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over samples S with $|S| \geq P((m+1), 1/\delta)$ points (where P is some fixed polynomial) the hypothesis h_S returned by L is consistent with Z since $R_{\mathcal{D}}(h_S) \leq 1/(m+1)$.

2.11 Senate laws

(a) The true error in the consistent case is bounded as follows:

$$R_{\mathcal{D}}(h) \leq \frac{1}{m}(\log|\mathcal{H}| + \log\frac{1}{\delta}). \tag{E.41}$$

For $\delta = .05$, $m = 200$ and $|\mathcal{H}| = 2800$, $R_{\mathcal{D}}(h) \leq 5.5\%$.

(b) The true error in the inconsistent case is bounded as:

$$R_{\mathcal{D}}(h) \leq \widehat{R}_{\mathcal{D}}(h) + \sqrt{\frac{1}{2m}(\log 2|\mathcal{H}| + \log\frac{1}{\delta})}. \tag{E.42}$$

For $\delta = .05$, $\widehat{R}_{\mathcal{D}}(h) = m'/m = .1$, $m = 200$ and $|\mathcal{H}| = 2800$, $R_{\mathcal{D}}(h) \leq 27.05\%$.

2.12 Bayesian bound. For any fixed $h \in \mathcal{H}$, by Hoeffding's inequality, for any $\delta > 0$,

$$\mathbb{P}\left[R(h) - \widehat{R}_S(h) \geq \sqrt{\frac{\log\frac{1}{p(h)\delta}}{2m}}\right] \leq p(h)\delta. \tag{E.43}$$

By the union bound,

$$\mathbb{P}\left[\exists h\colon R(h) - \widehat{R}_S(h) \geq \sqrt{\frac{\log \frac{1}{p(h)\delta}}{2m}}\right] \leq \sum_{h\in\mathcal{H}} \mathbb{P}\left[R(h) - \widehat{R}_S(h) \geq \sqrt{\frac{\log \frac{1}{p(h)\delta}}{2m}}\right]$$
$$\leq \sum_{h\in\mathcal{H}} p(h)\delta = \delta.$$

In the case of a finite hypothesis set and a uniform prior $p(h) = 1/|\mathcal{H}|$, the bound coincides with the one presented in the chapter.

2.13 Learning with an unknown parameter.

(a) By definition of acceptance,

$$\begin{aligned}
\mathbb{P}[h \text{ is accepted}] &= \mathbb{P}[\widehat{R}_S(h) \leq 3/4\epsilon] \\
&\leq \mathbb{P}[\widehat{R}_S(h) \leq 3/4\, R(h)] && (R(h) \geq \epsilon) \\
&\leq \exp\Big(-\frac{n}{2} R(h)(1/4)^2\Big) && \text{(Chernoff bound)} \\
&= \exp\Big(-\frac{R(h)}{\epsilon} \log \frac{2^{i+1}}{\delta}\Big) && \text{(def. of } n) \\
&= \exp\Big(-\log \frac{2^{i+1}}{\delta}\Big) = \frac{\delta}{2^{i+1}}. && (R(h) \geq \epsilon)
\end{aligned}$$

(b) By definition, $\mathbb{P}[h \text{ is rejected}] = \mathbb{P}[\widehat{R}_S(h) \geq \frac{3}{4}\epsilon]$. Since $R(h) \leq \epsilon/2$, $\mathbb{P}[h \text{ is rejected}] \leq \mathbb{P}[\widehat{R}_S(h) \geq \frac{3}{4}\epsilon \mid R(h) = \epsilon/2]$. By the Chernoff bounds, we can thus write

$$\begin{aligned}
\mathbb{P}[h \text{ is rejected}] &\leq \exp\Big(-\frac{n}{3}\frac{\epsilon}{2}(1/2)^2\Big) && \text{(Chernoff bound)} \\
&= \exp\Big(-\frac{4}{3} \log \frac{2^{i+1}}{\delta}\Big) && \text{(def. of } n) \\
&\leq \exp\Big(-\log \frac{2^{i+1}}{\delta}\Big) = \frac{\delta}{2^{i+1}}.
\end{aligned}$$

(c) The estimate $\widetilde{s}$ is then an upper bound on s and thus, by definition of algorithm $\mathcal{B}$, $\mathbb{P}[R(h_i) \leq \epsilon/2] \geq 1/2$. If a hypothesis h has error at least $\epsilon/2$ it is rejected with probability at most $\delta/2^{i+1} \leq \delta/4 \leq 1/4$, therefore, it is accepted with probability at 3/4. Thus, for $\widetilde{s} \geq s$, $\mathbb{P}[h_i \text{ is accepted}] \geq 1/2 \times 1/4 = 3/8$.

(d) By the previous question, the probability that algorithm $\mathcal{B}$ fails to halt while $\widetilde{s} \geq s$ is at most $1 - 3/8 = 5/8$. Thus, the probability that it does not halt after j iterations is at most $(5/8)^j \leq (5/8)^{\log \frac{2}{\delta} / \log \frac{8}{5}} = \exp\Big(\log \frac{2}{\delta} / \log \frac{8}{5} \log \frac{5}{8}\Big) = \delta/2$.

(e) By definition,

$$\begin{aligned}
\widetilde{s} \geq s &\iff \lfloor 2^{(i-1)/\log \frac{2}{\delta}} \rfloor \geq s \\
&\iff 2^{(i-1)/\log \frac{2}{\delta}} \geq s \\
&\iff \frac{i-1}{\log \frac{2}{\delta}} \geq \log_2 s \\
&\iff i \geq 1 + (\log_2 s) \log \frac{2}{\delta} \\
&\iff i \geq \lceil 1 + (\log_2 s) \log \frac{2}{\delta} \rceil.
\end{aligned}$$

(f) In view of the two previous questions, with probability at least $1 - \delta/2$, algorithm $\mathcal{B}$ halts after at most j' iterations. The probability that the hypothesis it returns be accepted while its error is greater than ϵ is at most $\delta/2^{j'+1} \leq \delta/2$. Thus, with probability $1 - \delta$, the algorithm halts and the hypothesis it returns has error at most ϵ. □

Since concepts can be identified with indicator functions, the intersection of two concepts can be identified with the product two such indicator functions. In view of that, by the result just proven and after taking expectations, the following holds:

$$\mathfrak{R}_m(\mathcal{C}) \leq \mathfrak{R}_m(\mathcal{C}_1) + \mathfrak{R}_m(\mathcal{C}_2).$$

3.10 Rademacher complexity of prediction vector

(a) The following proves the result:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{H}) &= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{f \in \{h, -h\}} \sum_{i=1}^m \sigma_i f(x_i) \right] \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} [\max\{\boldsymbol{\sigma} \cdot \mathbf{u}, -\boldsymbol{\sigma} \cdot \mathbf{u}\}] \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} [|\boldsymbol{\sigma} \cdot \mathbf{u}|] \\
&\leq \frac{1}{m} \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} [|\boldsymbol{\sigma} \cdot \mathbf{u}|^2]} && \text{(by Jensen's inequality)} \\
&= \frac{1}{m} \sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sum_{i,j=1}^m \sigma_i \sigma_j u_i u_j \right]} \\
&= \frac{1}{m} \sqrt{\sum_{i,j=1}^m \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} [\sigma_i \sigma_j] u_i u_j} \\
&= \frac{1}{m} \sqrt{\sum_{i=1}^m u_i^2} && (\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[\sigma_i \sigma_j] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[\sigma_i] \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[\sigma_j] = 0 \text{ for } i \neq j) \\
&= \frac{\|\mathbf{u}\|}{m}.
\end{aligned}$$

Thus, $\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \frac{\sqrt{n}}{m}$. For $n = 1$, $\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \frac{1}{m}$ while for $n = m$, $\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \frac{1}{\sqrt{m}}$.

(b) The empirical Rademacher complexity of $\mathcal{F} + h$ can be expressed as follows:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{F} + h) &= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(x_i) + \sigma_i h(x_i) \right] \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(x_i) \right] + \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sum_{i=1}^m \sigma_i h(x_i) \right] \\
&= \widehat{\mathfrak{R}}_S(\mathcal{F}) + \frac{1}{m} \sum_{i=1}^m \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[\sigma_i] h(x_i) = \widehat{\mathfrak{R}}_S(\mathcal{F}).
\end{aligned}$$

The empirical Rademacher complexity of $\mathcal{F} \pm h$ can be upper bounded as follows using the first question:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{F} + h) &= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(x_i) + \sup_{s \in \{-1, +1\}} s \sigma_i h(x_i) \right] \\
&= \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^m \sigma_i f(x_i) \right] + \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{s \in \{-1, +1\}} s \sum_{i=1}^m \sigma_i h(x_i) \right] \\
&\leq \widehat{\mathfrak{R}}_S(\mathcal{F}) + \frac{\|\mathbf{u}\|}{m}.
\end{aligned}$$

3.11 Rademacher complexity of regularized neural networks

(a)

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{H}) &= \frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_1\leq\Lambda', \|\mathbf{u}_j\|_2\leq\Lambda}\sum_{i=1}^m \sigma_i \sum_{j=1}^{n_2} w_j \sigma(\mathbf{u}_j\cdot\mathbf{x}_i)\right]\\
&= \frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{w}\|_1\leq\Lambda', \|\mathbf{u}_j\|_2\leq\Lambda}\sum_{j=1}^{n_2} w_j \sum_{i=1}^m \sigma_i \sigma(\mathbf{u}_j\cdot\mathbf{x}_i)\right]\\
&= \frac{\Lambda'}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{u}_j\|_2\leq\Lambda}\max_{j\in[n_2]}\left|\sum_{i=1}^m \sigma_i \sigma(\mathbf{u}_j\cdot\mathbf{x}_i)\right|\right] && \text{(all the weight put on largest term)}\\
&= \frac{\Lambda'}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{u}_j\|_2\leq\Lambda, j\in[n_2]}\left|\sum_{i=1}^m \sigma_i \sigma(\mathbf{u}_j\cdot\mathbf{x}_i)\right|\right]\\
&= \frac{\Lambda'}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{u}\|_2\leq\Lambda}\left|\sum_{i=1}^m \sigma_i \sigma(\mathbf{u}\cdot\mathbf{x}_i)\right|\right].
\end{aligned}$$

(b) By Talagrand's lemma, since σ is L-Lipschitz, the following inequality holds:

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{H}) &\leq \frac{\Lambda' L}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\left|\sum_{i=1}^m \sigma_i \mathbf{u}\cdot\mathbf{x}_i\right|\right]\\
&= \frac{\Lambda' L}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\sup_{s\in\{-1,+1\}} s\sum_{i=1}^m \sigma_i \mathbf{u}\cdot\mathbf{x}_i\right] && \text{(def. of abs. value)}\\
&= \Lambda' L\,\widehat{\mathfrak{R}}_S(\mathcal{H}').
\end{aligned}$$

(c)

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{H}') &= \frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{u}\|_2\leq\Lambda, s\in\{-1,+1\}}\sum_{i=1}^m \sigma_i s\mathbf{u}\cdot\mathbf{x}_i\right]\\
&= \frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{u}\|_2\leq\Lambda}\left|\sum_{i=1}^m \sigma_i \mathbf{u}\cdot\mathbf{x}_i\right|\right] && \text{(def. of abs. val.)}\\
&= \frac{1}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{\|\mathbf{u}\|_2\leq\Lambda}\left|\mathbf{u}\cdot\sum_{i=1}^m \sigma_i \mathbf{x}_i\right|\right]\\
&= \frac{\Lambda}{m}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^m \sigma_i \mathbf{x}_i\right\|_2\right] && \text{(Cauchy-Schwarz eq. case).}
\end{aligned}$$

The last equality holds by setting $\mathbf{u} = \frac{\Lambda\sum_{i=1}^m \sigma_i\mathbf{x}_i}{\|\sum_{i1=1}^m \sigma_i\mathbf{x}_i\|}$.

(d)

$$\begin{aligned}
\widehat{\mathfrak{R}}_S(\mathcal{H}') &= \frac{\Lambda}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{m} \sigma_i \mathbf{x}_i\right\|_2\right] \\
&\leq \frac{\Lambda}{m}\sqrt{\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{m} \sigma_i \mathbf{x}_i\right\|_2^2\right]} && \text{(Jensen's ineq.)} \\
&= \frac{1}{m}\sqrt{\sum_{i,j=1}^{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[\sigma_i\sigma_j]\,(\mathbf{x}_i \cdot \mathbf{x}_j)} \\
&= \frac{\Lambda}{m}\sqrt{\sum_{i,j=1}^{m} 1_{i=j}(\mathbf{x}_i \cdot \mathbf{x}_j)} && \text{(independence of } \sigma_i\text{s)} \\
&= \frac{\Lambda}{m}\sqrt{\sum_{i=1}^{m} \|\mathbf{x}_i\|_2^2}.
\end{aligned}$$

(e) In view of the previous questions,

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \leq \Lambda' L\, \widehat{\mathfrak{R}}_S(\mathcal{H}') \leq \frac{\Lambda'\Lambda L}{m}\sqrt{\sum_{i=1}^{m} \|\mathbf{x}_i\|_2^2} \leq \frac{\Lambda'\Lambda L}{m}\sqrt{mr^2} = \frac{\Lambda'\Lambda L r}{\sqrt{m}}.$$

3.12 Rademacher complexity

Consider the simple case where $\mathcal{H}$ is reduced to the constant hypothesis $h_1 \colon x \mapsto 1$ and $h_{-1} \colon x \mapsto -1$. Then, by definition of the empirical Rademacher complexity,

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) = \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[\max\{\sum_{i=1}^{m} \sigma_i, \sum_{i=1}^{m} -\sigma_i\}] = \frac{1}{m} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}[|\sum_{i=1}^{m} \sigma_i|]$$

Let $X = \sum_{i=1}^{m} \sigma_i$. Note that $\mathbb{E}[X^2] = \mathbb{E}[\sum_{i,j=1}^{m} \sigma_i\sigma_j]$. For any $i \neq j$, since σ_i and σ_j are independent, $\mathbb{E}[\sigma_i\sigma_j] = \mathbb{E}[\sigma_i]\,\mathbb{E}[\sigma_j] = 0$. Thus,

$$\mathbb{E}[X^2] = \sum_{i=1}^{m} \mathbb{E}[\sigma_i\sigma_i] = \sum_{i=1}^{m} \mathbb{E}[\sigma_i^2] = m.$$

Now, by Hölder's inequality,

$$m = \mathbb{E}[X^2] = \mathbb{E}[|X|^{2/3}|X|^{4/3}] \leq \mathbb{E}[|X|]^{2/3}\, \mathbb{E}[X^4]^{1/3}.$$

Thus,

$$\begin{aligned}
\mathbb{E}[|X|] &\geq \frac{m^{3/2}}{\mathbb{E}[X^4]^{1/2}} = \frac{m^{3/2}}{\sqrt{\mathbb{E}[\sum_{i=1}^{m} \sigma_i^4 + 3\sum_{i\neq j} \sigma_i^2\sigma_j^2]}} = \frac{m^{3/2}}{\sqrt{m + 3m(m-1)}} \\
&= \frac{m^{3/2}}{\sqrt{m(3m-2)}} \geq \frac{m^{3/2}}{\sqrt{m(3m)}} = \frac{\sqrt{m}}{\sqrt{3}}.
\end{aligned}$$

This shows that

$$\widehat{\mathfrak{R}}_S(\mathcal{H}) \geq \frac{\sqrt{m}}{\sqrt{3}}.$$

Since $\mathfrak{R}_m(\mathcal{H}) \geq \widehat{\mathfrak{R}}_S(\mathcal{H}) + O(\frac{1}{\sqrt{m}})$, it implies $\mathfrak{R}_m(\mathcal{H}) \geq O(\frac{1}{\sqrt{m}})$, which contradicts $\mathfrak{R}_m(\mathcal{H}) \leq O(\frac{1}{m})$.

Note that for the lower bound, we could have used instead a more general result (Khinchine's inequality) which states that for any $\mathbf{a} \in \mathbb{R}^m$,

$$\mathbb{E}[|\ \boldsymbol{\sigma} \cdot \mathbf{a}|] \geq \frac{\|\mathbf{a}\|_2}{\sqrt{2}}.$$

3.13 VC-dimension of union of k intervals

The VC-dimension of this class is $2k$. It is not hard to see that any $2k$ distinct points on the real line can be shattered using k intervals; it suffices to shatter each of the k pairs of consecutive points with an interval. Assume now that $2k+1$ distinct points $x_1 < \cdots < x_{2k+1}$ are given. For any $i \in [2k+1]$, label x_i with $(-1)^{i+1}$, that is alternatively label points with 1 or -1. This leads to $k+1$ points labeled positively and requires $2k+1$ intervals to shatter the set, since no interval can contain two consecutive points. Thus, no set of $2k+1$ points can be shattered by k intervals, and the VC-dimension of the union of k intervals is $2k$.

3.14 VC-dimension of finite hypothesis sets

With a finite set $\mathcal{H}$, at most $2^{|\mathcal{H}|}$ dichotomies can be defined.

3.15 VC-dimension of subsets

The set of three points $\{0, 3/4, 3/2\}$ can be fully shattered as follows:

$$\begin{array}{ll}
+++ & \alpha = -2 \\
++- & \alpha = 0 \\
+-+ & \alpha = -1 \\
+-- & \alpha = 3/2 - 2 + \epsilon \\
-++ & \alpha = 3/4 - 2 \\
-+- & \alpha = \epsilon \\
--+ & \alpha = 3/2 \\
--- & \alpha = 3/2 + \epsilon,
\end{array}$$

where e is a small number, e.g., $\epsilon = .1$. No set of four points $x_1 < x_2 < x_3 < x_4$ can be labeled by $+-+-$. This is because the three leftmost labels $+-+$ imply that $\alpha + 2 \leq x_3$ and thus also $\alpha + 2 < x_4$. Thus, the VC-dimension of the set of subsets I_α is 3. Note that this does not coincide with the number of parameters used to describe the class.

3.16 VC-dimension of axis-aligned squares and triangles

(a) It is not hard to see that the set of 3 points with coordinates $(1, 0)$, $(0, 1)$, and $(-1, 0)$ can shattered by axis-aligned squares: e.g., to label positively two of these points, use a square defined by the axes and with those to points as corners. Thus, the VC-dimension is at least 3. No set of 4 points can be fully shattered. To see this, let P_T be the highest point, P_B the lowest, P_L the leftmost, and P_R the rightmost, assuming for now that these can be defined in a unique way (no tie) – the cases where there are ties can be treated in a simpler fashion. Assume without loss of generality that the difference d_{BT} of y-coordinates between P_T and P_B is greater than the difference d_{LR} of x-coordinates between P_L and P_R. Then, P_T and P_B cannot be labeled positively while P_L and P_R are labeled negatively. Thus, the VC-dimension of axis-aligned squares in the plane is 3.

(b) Check that the set of 4 points with coordinates $(1, 0)$, $(0, 1)$, $(-1, 0)$, and $(0, -1)$ can be shattered by such triangles. This is essentially the same as the case with axis aligned rectangles. To see that no five points can be shattered, the same example or argument as for axis-aligned rectangles can be used: labeling all points positively except from the one within the interior of the convex hull is not possible (for the degnerate cases where no point is in the interior of the convex hull is simpler, this is even easier to see). Thus, the VC-dimension of this family of triangles is 4.

3.17 VC-dimension of closed balls in $\mathbb{R}^n$.

Let $B(a, r)$ be the ball of radius r centered at $a \in \mathbb{R}^n$. Then $x \in B(a, r)$ iff

$$\sum_{i=1}^{n} \|x_i\|^2 - 2\sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} a_i^2 - r \leq 0, \tag{E.47}$$

which is equivalent to

$$\langle W, X\rangle + B \leq 0, \tag{E.48}$$

with $W = \begin{bmatrix} 1 \\ -2a_1 \\ \cdots \\ -2a_n \end{bmatrix}$, $X = \begin{bmatrix} \sum_{i=1}^n \|x_i\|^2 \\ x_1 \\ \cdots \\ x_n \end{bmatrix}$, and $B = \sum_{i=1}^n a_i^2 - r$. The VC-dimension of closed balls in R^n is thus at most equal to the VC-dimension of hyperplanes in $\mathbb{R}^{n+1}$, that is, $n+2$.

3.18 VC-dimension of ellipsoids

The general equation of ellipsoids in $\mathbb{R}^n$ is

$$(\mathbf{X} - \mathbf{X}_0)^\top \mathbf{A}(\mathbf{X} - \mathbf{X}_0) \leq 1, \tag{E.49}$$

where $\mathbf{X}, \mathbf{X}_0 \in \mathbb{R}^n$ and $\mathbf{A} = (a_{ij}) \in \mathbb{S}_+^n$ is a positive semidefinite symmetric matrix. This can be rewritten as

$$\mathbf{X}^\top \mathbf{A}\mathbf{X} - 2\mathbf{X}^\top \mathbf{A}\mathbf{X}_0 + \mathbf{X}_0^\top \mathbf{A}\mathbf{X}_0 \leq 1, \tag{E.50}$$

or $\sum_{i,j=1}^n a_{ij}(x_i x_j + x_j x_i) - \sum_{i=1}^n 2(\mathbf{A}\mathbf{X}_0)_i\, x_i + \big(\mathbf{X}_0^\top \mathbf{A}\mathbf{X}_0 - 1\big) \leq 0$ using the fact that $\mathbf{A}$ is symmetric. Let $a_i = -2(\mathbf{A}\mathbf{X}_0)_i$ for $i \in [n]$ and let $b = \mathbf{X}_0^\top \mathbf{A}\mathbf{X}_0 - 1$. Then this can be viewed in terms of the following equations of hyperplanes in $\mathbb{R}^{n(n+1)/2+n}$

$$\mathbf{W}^\top \mathbf{Z} + b \leq 0, \tag{E.51}$$

with

$$\mathbf{W} = \begin{bmatrix} a_1 \\ \cdots \\ a_n \\ a_{11} \\ \cdots \\ a_{ij} \\ \cdots \\ a_{nn} \end{bmatrix} \quad \mathbf{Z} = \left.\begin{bmatrix} x_1 \\ \cdots \\ x_n \\ x_1x_1 + x_1x_1 \\ \cdots \\ x_ix_j + x_jx_i \\ \cdots \\ x_nx_n + x_nx_n. \end{bmatrix}\right\} n(n+1)/2 + n \tag{E.52}$$

Since the VC-dimension of hyperplanes in $\mathbb{R}^{n(n+1)/2+n}$ is $n(n+1)/2+n+1 = (n+1)(n/2+1)$, the VC-dimension of ellipsoids in $\mathbb{R}^n$ is bounded by $(n+1)(n+2)/2$.

3.19 VC-dimension of a vector space of real functions

Show that no set of size $m = r+1$ can be shattered by $\mathcal{H}$. Let $x_1, \ldots, x_m$ be m arbitrary points. Define the linear mapping $l\colon F \to \mathbb{R}^m$ defined by:

$$l(f) = (f(x_1), \ldots, f(x_m))$$

Since the dimension of $dim(F) = m-1$, the rank of l is at most $m-1$, and there exists $\alpha \in \mathbb{R}^m$ orthogonal to $l(F)$:

$$\forall f \in F, \sum_{i=1}^m \alpha_i f(x_i) = 0$$

We can assume that at least one α_i is negative. Then,

$$\forall f \in F, \sum_{i\colon \alpha_i \geq 0}^m \alpha_i f(x_i) = -\sum_{i\colon \alpha_i < 0}^m \alpha_i f(x_i)$$

Now, assume that there exists a set $\{x\colon f(x) \geq 0\}$ selecting exactly the x_is on the left-hand side. Then all the terms on the left-hand side are non-negative, while those on the right-hand side are negative, which cannot be. Thus, $\{x_1, \ldots, x_m\}$ cannot be shattered.

3.20 VC-dimension of sine functions

(a) Fix $x \in \mathbb{R}$ and suppose there exists an ω that realizes the labeling $--+-$. Thus $\sin(\omega x) < 0$, $\sin(2\omega x) < 0$, $\sin(3\omega x) \geq 0$ and $\sin(4\omega x) < 0$. We will show that this implies $\sin^2(\omega x) < \frac{1}{2}$ and $\sin^2(\omega x) \geq \frac{3}{4}$, a contradiction.

Using the identity $\sin(2\theta) = 2\sin(\theta)\cos(\theta)$ and the fact that $\sin(4\omega x) < 0$ we have

$$2\sin(2\omega x)\cos(2\omega x) = \sin(4\omega x) < 0.$$

Since $\sin(2\omega x) < 0$ we can divide both sides of this inequality by $2\sin(2\omega x)$ to conclude $\cos(2\omega x) > 0$. Applying the identity $\cos(2\theta) = 1 - 2\sin^2(\theta)$ yields $1 - 2\sin^2(\omega x) > 0$, or $\sin^2(\omega x) < \frac{1}{2}$.

Using the identity $\sin(3\theta) = 3\sin(\theta) - 4\sin^3(\theta)$ and the fact that $\sin(3\omega x) \geq 0$ we have

$$3\sin(\omega x) - 4\sin^3(\omega x) = \sin(3\omega x) \geq 0$$

Since $\sin(\omega x) < 0$ we can divide both sides of this inequality by $\sin(\omega x)$ to conclude $3 - 4\sin^2(\omega x) \leq 0$, or $\sin^2(\omega x) \geq \frac{3}{4}$.

(b) For any $m > 0$, consider the set of points $(x_1, \ldots, x_m)$ with arbitrary labels $(y_1, \ldots, y_m) \in \{-1, +1\}^m$. Now, choose the parameter $\omega = \pi(1 + \sum_{i=1}^m 2^i y'_i)$ where $y'_i = \frac{1-y_i}{2}$. We show that this single parameter will always correctly classify the entire sample for any $m > 0$ and choice of labels. For any $j \in [m]$ we have,

$$\omega x_j = \omega 2^{-j} = \pi(2^{-j} + \sum_{i=1}^{m} 2^{i-j} y'_i) = \pi(2^{-j} + (\sum_{i=1}^{j-1} 2^{i-j} y'_i) + y'_j + (\sum_{i=1}^{m-j} 2^i y'_i)) \,.$$

The last term can be dropped from the sum, since it contributes only multiples of 2π. Since $y'_i \in \{0, 1\}$ the remaining term $\pi(2^{-j} + (\sum_{i=1}^{j-1} 2^{i-j} y'_i) + y'_j) = \pi(\sum_{i=1}^{j-1} 2^{-i} y'_i + 2^{-j} + y'_j)$ can be upper and lower bounded as follows:

$$\pi(\sum_{i=1}^{j-1} 2^{-i} y'_i + 2^{-j} + y'_j) \leq \pi(\sum_{i=1}^{j} 2^{-i} + y'_j) < \pi(1 + y'_j) \,,$$

$$\pi(\sum_{i=1}^{j-1} 2^{-i} y'_i + 2^{-j} + y'_j) > \pi y'_j \,.$$

Thus, if $y_j = 1$ we have $y'_j = 0$ and $0 < \omega \mathbf{x}_j < \pi$, which implies $\text{sgn}(\omega x_j) = 1$. Similarly, for $y_j = -1$ we have $\text{sgn}(\omega x_j) = -1$.

3.21 VC-dimension of union of halfspaces

3.22 VC-dimension of intersection of halfspaces

Let $m \geq 0$. Note the general fact that for any concept class $\mathcal{C} = \{c_1 \cap c_2 : c_1 \in \mathcal{C}_1, c_2 \in \mathcal{C}_2\}$,

$$\Pi_{\mathcal{C}}(m) \leq \Pi_{\mathcal{C}_1}(m)\, \Pi_{\mathcal{C}_2}(m). \tag{E.53}$$

Indeed, fix a set $\mathcal{X}$ of m points. Let $\mathcal{Y}_1, \ldots, \mathcal{Y}_k$ be the traces of $\mathcal{C}_1$ on $\mathcal{X}$. By definition of $\Pi_{\mathcal{C}_1}(\mathcal{X})$, $k \leq \Pi_{\mathcal{C}_1}(\mathcal{X}) \leq \Pi_{\mathcal{C}_1}(m)$. By definition of $\Pi_{\mathcal{C}_2}(\mathcal{Y}_i)$, the traces of $\mathcal{C}_2$ on a subset $\mathcal{Y}_i$ are at most $\Pi_{\mathcal{C}_2}(\mathcal{Y}_i) \leq \Pi_{\mathcal{C}_2}(m)$. Thus, the traces of $\mathcal{C}$ on $\mathcal{X}$ are at most

$$k\Pi_{\mathcal{C}_2}(\mathcal{Y}_i) \leq \Pi_{\mathcal{C}_1}(m)\, \Pi_{\mathcal{C}_2}(m). \tag{E.54}$$

For the particular case of $\mathcal{C}_k$, using Sauer's lemma, this implies that

$$\Pi_{\mathcal{C}_k}(m) \leq (\Pi_{\mathcal{C}_1}(m))^k \leq \left(\frac{em}{n+1}\right)^{k(n+1)}. \tag{E.55}$$

If $(em/(n+1))^{k(n+1)} < 2^m$, then the VC-dimension of $\mathcal{C}_k$ is less than m. If the VC-dimension of $\mathcal{C}_k$ is m, then $\Pi_{\mathcal{C}_k}(m) = 2^m \leq (em/(n+1))^{k(n+1)}$. These inequalities give an upper bound and a lower bound on $\text{VCdim}(\mathcal{C}_k)$. As an example, using the inequality: $\forall x \in \mathbb{N} - \{3\}, \log_2(x) \leq x/2$, one can verify that:

$$\text{VCdim}(\mathcal{C}_k) \leq 2(n+1)k\log(3k). \tag{E.56}$$

3.23 VC-dimension of intersection concepts

(a) Fix a set $\mathcal{X}$ of m points. Let $\mathcal{Y}_1, \ldots, \mathcal{Y}_k$ be the set of intersections of the concepts of $\mathcal{C}_1$ with $\mathcal{X}$. By definition of $\Pi_{\mathcal{C}_1}(\mathcal{X})$, $k \leq \Pi_{\mathcal{C}_1}(\mathcal{X}) \leq \Pi_{\mathcal{C}_1}(m)$. By definition of $\Pi_{\mathcal{C}_2}(\mathcal{Y}_i)$, the intersection of the concepts of $\mathcal{C}_2$ with $\mathcal{Y}_i$ are at most $\Pi_{\mathcal{C}_2}(\mathcal{Y}_i) \leq \Pi_{\mathcal{C}_2}(m)$. Thus, the number of sets intersections of concepts of $\mathcal{C}$ with $\mathcal{X}$ is at most

$$k\Pi_{\mathcal{C}_2}(\mathcal{Y}_i) \leq \Pi_{\mathcal{C}_1}(m)\,\Pi_{\mathcal{C}_2}(m). \tag{E.57}$$

(b) In view of the result proved in the previous question, $\Pi_{\mathcal{C}_s}(m) \leq (\Pi_{\mathcal{C}_1}(m))^s$. By Sauer's lemma, this implies

$$\Pi_{\mathcal{C}_s}(m) \leq \Big(\frac{em}{d}\Big)^{sd}. \tag{E.58}$$

If $\big(\frac{em}{d}\big)^{sd} < 2^m$, then the VC-dimension of $\mathcal{C}_s$ is less than m. Thus, it suffices to show this inequality holds with $m = 2ds\log_2(3s)$. Plugging in that value for m and taking the $\log_2$ yield:

$$\begin{aligned}
ds\log_2\left(2es\log_2(3s)\right) &< 2ds\log_2(3s) && \text{(E.59)}\\
\Leftrightarrow \log_2\left(2es\log_2(3s)\right) &< 2\log_2(3s) = \log_2(9s^2) && \text{(E.60)}\\
\Leftrightarrow 2es\log_2(3s) &< 9s^2 && \text{(E.61)}\\
\Leftrightarrow \log_2(3s) &< \frac{9s}{2e}. && \text{(E.62)}
\end{aligned}$$

This last inequality holds for $s = 2$: $\log_2(6) \approx 2.6 < 9/(2e) \approx 3.3$. Since the functions corresponding to the left-hand-side grows more slowly than the one corresponding to the right-hand-side (compare derivatives for example), this implies that the inequality holds for all $s \geq 2$.

3.24 VC-dimension of union of concepts

(a) When $\mathcal{C} = \mathcal{A} \cup \mathcal{B}$, $\Pi_{\mathcal{C}}(\mathcal{X}) \leq \Pi_{\mathcal{A}}(\mathcal{X}) + \Pi_{\mathcal{B}}(\mathcal{X})$ for any set $\mathcal{X}$, since dichotomies in $\Pi_{\mathcal{C}}(\mathcal{X})$ can be generated by $\mathcal{A}$ or by $\mathcal{B}$. Thus, for all m, $\Pi_{\mathcal{C}}(m) \leq \Pi_{\mathcal{A}}(m) + \Pi_{\mathcal{B}}(m)$.

(b) For $m \geq d_{\mathcal{A}} + d_{\mathcal{B}} + 2$, by Sauer's lemma,

$$\begin{aligned}
\Pi_{\mathcal{C}}(m) &\leq \sum_{i=0}^{d_{\mathcal{A}}}\binom{m}{i} + \sum_{i=0}^{d_{\mathcal{B}}}\binom{m}{i} = \sum_{i=0}^{d_{\mathcal{A}}}\binom{m}{i} + \sum_{i=0}^{d_{\mathcal{B}}}\binom{m-i}{i}\\
&= \sum_{i=0}^{d_{\mathcal{A}}}\binom{m}{i} + \sum_{i=m-d_{\mathcal{B}}}^{d_{\mathcal{B}}}\binom{m}{i} && \text{(E.63)}\\
&\leq \sum_{i=0}^{d_{\mathcal{A}}}\binom{m}{i} + \sum_{i=d_{\mathcal{A}}+2}^{d_{\mathcal{B}}}\binom{m}{i} && \text{(E.64)}\\
&< \sum_{i=0}^{m}\binom{m}{i} = 2^m. && \text{(E.65)}
\end{aligned}$$

Thus, the VC-dimension of $\mathcal{C}$ is strictly less than $d_{\mathcal{A}} + d_{\mathcal{B}} + 2$:

$$\text{VCdim}(\mathcal{C}) \leq d_{\mathcal{A}} + d_{\mathcal{B}} + 1. \tag{E.66}$$

3.25 VC-dimension of symmetric difference of concepts

Fix a set $\mathcal{S}$. We can show that the number of classifications of $\mathcal{S}$ using $\mathcal{H}$ is the same as when using $\mathcal{H}\Delta A$. The set of classifications obtained using $\mathcal{H}$ can be identified with $\{\mathcal{S} \cap h \colon h \in \mathcal{H}\}$ and the set of classifications using $\mathcal{H}\Delta A$ can be identified with $\{\mathcal{S} \cap (h\Delta A) \colon h \in \mathcal{H}\}$. Observe that for any $h \in \mathcal{H}$,

$$\mathcal{S} \cap (h\Delta A) = (\mathcal{S} \cap h)\Delta(\mathcal{S} \cap A). \tag{E.67}$$

Figure E.7 helps illustrate this equality in a special case. Now, in view of this inequality, if $\mathcal{S} \cap (h\Delta A) = \mathcal{S} \cap (h'\Delta A)$ for $h, h' \in \mathcal{H}$, then

$$(\mathcal{S} \cap h)\Delta\mathcal{B} = (\mathcal{S} \cap h')\Delta\mathcal{B}, \tag{E.68}$$

with $\mathcal{B} = \mathcal{S} \cap \mathcal{A}$. Since two sets that have the same symmetric differences with respect to a set $\mathcal{B}$ must be equal, this implies

$$\mathcal{S} \cap h = \mathcal{S} \cap h'. \tag{E.69}$$

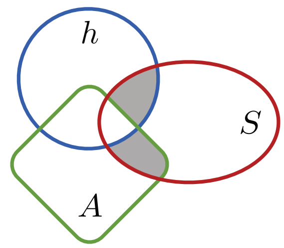

Figure E.7
Illustration of $(h\Delta\mathcal{A}) \cap \mathcal{S} = (h \cap \mathcal{S})\Delta(\mathcal{A} \cap \mathcal{S})$ shown in gray.

This shows that ϕ defined by

$$\phi\colon \mathcal{S} \cap \mathcal{H} \to \mathcal{S} \cap (\mathcal{H}\Delta\mathcal{A})$$
$$\mathcal{S} \cap h \mapsto \mathcal{S} \cap (h\Delta\mathcal{A})$$

is a bijection, and thus that the sets $\mathcal{S} \cap \mathcal{H}$ and $\mathcal{S} \cap (\mathcal{H}\Delta A)$ have the same cardinality.

3.26 Symmetric functions

(a) For $i = 0, \ldots, n$, let $x_i \in \{0,1\}^n$ be defined by $x_i = (\underbrace{1, \ldots, 1}_{i \text{ 1's}}, 0, \ldots, 0)$. Then, $\{x_0, \ldots, x_n\}$ can be shattered by $\mathcal{C}$. Indeed, let $y_0, \ldots, y_n \in 0, 1$ be an arbitrary labeling of these points. Then, the function h defined by:

$$h(x) = y_i \tag{E.70}$$

for all x with i 1's is symmetric and $h(x_i) = y_i$. Thus, $\text{VCdim}(\mathcal{C}) \geq n + 1$. Conversely, a set of $n + 2$ points cannot be shattered by $\mathcal{C}$, since at least two points would then have the same number of 1's and will not be distinguishable by $\mathcal{C}$. Thus,

$$\text{VCdim}(\mathcal{C}) = n + 1. \tag{E.71}$$

(b) Thus, in view of the theorems presented in class, a lower bound on the number of training examples needed to learn symmetric functions with accuracy $1 - \epsilon$ and confidence $1 - \delta$ is

$$\Omega(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{n}{\epsilon}), \tag{E.72}$$

and an upper bound is:

$$O(\frac{1}{\epsilon} \log \frac{1}{\delta} + \frac{n}{\epsilon} \log \frac{1}{\epsilon}), \tag{E.73}$$

which is only within a factor $\frac{1}{\epsilon}$ of the lower bound.

(c) For a training data $(z_0, t_0), \ldots, (z_m, t_m) \in \{0,1\}^n \times \{0,1\}$ define h as the symmetric function such that $h(z_i) = t_i$ for all $i = 0, \ldots, m$.

3.27 VC-dimension of neural networks

(a) Let $\Pi_u(m)$ denote the growth function at a node u in the intermediate layer. For a fixed set of values at the intermediate layer, using the concept class $\mathcal{C}$ the output node can generate at most $\Pi_{\mathcal{C}}(m)$ distinct labelings. There are $\prod_u \Pi_u(m)$ possible sets of values at the intermediate layer since, by definition, for a sample of size m, at most $\Pi_u(m)$ distinct values are possible at each u. Thus, at most $\Pi_{\mathcal{C}}(m) \times \prod_u \Pi_u(m)$ labelings can be generated by the neural network and $\Pi_{\mathcal{H}}(m) \leq \Pi_{\mathcal{C}}(m) \prod_u \Pi_u(m)$.

(b) For any intermediate node u, $\Pi_u(m) = \Pi_{\mathcal{C}}(m)$. Thus, for $\tilde{k} = k + 1$, $\Pi_{\mathcal{H}}(m) \leq \Pi_{\mathcal{C}}(m)^{\tilde{k}}$. By Sauer's lemma, $\Pi_{\mathcal{C}}(m) \leq \left(\frac{em}{d}\right)^d$, thus $\Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^{d\tilde{k}}$. Let $m = 2\tilde{k}d \log_2(e\tilde{k})$. In view of the inequality given by the hint and $e\tilde{k} > 4$, this implies $m > d\tilde{k}\log_2\left(\frac{em}{d}\right)$, that is $2^m > \left(\frac{em}{d}\right)^{d\tilde{k}}$. Thus, the VC-dimension of $\mathcal{H}$ is less than

$$2\tilde{k}d\log_2(e\tilde{k}) = 2(k+1)d\log_2(e(k+1)).$$

(c) For threshold functions, the VC-dimension of $\mathcal{C}$ is r, thus, the VC-dimension of $\mathcal{H}$ is upper bounded by

$$2(k+1)r\log_2(e(k+1)).$$

3.28 VC-dimension of convex combinations

Following the hint, we can think of this family of functions as a one hidden layer neural network, where the hidden layer is represented by the functions $h_t \in \mathcal{H}$, and the top layer is a threshold function characterized by $(\alpha_1, \ldots, \alpha_T)$. Denote this class of threshold functions by Δ_T. From the solution of exercise 3.27(a) we can bound the growth function of $\mathcal{F}_T$ by:

$$\Pi_{\mathcal{F}_T}(m) \leq \Pi_{\Delta_T}(m)\left(\Pi_{\mathcal{H}}(m)\right)^T.$$

From the solution to exercise 3.27(c), the VC dimension of Δ_T is at most T, and we may further denote the VC dimension of $\mathcal{H}$ by d. Applying Sauer's lemma to the growth function yields:

$$\Pi_{\Delta_T}(m) \leq \left(\frac{em}{T}\right)^T, \quad \Pi_{\mathcal{H}}(m) \leq \left(\frac{em}{d}\right)^d.$$

Thus, we have that

$$\Pi_{\mathcal{F}_T}(m) \leq \left(\frac{em}{T}\right)^T \left(\frac{em}{d}\right)^{Td}.$$

Finally, we may apply the hint in exercise 3.27(b) with $m = \max\{4T\log_2(2e), 2Td\log_2(eT)\}$ to see that

$$\left(\frac{em}{T}\right)^T \left(\frac{em}{d}\right)^{Td} < 2^{4T\log_2(2e)+2Td\log_2(eT)},$$

so that the VC Dimension of $\mathcal{F}_T$ is bounded by:

$$2T(2\log_2(2e) + d\log_2(eT)).$$

Note that a coarser but relatively simpler bound would be to write:

$$\left(\frac{em}{T}\right)^T \left(\frac{em}{d}\right)^{Td} < (em)^{T(d+1)},$$

and to apply the hint in exercise 3.27(b) with $m = 2T(d+1)\log_2(eT(d+1))$. Notice that this is actually asymptotically optimal in T and d up to log terms.

3.29 Infinite VC-dimension

(a) Theorem 3.20 shows that there exists a distribution that can force an error of $\Omega(\frac{d}{m})$. Thus, for an infinite VC-dimension, this lower bound requires an infinite number of points to achieve a bounded error and thus implies that PAC-learning is not possible.

(b) Here is a description of the algorithm. Let M be the maximum value observed after drawing m points and let p be the probability that a point greater than M be drawn. The probability that all points drawn be smaller than or equal to M is

$$(1-p)^m \leq e^{-pm}. \tag{E.74}$$

Setting $\delta/2$ to match the upper bound, yields $\delta/2 = e^{-pm}$, that is

$$p = \frac{1}{m}\log\frac{2}{\delta}. \tag{E.75}$$

To bound p by $\epsilon/2$, we can impose the following

$$\frac{1}{m}\log\frac{2}{\delta} \leq \frac{\epsilon}{2}. \tag{E.76}$$

Thus, with confidence at least $1-\delta/2$, the probability that a point greater than M be drawn is at most $\epsilon/2$ if L draws $m \geq \frac{2}{\epsilon}\log\frac{2}{\delta}$ points.

In the second stage, the problem is reduced to a finite VC-dimension M. Since PAC-learning with $(\epsilon/2, \delta/2)$ is possible for a finite dimension, this guarantees the (ϵ, δ)-PAC-learning of the full algorithm.

3.30 VC-dimension generalization bound – realizable case

(a) Let $h_0 \in \mathcal{H}_S$, then we have the following set of inequalities:

$$\begin{aligned}\mathbb{P}\Big[\sup_{h\in\mathcal{H}_S} |\widehat{R}_S(h) - \widehat{R}_{S'}(h)| > \frac{\epsilon}{2}\Big] &\geq \mathbb{P}\Big[|\widehat{R}_S(h_0) - \widehat{R}_{S'}(h_0)| > \frac{\epsilon}{2}\Big]\\ &= \mathbb{P}\Big[\widehat{R}_{S'}(h_0) > \frac{\epsilon}{2}\Big]\\ &\geq \mathbb{P}\Big[\widehat{R}(h_0) > \frac{\epsilon}{2} \mid R(h_0) > \epsilon\Big]\, \mathbb{P}[R(h_0) > \epsilon]\\ &> \mathbb{P}\Big[B(m,\epsilon) > \frac{m\epsilon}{2}\Big]\, \mathbb{P}[R(h_0) > \epsilon]\,.\end{aligned}$$

The second inequality follows from the fact that for any two random events A and B, $\mathbb{P}[A] \geq \mathbb{P}[A \wedge B] = \mathbb{P}[A|B]\,\mathbb{P}[B]$. The final equality follows, since the event we are concerned with is the probability that we get at least a fraction of $\epsilon/2$ errors on a sample of size m when the true probability of error is at least ϵ. In the case the true error rate equals ϵ, this exactly describes the probability that $B(m,\epsilon) \geq m\epsilon/2$.

(b) We apply Chebyshev's inequality to the binomial random variable $B(m,\epsilon)$, which has mean $m\epsilon$ variance $m\epsilon(1-\epsilon)$.

$$\begin{aligned}\mathbb{P}\Big[B(m,\epsilon) \leq \frac{m\epsilon}{2}\Big] &= \mathbb{P}\Big[m\epsilon - B(m,\epsilon) \geq \frac{m\epsilon}{2}\Big]\\ &\leq \frac{m\epsilon(1-\epsilon)}{(m\epsilon/2)^2} = \frac{4(1-\epsilon)}{m\epsilon} \leq \frac{4}{m\epsilon} \leq \frac{1}{2}\end{aligned}$$

where the last inequality uses the assumption $m\epsilon \geq 8$. Thus, this shows that $\mathbb{P}[B(m,\epsilon) > m\epsilon/2] > 1 - 1/2 = 1/2$. Plugging the bound into part (a) completes the question.

(c) There are $\binom{2m}{l}$ total ways to distribute the l error over the sample T and $\binom{m}{l}$ way to distribute the error such that the only hit S'. Thus, the probability of all error falling only into S' is bounded as

$$\frac{\binom{m}{l}}{\binom{2m}{l}} = \Pi_{i=0}^{l-1}\frac{m-i}{2m-i} \leq \Pi_{i=0}^{l-1}\frac{m-i}{2m-2i} \leq \frac{1}{2^l}\,.$$

(d) This follows from

$$\begin{aligned}&\mathop{\mathbb{P}}_{\substack{T\sim\mathcal{D}^{2m}:\\ T\to(S,S')}}\Big[\widehat{R}_S(h) = 0 \ \wedge\ \widehat{R}_{S'}(h) > \frac{\epsilon}{2}\Big]\\ &= \mathop{\mathbb{P}}_{\substack{T\sim\mathcal{D}^{2m}:\\ T\to(S,S')}}\Big[\widehat{R}_S(h) = 0 \ \wedge\ \widehat{R}_{S'}(h) > \frac{\epsilon}{2} \ \wedge\ \widehat{R}_T(h) > \frac{\epsilon}{2}\Big]\\ &= \mathop{\mathbb{P}}_{\substack{T\sim\mathcal{D}^{2m}:\\ T\to(S,S')}}\Big[\widehat{R}_S(h) = 0 \ \wedge\ \widehat{R}_{S'}(h) > \frac{\epsilon}{2} \mid \widehat{R}_T(h) > \frac{\epsilon}{2}\Big]\,\mathbb{P}[\widehat{R}_T(h) > \frac{\epsilon}{2}]\\ &\leq \mathop{\mathbb{P}}_{\substack{T\sim\mathcal{D}^{2m}:\\ T\to(S,S')}}\Big[\widehat{R}_S(h) = 0 \ \wedge\ \widehat{R}_{S'}(h) > \frac{\epsilon}{2} \mid \widehat{R}_T(h) > \frac{\epsilon}{2}\Big]\\ &\leq 2^{-l} \leq 2^{-\frac{m\epsilon}{2}}\,.\end{aligned}$$

(e) Using the definition of the growth function, we can provide the following union bound that is then in turn bounded using corollary 3.18:

$$\mathop{\mathbb{P}}_{\substack{T\sim\mathcal{D}^{2m}:\\ T\to(S,S')}}\Big[\exists h \in \mathcal{H}\colon \widehat{R}_S(h) = 0 \ \wedge\ \widehat{R}_{S'}(h) > \frac{\epsilon}{2}\Big] \leq \Pi_{\mathcal{H}}(2m)2^{-\frac{m\epsilon}{2}} \leq \Big(\frac{2em}{d}\Big)^d 2^{-m\epsilon/2}\,.$$

Combining part (a) through (e), we finally have,

$$\mathbb{P}[R(h) > \epsilon] \leq 2\Big(\frac{2em}{d}\Big)^d 2^{-m\epsilon/2}\,. \tag{E.77}$$

Setting the right-hand side equal to δ and solving for ϵ show that with probability at least $1-\delta$

$$\epsilon \leq \frac{1}{m}\Big(d \log \frac{2em}{d} + \log \frac{1}{\delta} + \log 2\Big)\frac{2}{\log 2}\,.$$

3.31 Covering number generalization bound.

(a) First split the term into two separate terms:

$$|L_S(h_1) - L_S(h_2)| \leq |R(h_1) - R(h_2)| + |\widehat{R}_S(h_1) - \widehat{R}_S(h_2)|$$
$$= \Big| \mathop{\mathbb{E}}_{x,y}[(h_1(x)-y)^2 - (h_2(x)-y)^2]\Big| + \Big|\frac{1}{m}\sum_{i=1}^{m}(h_1(x_i)-y_i)^2 - (h_2(x_i)-y_i)^2\Big|\,.$$

Then, expanding the term

$$(h_1(x)-y)^2 - (h_2(x)-y)^2 = (h_1(x) - h_2(x))(h_1 + h_2 - 2y)$$
$$= (h_1(x)-h_2(x))\big((h_1 - y) + (h_2 - y)\big) \leq \|h_1 - h_2\|_\infty 2M\,,$$

allows us to bound both the empirical and true error, resulting in a total bound of $4M\|h_1 - h_2\|_\infty$.

(b) This follows by splitting the event into the union of several smaller events and then using the sum rule,

$$\mathop{\mathbb{P}}_{S}\Big[\sup_{h\in\mathcal{H}} |L_S(h)| \geq \epsilon\Big]$$
$$= \mathop{\mathbb{P}}_{S}\Big[\bigvee_{i=1}^{k} \sup_{h\in B_i} |L_S(h)| \geq \epsilon\Big] \leq \sum_{i=1}^{k} \mathop{\mathbb{P}}_{S}\Big[\sup_{h\in B_i} |L_S(h)| \geq \epsilon\Big]\,.$$

(c) For any i let h_i be the center of ball B_i with radius $\frac{\epsilon}{8M}$. Note that for any $h \in \mathcal{H}$ we have $|L_S(h) - L_S(h_i)| \leq 4M\|h - h_i\|_\infty \leq \epsilon/2$. Thus, if for any $h \in B_i$ we have $|L_S(h)| \geq \epsilon$ it must be the case that $|L_S(h_i) \geq \epsilon_2|$, which shows the inequality.

To complete the bound, we use Hoeffding's inequality applied to the random variables $(h(x_i) - y_i)^2/m \leq M^2/m$, which guarantees

$$\mathop{\mathbb{P}}_{S}\Big[|L_S(h_i)| \geq \frac{\epsilon}{2}\Big] \leq 2\exp\Big(\frac{-m\epsilon^2}{2M^4}\Big)\,.$$

Chapter 5

5.1 Soft margin hyperplanes

(a) The corresponding dual problem is:

$$\max_{\alpha,\beta} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j (x_i\cdot x_j) - \sum_{i=1}^{m}\frac{(\alpha_i+\beta_i)^{k/(k-1)}}{(kC)^{1/(k-1)}}(1-\frac{1}{k})$$

subject to:

$$\sum_{i=1}^{m}\alpha_i y_i = 0 \quad \alpha \geq 0 \quad \beta \geq 0.$$

(b) Here we see that the objective function is more complex requiring an optimization over both α and β and there is the additional constraint $\beta \geq 0$.

For $k = 2$ the additional term of interest is $-\sum_{i=1}^{m}(\alpha_i + \beta_i)^2$ (to see this, note that the Hessian is negative semidefinite), which is jointly concave in α_i and β_i, which allows for convex optimization techniques.

5.2 Tighter Rademacher bound

Proceed as in the proof of theorem 5.9, but choose $\rho_k = 1/\gamma^k$. For any $\rho \in (0,1)$, there exists $k \geq 1$ such that $\rho \in (\rho_k, \rho_{k-1}]$, with $\rho_0 = 1$. For that k, $\rho \leq \rho_{k-1} = \gamma\rho_k$, thus $1/\rho_k \leq \gamma/\rho$ and $\log k = \sqrt{\log\log_\gamma(1/\rho_k)} \leq \sqrt{\log\log_2(\gamma/\rho)}$. Furthermore, for any $h \in \mathcal{H}$, $\widehat{R}_{S,\rho_k}(h) \leq \widehat{R}_{S,\rho}(h)$. Thus,

$$\mathbb{P}\left[\exists k\colon R(h) - \widehat{R}_{S,\rho}(h) > \frac{2\gamma}{\rho}\mathfrak{R}_m(\mathcal{H}) + \sqrt{\frac{\log\log_\gamma(\gamma/\rho)}{m}} + \epsilon\right] \leq 2\exp(-2m\epsilon^2),$$

which proves the statement.

5.3 Importance weighted SVM

The modified primal optimization problem can be written as

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + C\sum_{i=1}^m \xi_i p_i \\ \text{subject to} & y_i[w \cdot x_i + b] \geq 1 - \xi_i\,. \end{array}$$

The Lagrangian holding for all $w, b, \alpha_i \geq 0, \beta_i \geq 0$ is then

$$\begin{aligned} L(w,b,\alpha) \quad &= \quad \frac{1}{2}||w||^2 + C\sum_{i=1}^m \xi_i p_i \\ &\quad - \sum_{i=1}^m \alpha_i[y_i(w \cdot x_i + b) - 1 + \xi_i] - \sum_{i=1}^m \beta_i \xi_i\,. \end{aligned} \tag{E.78}$$

Then $\frac{\partial L}{\partial w}$ and $\frac{\partial L}{\partial b}$ are the same as for the regular non-separable SVM optimization problem. We also have $\frac{\partial L}{\partial \xi_i} = Cp_i - \alpha_i - \beta_i$. Thus, to satisfy the KKT conditions we have for all $i \in [m]$,

$$w = \sum_{i=1}^m \alpha_i y_i x_i \tag{E.79}$$

$$\sum_{i=1}^m \alpha_i y_i = 0 \tag{E.80}$$

$$\alpha_i + \beta_i = Cp_i \tag{E.81}$$

$$\alpha_i[y_i(w \cdot x_i + b) - 1 + \xi_i] = 0 \tag{E.82}$$

$$\beta_i \xi_i = 0\,. \tag{E.83}$$

Plugging equation E.79 into equation E.78, we get

$$\begin{aligned} L \quad &= \quad \frac{1}{2}||\sum_{i=1}^m \alpha_i y_i x_i||^2 + C\sum_{i=1}^m \xi_i p_i - \sum_{i,j=1}^m \alpha_i\alpha_j y_i y_j (x_i \cdot x_j) \\ &\quad - \sum_{i=1}^m \alpha_i y_i b + \sum_{i=1}^m \alpha_i - \sum \alpha_i \xi_i - \sum_{i=1}^m \beta_i \xi_i\,. \end{aligned} \tag{E.84}$$

Using equation E.81, we can simplify:

$$L = \sum_{i=1}^m \alpha_i - \frac{1}{2}||\sum_{i=1}^m \alpha_i y_i x_i||^2\,,$$

meaning that the objective function is the same as in the regular SVM problem. The difference is in the constraints on the optimization. Recall that our dual form holds for $\beta_i \geq 0$. Using again equation E.81, our optimization problem is to maximize L subject to the constraints:

$$\forall i \in [m], 0 \leq \alpha_i \leq Cp_i \wedge \sum_{i=1}^m \alpha_i y_i = 0.$$

5.4 Sequential Minimal Optimization (SMO).

(a) Starting from equation (5.33) and removing all terms that are constant with respect to α_1 and α_2 yields the desired result.

(b) Substituting into Ψ_1, we have:

$$\Psi_2 = \gamma - s\alpha_2 + \alpha_2 - \frac{1}{2}K_{11}(\gamma - s\alpha_2)^2 - \frac{1}{2}K_{22}\alpha_2^2 - sK_{12}(\gamma - s\alpha_2)\alpha_2 - y_1(\gamma - s\alpha_2)v_1 - y_2\alpha_2 v_2 .$$

We take the derivative to find the equation for the stationary point as follows:

$$\frac{d\Psi_2}{d\alpha_2} = -s+1+sK_{11}(\gamma-s\alpha_2)-K_{22}\alpha_2-sK_{12}(\gamma-s\alpha_2)+s^2K_{12}\alpha_2+y_1sv_1-y_2v_2 = 0 .$$

Noting that $s^2 = 1$ and rearranging terms yields the statement of interest.

(c) By definition of $f(\cdot)$ we see that $v_1 = f(\mathbf{x}_1) - y_1\alpha_1^* K_{11} - y_2\alpha_2^* K_{12}$ and similarly, $v_2 = f(\mathbf{x}_2) - y_1\alpha_1^* K_{12} - y_2\alpha_2^* K_{22}$. Using these equations along with the identities $\alpha_1^* = \gamma - s\alpha_2^*$ and $y_1 = sy_2$ we have

$$\begin{aligned}
v_1 - v_2 &= f(\mathbf{x}_1) - f(\mathbf{x}_2) - y_1\alpha_1^*(K_{11} - K_{12}) - y_2\alpha_2^*(K_{12} - K_{22}) \\
&= f(\mathbf{x}_1) - f(\mathbf{x}_2) - y_1(\gamma - s\alpha_2^*)(K_{11} - K_{12}) - y_2\alpha_2^*(K_{12} - K_{22}) \\
&= f(\mathbf{x}_1) - f(\mathbf{x}_2) - sy_2(\gamma - s\alpha_2^*)(K_{11} - K_{12}) - y_2\alpha_2^*(K_{12} - K_{22}) \\
&= f(\mathbf{x}_1) - f(\mathbf{x}_2) - sy_2\gamma(K_{11} - K_{12}) + y_2\alpha_2^*(K_{11} - K_{12}) - y_2\alpha_2^*(K_{12} - K_{22}) \\
&= f(\mathbf{x}_1) - f(\mathbf{x}_2) - sy_2\gamma(K_{11} - K_{12}) + y_2\alpha_2^*\eta .
\end{aligned}$$

(d) Combining the results from (b) and (c), we have

$$\begin{aligned}
\eta\alpha_2 &= s(K_{11} - K_{12})\gamma + y_2\Big[f(\mathbf{x}_1) - f(\mathbf{x}_2) - sy_2\gamma(K_{11} - K_{12}) + y_2\alpha_2^*\eta\Big] - s + 1 \\
&= y_2\Big[f(\mathbf{x}_1) - f(\mathbf{x}_2) + y_2\alpha_2^*\eta\Big] - s + 1 \\
&= \alpha_2^*\eta + y_2\Big[f(\mathbf{x}_1) - f(\mathbf{x}_2) - y_1 + y_2\Big] \\
&= \alpha_2^*\eta + y_2\Big[(y_2 - f(\mathbf{x}_2)) - (y_1 - f(\mathbf{x}_1))\Big] .
\end{aligned}$$

Dividing both sides by η yields the desired result.

(e) Clipping is required to ensure that the new values of α_1 and α_2 satisfy the inequality constraints $0 \leq \alpha_1, \alpha_2 \leq C$. The lower bound of 0 follows directly from this inequality constraint, as does the upper bound of C. Moreover, when $s = +1$, the lower bound of $\gamma - C$ ensures that α_2^{clip} is large enough such that $\alpha_2^{clip} + \alpha_1 = \gamma$ while respecting the constraint $\alpha_1 \leq C$. Similarly, the upper bound γ ensures that α_2^{clip} is small enough such that $\alpha_2^{clip} + \alpha_1 = \gamma$ while respecting the constraint $\alpha_1 \geq 0$.

5.5 SVM hands-on

(a) Download and install the `libsvm` software library from:

`http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

(b) Concatenate and scale data.

```
cat \
  satimage/satimage.scale.t \
  satimage/satimage.scale.val > data/train
libsvm-2.88/svm-scale \
  data/train > data/train.scaled
```

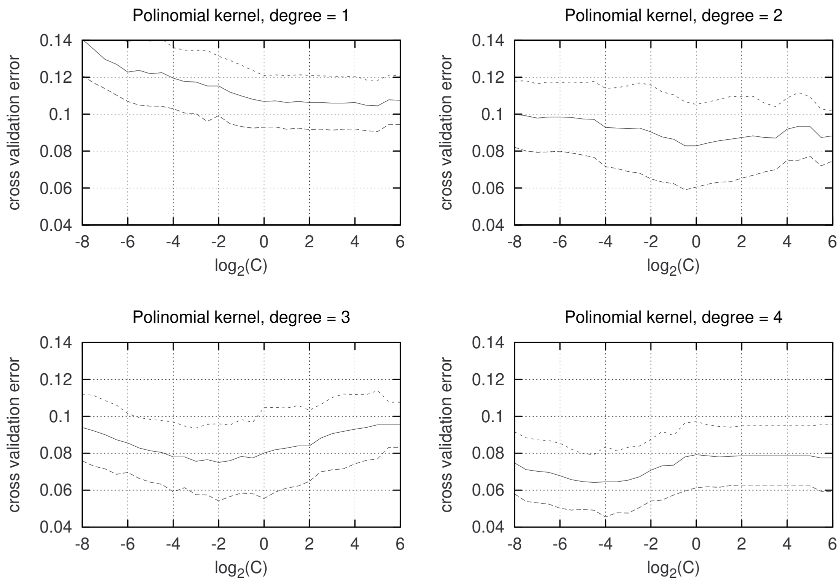

Figure E.8
Average cross-validation error plus or minus one standar deviation for different values of the trade-off constant C and of the degree of the polinomial kernel

(c) Run 10-fold cross-validation, for different values of the degree d of the polynomial kernel and of the trade-off constant C. We test $d = 1, 2, 3, 4$ and $log_2(C) = -8, -7.5, \cdots, 5.5, 6$. We step the value of the trade-off constant logarithmically as suggested by:

```
http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf
```

Which gives the cross-validation error plots shown in figure E.8.

The best values of trade-off constant C are:

```
d = 1 \colon C* = 2^(+5.0) = 32     cv-err = 10.4% +/- 1.4%
d = 2 \colon C* = 2^(+0.0) = 1      cv-err =  8.3% +/- 2.2%
d = 3 \colon C* = 2^(-2.0) = .25    cv-err =  7.5% +/- 2.1%
d = 4 \colon C* = 2^(-4.5) = .0442  cv-err =  6.4% +/- 1.5%
```

The best C measured on the validation set is $C^* = 2^{-4.5}$, with degree $d^* = 4$, which gives an average error rate of $6.4\% \pm 1.5\%$.

(d) The trade-off constant is fixed to $C^* = 2^{-4.5}$, and 10-fold cross-validation is run for degrees 1 through 4. In figure E.9 we plot the resulting cross-validation training and test errors and the average number of support vectors (`nSV` is the number of support vectors, `nBSV` is the number of bounded support vectors, i.e. whose dual variable is equal to the trade-off constant).

(e) Support vectors always lie on the margin hyperplanes when their dual variable is smaller than C. This happens for all the support vectors (`SV`) that are not bounded (`BSV`). Our measurement gives the following averages:

```
d = 1 \colon nSV - nBSV = 8.5
```

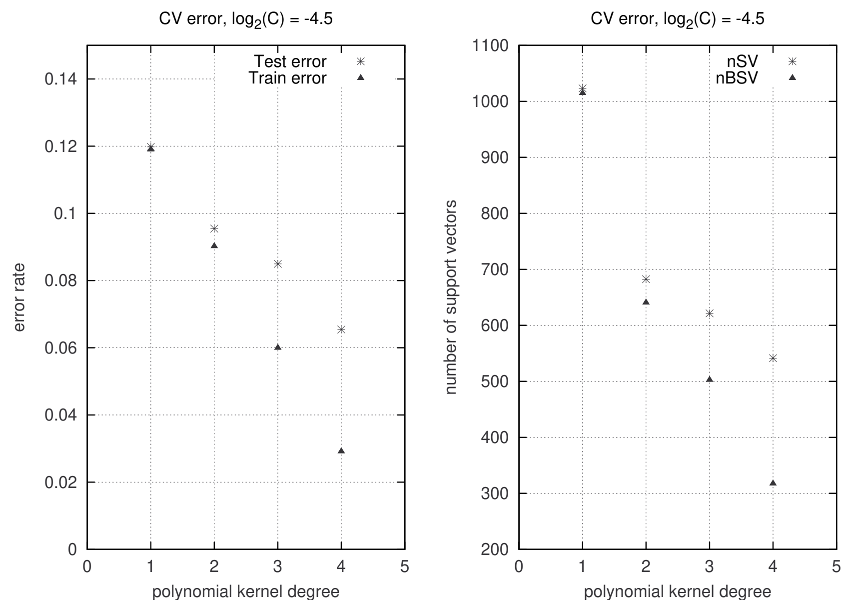

Figure E.9
Average cross-validation error rates and average number of support vectors (nSV) and of bounded support vectors(nSV) as a function of the degree of the polynomial kernel.

```
d = 2 \colon nSV - nBSV = 41.8
d = 3 \colon nSV - nBSV = 118.9
d = 4 \colon nSV - nBSV = 223.8
```

(f) We can consider the more general problem of assigning a weight k_i to every sample that will multiply its misclassification penalty. The optimization problem becomes:

$$\begin{aligned}&\text{minimize} \quad \frac{1}{2}||w||^2 + \sum_{i=1}^{m} k_i \xi_i \\ &\text{subject to} \quad y_i(wx_i + b) \leq 1 - \xi_i,\ \xi_i \geq 0,\ \forall i \in 1, \ldots, m\,.\end{aligned}$$

Moving to the dual exactly as shown in the chapter we obtain:

$$\begin{aligned}&\text{maximize} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i \cdot x_j \\ &\text{subject to} \quad 0 \leq \alpha_i \leq C k_i,\ \forall i \in 1, \ldots, m\,.\end{aligned}$$

Setting $k_i = k$ for $y_i = -1$ and $k_i = 1$ for $y_i = 1$ gives the desired problem.

(g) If k is an integer we can repeat every negative training point in the data k times. False positives will thus get penalized k times as much as false negatives.

(h) Repeating the training for $k = 1, 2, 4, 8, 16$, we find the following results:

```
k = 1 \colon  (d*, C*) = (4, 2^(-4.5)),  cv-err = 6.4% +/- 1.5%
k = 2 \colon  (d*, C*) = (4, 2^(-5.0)),  cv-err = 6.4% +/- 0.9%
```

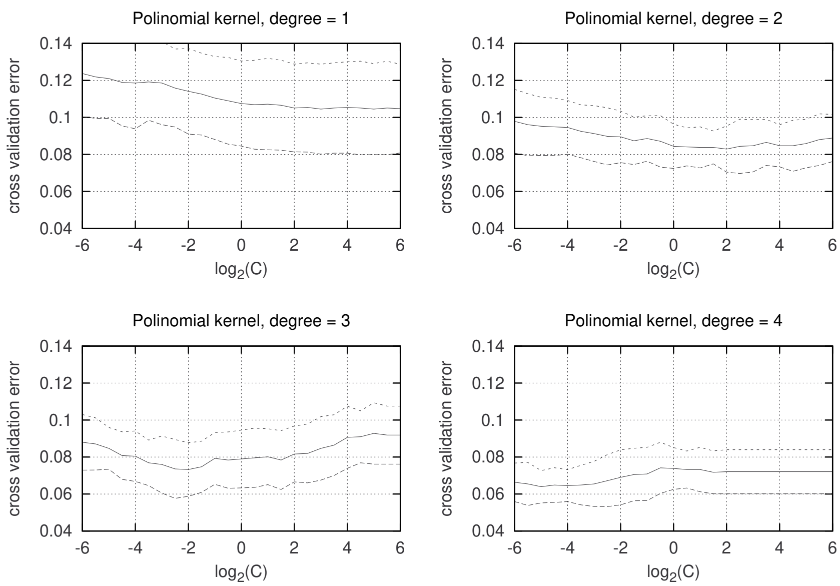

Figure E.10
As in figure E.8, but false positives are penalized twice as much as false negatives.

```
k = 4 \colon  (d*, C*) = (4, 2^(-5.5)),  cv-err = 6.1% +/- 1.7%
k = 8 \colon  (d*, C*) = (4, 2^(-3.5)),  cv-err = 6.2% +/- 1.0%
k = 16 \colon (d*, C*) = (4, 2^(-3.5)),  cv-err = 6.2% +/- 1.4%
```

Plots equivalent to the ones in figure E.8 are given in figure E.10, figure E.11, figure E.12, and figure E.13. We obtain the best average accuracy for $k = 4$.

5.6 Sparse SVM

(a) Let

$$\mathbf{x}'_i = (y_1 \mathbf{x}_i \cdot \mathbf{x}_1, \ldots, y_m \mathbf{x}_i \cdot \mathbf{x}_m)\,.$$

Then the optimization problem becomes

$$\min_{\boldsymbol{\alpha}, b, \xi} \; \frac{1}{2}||\boldsymbol{\alpha}||^2 + C \sum_{i=1}^{m} \xi_i$$
$$\text{subject to } \; y_i \left(\boldsymbol{\alpha} \cdot \mathbf{x}'_i + b\right) \geq 1 - \xi$$
$$\xi_i, \alpha_i \geq 0, i \in [m]\,,$$

which is the standard formulation of the primal SVM optimization problem on samples $\mathbf{x}'_i$, modulo the non-negativity constraints on $\boldsymbol{\alpha}_i$.

(b) The Lagrangian of (1) for all $\alpha_i \geq 0, \xi_i \geq 0, b, \alpha'_i \geq 0, \beta_i \geq 0, \gamma_i \geq 0, i \in [m]$ is

$$L = \frac{1}{2}||\boldsymbol{\alpha}||^2 + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha'_i (y_i(\boldsymbol{\alpha} \cdot \mathbf{x}'_i + b) - 1 + \xi_i) - \sum_{i=1}^{m} \beta_i \xi_i - \sum_{i=1}^{m} \gamma_i \alpha_i\,,$$

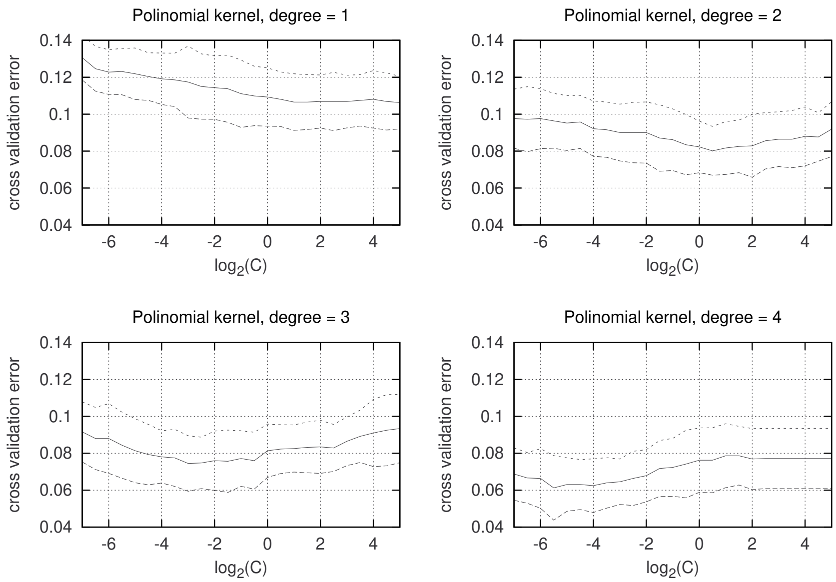

Figure E.11
As in figure E.8, but false positives are penalized four times as much as false negatives.

and the KKT conditions are

$$\nabla_\alpha L = 0 \quad \Leftrightarrow \quad \boldsymbol{\alpha} = \sum_{i=1}^{m} \alpha'_i y_i \mathbf{x}'_i + \boldsymbol{\gamma}$$

$$\nabla_b L = 0 \quad \Leftrightarrow \quad \sum_{i=1}^{m} \alpha'_i y_i = 0$$

$$\nabla_{\xi_i} L = 0 \quad \Leftrightarrow \quad \alpha'_i + \beta_i = C$$

and

$$\alpha'_i(y_i(\boldsymbol{\alpha} \cdot \mathbf{x}'_i + b) - 1 - \xi_i) = 0$$
$$\beta_i \xi_i = 0$$
$$\gamma_i \alpha_i = 0.$$

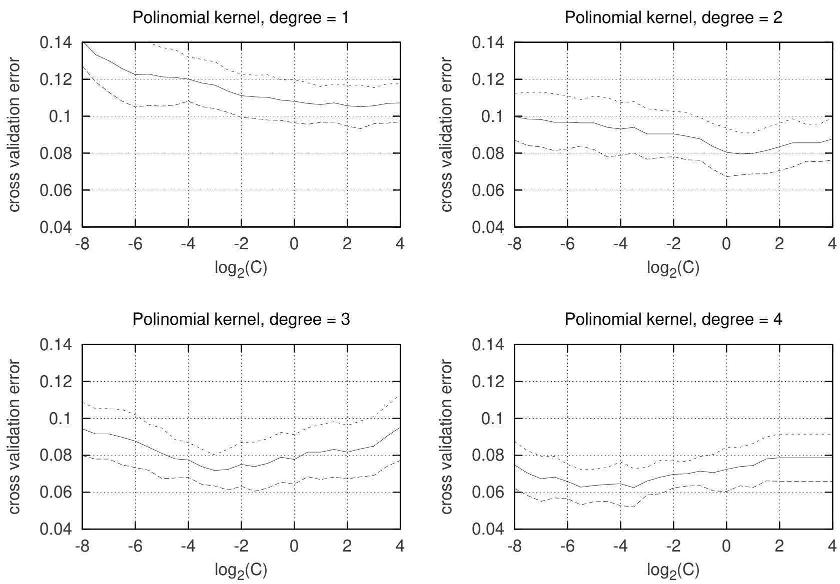

Figure E.12
As in figure E.8, but false positives are penalized eight times as much as false negatives.

Using the KKT conditions on L we get

$$\begin{aligned}
L &= \frac{1}{2}\left(\sum_{i=1}^{m}\alpha_i' y_i \mathbf{x}_i' + \boldsymbol{\gamma}\right)\cdot\left(\sum_{j=1}^{m}\alpha_j' y_j \mathbf{x}_j' + \boldsymbol{\gamma}\right) + C\sum_{i=1}^{m}\xi_i \\
&\quad - \sum_{i=1}^{m}\alpha_i'\left(y_i\left(\left(\left(\sum_{j=1}^{m}\alpha_j' y_j \mathbf{x}_j' + \boldsymbol{\gamma}\right)\cdot \mathbf{x}_i' + b\right) - 1 + \xi_i\right)\right) \\
&\quad \cancel{-\sum_{i=1}^{m}\beta_i\xi_i} \cancel{- \sum_{i=1}^{m}\gamma_i\alpha_i} \\
&= -\frac{1}{2}\sum_{i=1}^{m}\alpha_i' y_i \mathbf{x}_i' \cdot \left(\sum_{j=1}^{m}\alpha_j' y_j \mathbf{x}_j' + \gamma\right) + \cancel{\frac{1}{2}\boldsymbol{\gamma}\cdot\boldsymbol{\alpha}} \\
&\quad + \sum_{i=1}^{m} C\xi_i - \alpha_i'\left(y_i b - 1 + \xi_i\right) \\
&= \sum_{i=1}^{m}\alpha_i' - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i'\alpha_j' y_i y_j \mathbf{x}_i'^{\top}\left(\mathbf{x}_j' + \boldsymbol{\gamma}\right) \\
&\quad \cancel{+\sum_{i=1}^{m}(C - \alpha_i')\xi_i} \cancel{- \sum_{i=1}^{m}\alpha_i' y_i b}.
\end{aligned}$$

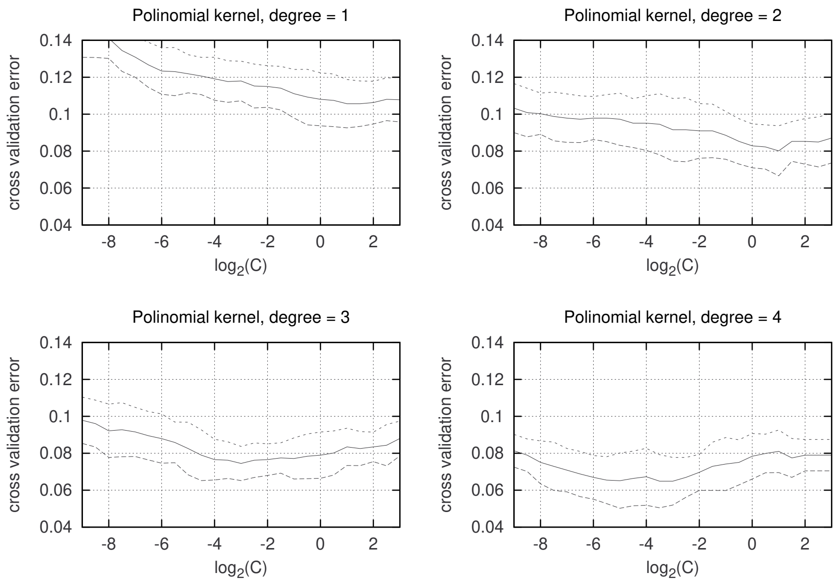

Figure E.13
As in figure E.8, but false positives are penalized sixteen times as much as false-negatives.

Thus the dual optimization problem is

$$\begin{aligned}\max_{\alpha',\gamma} \quad & \sum_{i=1}^{m} \alpha'_i - \frac{1}{2}\sum_{i,j=1}^{m} \alpha'_i\alpha'_j y_i y_j \mathbf{x}'_i \cdot \big(\mathbf{x}'_j + \boldsymbol{\gamma}\big) \\ \text{subject to} \quad & \sum_{i=1}^{m} \alpha'_i y_i = 0 \\ & 0 \leq \alpha'_i \leq C, \gamma_i \geq 0, i \in [m]\,.\end{aligned}$$

5.7 VC-dimension of canonical hyperplanes

(a) By definition of $\{\mathbf{x}_1, \ldots, \mathbf{x}_d\}$, for all $\mathbf{y} = (y_1, \ldots, y_d) \in \{-1, +1\}^d$, there exists $\mathbf{w}$ such that,

$$\forall i \in [d], 1 \leq y_i(\mathbf{w} \cdot \mathbf{x}_i)\,.$$

Summing up these inequalities yields

$$d \leq \mathbf{w} \cdot \sum_{i=1}^{d} y_i \mathbf{x}_i \leq \|\mathbf{w}\| \left\|\sum_{i=1}^{d} y_i \mathbf{x}_i\right\| \leq \Lambda \left\|\sum_{i=1}^{d} y_i \mathbf{x}_i\right\|\,.$$

(b) Since this inequality holds for all $\mathbf{y} \in \{-1, +1\}^d$, it also holds on expectation over $y_1, \ldots, y_d$ drawn i.i.d. according to a uniform distribution over $\{-1, +1\}$. In view of the independence assumption, for $i \neq j$ we have $\mathbb{E}[y_i y_j] = \mathbb{E}[y_i]\,\mathbb{E}[y_j]$. Thus, since the distribution is uniform, $\mathbb{E}[y_i y_j] = 0$ if $i \neq j$, $\mathbb{E}[y_i y_j] = 1$ otherwise. This gives

$$\begin{aligned}
d &\leq \Lambda \mathop{\mathbb{E}}_{\mathbf{y}} \left[\left\| \sum_{i=1}^{d} y_i \mathbf{x}_i \right\| \right] && \text{(taking expectations)} \\
&\leq \Lambda \left[\mathop{\mathbb{E}}_{\mathbf{y}} \left[\left\| \sum_{i=1}^{d} y_i \mathbf{x}_i \right\|^2 \right] \right]^{\frac{1}{2}} && \text{(Jensen's inequality)} \\
&= \Lambda \left[\sum_{i,j=1}^{d} \mathop{\mathbb{E}}_{\mathbf{y}} [y_i y_j] \left(\mathbf{x}_i \cdot \mathbf{x}_j \right) \right]^{\frac{1}{2}} \\
&= \Lambda \left[\sum_{i=1}^{d} (\mathbf{x}_i \cdot \mathbf{x}_i) \right]^{\frac{1}{2}}.
\end{aligned}$$

Thus, $\sqrt{d} \leq \Lambda r$, which completes the proof.

(c) In view of the previous inequality, we can write $d \leq \Lambda \big[d r^2 \big]^{\frac{1}{2}} = \Lambda r \sqrt{d}$.

Chapter 6

6.1 This follows directly the Cauchy-Schwarz inequality:

$$|\Phi(x)^\top \Phi(y)| \leq \|\Phi(x)\| \|\Phi(y)\|, \tag{E.85}$$

where Φ is a feature mapping associated to K.

6.2 (a) Since $\cos(x-y) = \cos x \cos y + \sin x \sin y$, $K(x,y)$ can be written as the dot product of the vectors

$$\Phi(x) = \begin{bmatrix} \cos x \\ \sin x \end{bmatrix} \quad \text{and} \quad \Phi(y) = \begin{bmatrix} \cos y \\ \sin y \end{bmatrix}; \tag{E.86}$$

thus it is PDS.

(b) This is a consequence of the fact that the kernel of the previous question is PDS since $\sum_{ij} c_i c_j \cos(x_i^2 - x_j^2) = \sum_{ij} c_i c_j \cos(x'_i - x'_j)$, with $x'_i = x_i^2$ for all i.
The solution for this question is similar to that of the previous question.

(c) Since the product and sum of PDS kernels is PDS, it suffices to show that $k \colon x \mapsto \cos(x^2 - y^2)$ is PDS over $\mathbb{R} \times \mathbb{R}$, which was proven in part (b).

(d) For any $x_1, \ldots, x_m \in \mathbb{R}$, $c_1, \ldots, c_m \in \mathbb{R}$, and $a \geq 0$, let $f(a)$ be defined by

$$f(a) = \sum_{i,j} c_i c_j K(x_i, x_j) a^{x_i + x_j}. \tag{E.87}$$

Then, $f'(a) = \sum_{i,j} c_i c_j a^{x_i + x_j - 1} = \frac{1}{a} \|(c_i a^{x_i})_i\|^2 \geq 0$. Therefore f is monotonically increasing. Thus, $f(1) \geq f(0)$, that is $\sum_{i,j} c_i c_j K(x_i, x_j) \geq 0$.

(e) Rewriting the cosine in terms of the dot product, we have

$$K(\mathbf{x}, \mathbf{x}') = \cos \angle(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x} \cdot \mathbf{x}'}{\|\mathbf{x}\| \|\mathbf{x}'\|}.$$

Thus, the cosine kernel is just a scaling of the standard dot product, which is a PDS kernel. Hence, the cosine kernel is also PDS.

(f) For all $x, x' \in \mathbb{R}$,

$$\begin{aligned}
[\sin(x' - x)]^2 &= 1 - [\cos(x' - x)]^2 \\
&= 1 - [\cos x' \cos x + \sin x' \sin x]^2 \\
&= 1 - (\mathbf{u}(x') \cdot \mathbf{u}(x))^2,
\end{aligned}$$

iii. Using the hint and the result of the previous question, we can write

$$\begin{aligned}\widehat{R}_{S,\rho}(f) &\leq \prod_t \left(1 - 2\frac{\left(\frac{1-\rho}{2} - \epsilon_t\right)^2}{1-\rho^2}\right)\\ &\leq \prod_t \left(\exp\left(-2\frac{\left(\frac{1-\rho}{2} - \epsilon_t\right)^2}{1-\rho^2}\right)\right)\\ &= \exp\left(-2\frac{\left(\frac{1-\rho}{2} - \epsilon_t\right)^2}{1-\rho^2}T\right)\\ &\leq \exp\left(-\frac{2\gamma^2 T}{1-\rho^2}\right).\end{aligned}$$

Thus, if the upper bound is less that $1/m$, then $\widehat{R}_\rho(f) = 0$ and every training point has margin at least ρ. The inequality $\exp\left(-\frac{2\gamma^2 T}{1-\rho^2}\right) < 1/m$ is equivalent to $T > \frac{(\log m)(1-\rho^2)}{2\gamma^2}$.

Chapter 8

8.1 Perceptron lower bound

Let w be the weight vector. Since each update is of the form $w \leftarrow w + y_i x_i$ and since the components of the sample points are integers, the components of w are also integers.

Let $n_1, \ldots, n_N \in \mathbb{Z}$ denote the components of w. w correctly classifies all points iff $y_i(w \cdot x_i) > 0$ for $i = 1, \ldots, m$, that is,

$$\begin{cases} n_1 > 0 \\ n_1 - n_2 < 0 \\ -n_1 - n_2 + n_3 > 0 \\ \ldots \\ (-1)^N(n_1 + n_2 + \ldots + n_{N-1} - n_N) < 0 \end{cases} \Leftrightarrow \begin{cases} n_1 > 0 \\ n_2 > n_1 \\ n_3 > n_1 + n_2 \\ \ldots \\ n_N > n_1 + n_2 + \ldots + n_{N-1}. \end{cases}$$

These last inequalities show that the data is linearly separable with $w = (1, 2, \ldots, 2^{N-1})$. They also imply that $n_1 \geq 1, n_2 \geq 2, n_3 \geq 4, \ldots, n_N \geq 2^{N-1}$. Since each update can at most increment n_N by 1, the number of updates is at least $2^{N-1} = \Omega(2^N)$.

8.2 Generalized mistake bound

The bound is unaffected, as shown by the following, using the same definitions and steps as in this chapter:

$$\begin{aligned}
M\rho &\leq \frac{\mathbf{v}\cdot\sum_{t\in I} y_t\mathbf{x}_t}{\|\mathbf{v}\|}\\
&= \frac{\mathbf{v}\cdot\sum_{t\in I}(\mathbf{w}_{t+1}-\mathbf{w}_t)/\eta}{\|\mathbf{v}\|} \quad \text{(definition of updates)}\\
&= \frac{\mathbf{v}\cdot\mathbf{w}_{T+1}}{\eta\|\mathbf{v}\|}\\
&\leq \|\mathbf{w}_{T+1}\|/\eta \qquad \text{(Cauchy-Schwarz ineq.)}\\
&= \|\mathbf{w}_{t_m}+\eta y_{t_m}\mathbf{x}_{t_m}\|/\eta \qquad (t_m \text{ largest } t \text{ in } I)\\
&= \Big[\|\mathbf{w}_{t_m}\|^2+\eta^2\|\mathbf{x}_{t_m}\|^2+\underbrace{\eta y_{t_m}\mathbf{w}_{t_m}\cdot\mathbf{x}_{t_m}}_{\leq 0}\Big]^{1/2}/\eta\\
&\leq \Big[\|\mathbf{w}_{t_m}\|^2+\eta^2R^2\Big]^{1/2}/\eta\\
&\leq \Big[M\eta^2R^2\Big]^{1/2}/\eta = \sqrt{M}R. \quad \text{(applying the same to previous } t\text{s in } I).
\end{aligned}$$

8.3 Sparse instances

Clearly, it takes T updates and leads to $\mathbf{w}=\sum_{t=1}^T y_t\mathbf{x}_t$. Let $\mathbf{u}\in\mathbb{R}^T$ be a vector of norm 1 defining a separating hyperplane, thus $y_t\mathbf{u}\cdot\mathbf{x}_t = y_tu_t \geq 0$ for all $t\in[T]$. To obtain the maximum margin ρ, we seek a vector $\mathbf{u}$ maximizing the minimum of y_tu_t with $y_tu_t\geq 0$ for all t and $\|\mathbf{u}\|=1$. By symmetry, all y_tu_ts are equal, thus $u_t = y_t/\sqrt{T}$ for all $t\in[T]$ and $\rho = 1/\sqrt{T}$. Thus, Novikoff's bound gives $R^2/\rho^2 = 1/(1/T) = T$.

8.4 Tightness of lower bound

The lower bound is tight. This follows from the tightness of the Khintchine-Kahane inequality, which is the only inequality used in the proof.

8.5 On-line SVM algorithm

First we write the optimization in the following equivalent form:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^m \max\big(0, 1-y_i(\mathbf{w}\cdot\mathbf{x}_i)\big)\,.$$

Then, using the general update rule in equation (8.22), we get the update rule,

$$\mathbf{w}_{t+1} \leftarrow \begin{cases}\mathbf{w}_t - \eta(\mathbf{w}_t - Cy_t\mathbf{x}_t) & \text{if } y_t(\mathbf{w}_t\cdot\mathbf{x}_t) < 1,\\ \mathbf{w}_t - \eta\mathbf{w}_t & \text{if } y_t(\mathbf{w}_t\cdot\mathbf{x}_t) > 1,\\ \mathbf{w}_t & \text{otherwise,}\end{cases}$$

which corresponds exactly to the update in the pseudocode.

8.6 Margin Perceptron

(a) By assumption, there exists $\mathbf{v}\in\mathbb{R}^N$ such that for all $t\in[T]$, $\rho\leq\frac{y_t(\mathbf{v}\cdot\mathbf{x}_t)}{\|\mathbf{v}\|}$, where ρ is the maximum margin achievable on S. Summing up these inequalities gives

$$\begin{aligned}
M\rho \leq \frac{\mathbf{v}\cdot\sum_{t\in I}y_t\mathbf{x}_t}{\|\mathbf{v}\|} &\leq \Big\|\sum_{t\in I}y_t\mathbf{x}_t\Big\| && \text{(Cauchy-Schwarz inequality)}\\
&= \Big\|\sum_{t\in I}(\mathbf{w}_{t+1}-\mathbf{w}_t)\Big\| && \text{(definition of updates)}\\
&= \|\mathbf{w}_{T+1}\| && \text{(telescoping sum, } \mathbf{w}_0 = 0).
\end{aligned}$$

(b) For any $t \in I$, by definition of the update, $\mathbf{w}_{t+1} = \mathbf{w}_t + y_t\mathbf{x}_t$; thus,

$$\begin{aligned}
\|\mathbf{w}_{t+1}\|^2 &= \|\mathbf{w}_t\|^2 + \|\mathbf{x}_t\|^2 + 2y_t\mathbf{w}_t \cdot \mathbf{x}_t \\
&\leq \|\mathbf{w}_t\|^2 + \|\mathbf{x}_t\|^2 + \|\mathbf{w}_t\|\rho \qquad \text{(def. of update condition)} \\
&\leq \|\mathbf{w}_t\|^2 + R^2 + \|\mathbf{w}_t\|\rho + \rho^2/4 \\
&= (\|\mathbf{w}_t\| + \rho/2)^2 + R^2.
\end{aligned}$$

(c) In view of the previous result, $\|\mathbf{w}_{t+1}\|^2 - (\|\mathbf{w}_t\| + \rho/2)^2 \leq R^2$, that is

$$\begin{aligned}
&(\|\mathbf{w}_{t+1}\| - \|\mathbf{w}_t\| - \rho/2)(\|\mathbf{w}_{t+1}\| + \|\mathbf{w}_t\| + \rho/2) \leq R^2 \\
\Longrightarrow\ &(\|\mathbf{w}_{t+1}\| - \|\mathbf{w}_t\| - \rho/2) \leq \frac{R^2}{\|\mathbf{w}_{t+1}\| + \|\mathbf{w}_t\| + \rho/2} \\
\Longrightarrow\ &\|\mathbf{w}_{t+1}\| \leq \|\mathbf{w}_t\| + \rho/2 + \frac{R^2}{\|\mathbf{w}_{t+1}\| + \|\mathbf{w}_t\| + \rho/2}.
\end{aligned}$$

(d) If $\|\mathbf{w}_t\| \geq \frac{4R^2}{\rho}$ or $\|\mathbf{w}_{t+1}\| \geq \frac{4R^2}{\rho}$, then $\|\mathbf{w}_{t+1}\| + \|\mathbf{w}_t\| + \rho/2 \geq \frac{4R^2}{\rho}$, thus

$$\frac{R^2}{\|\mathbf{w}_{t+1}\| + \|\mathbf{w}_t\| + \rho/2} \leq \frac{R^2}{4R^2/\rho} = \frac{\rho}{4}.$$

In view of this, the inequality of the previous question implies

$$\begin{aligned}
&\|\mathbf{w}_{t+1}\| \leq \|\mathbf{w}_t\| + \rho/2 + \frac{R^2}{\|\mathbf{w}_{t+1}\| + \|\mathbf{w}_t\| + \rho/2} \\
\Longrightarrow\ &\|\mathbf{w}_{t+1}\| \leq \|\mathbf{w}_t\| + \rho/2 + \frac{\rho}{4} = \|\mathbf{w}_t\| + \frac{3}{4}\rho.
\end{aligned}$$

(e) Since $\mathbf{w}_1 = y_1\mathbf{x}_1$, $\|\mathbf{w}_1\| = \|\mathbf{x}_1\| \leq R$. The margin ρ is at most twice the radius R, thus, $\rho \leq 2R$ and $2R/\rho \geq 1$. This implies that $\|\mathbf{w}_1\| \leq R \leq 2R^2/\rho$. Since $\|\mathbf{w}_1\| \leq 2R^2/\rho$ and $\|\mathbf{w}_{T+1}\| \geq \frac{4R^2}{\rho}$, there must exist at least one update time $t \in I$ at which $\|\mathbf{w}_t\| \leq \frac{4R^2}{\rho}$ and $\|\mathbf{w}_{t+1}\| \geq \frac{4R^2}{\rho}$. The set of such times t is non empty and thus admits a largest element t_0.

(f) By definition of t_0, for any $t \geq t_0$, $\|\mathbf{w}_{t+1}\| \geq \frac{4R^2}{\rho}$. Thus, by the inequality of part (d), the following holds for any $t \geq t_0$,

$$\|\mathbf{w}_{t+1}\| \leq \|\mathbf{w}_t\| + \frac{3}{4}\rho.$$

This implies that

$$\begin{aligned}
\|\mathbf{w}_{T+1}\| &\leq \|\mathbf{w}_{t_0}\| + \big|\{t_0, \ldots, T\} \cap I\big|\frac{3}{4}\rho \\
&\leq \|\mathbf{w}_{t_0}\| + M\frac{3}{4}\rho \\
&\leq \frac{4R^2}{\rho} + M\frac{3}{4}\rho.
\end{aligned}$$

By the first question $M\rho \leq \|\mathbf{w}_{T+1}\|$; therefore,

$$M\rho \leq \frac{4R^2}{\rho} + M\frac{3}{4}\rho \iff M\rho/4 \leq 4R^2/\rho \iff M \leq 16R^2/\rho^2.$$

8.7 Generalized RWM

(a) Observe that:

$$\begin{aligned}
W_{t+1} &= \sum_{i=1}^{N}(1 - (1-\beta)l_{t,i})w_{t,i} \\
&= W_t - W_t(1-\beta)l_{t,i})w_{t,i}/W_t = W_t(1 - (1-\beta)L_t).
\end{aligned}$$

Thus, $W_{T+1} = N \prod_{t=1}^{T}(1 - (1-\beta)L_t)$ and

$$\begin{aligned}\log W_{T+1} &= \log N + \sum_{t=1}^{T} \log(1 - (1-\beta)L_t)\\ &\le \log N + \sum_{t=1}^{T} -(1-\beta)L_t\\ &= \log N - (1-\beta)\mathcal{L}_T.\end{aligned}$$

(b) For all $i \in [N]$,

$$\begin{aligned}\log W_{T+1} &\ge \log w_{T+1,i}\\ &= \log\big(\prod_{t=1}^{T}(1 - (1-\beta)l_{t,i}\big)\\ &= \sum_{t=1}^{T} \log(1 - (1-\beta)l_{t,i}\\ &\ge \sum_{t=1}^{T} -(1-\beta)l_{t,i} - (1-\beta)^2 l_{t,i}^2\\ &= -(1-\beta)\mathcal{L}_{T,i} - (1-\beta)^2 \sum_{t=1}^{T} l_{t,i}^2.\end{aligned}$$

(c) Comparing the lower and upper bounds gives:

$$\begin{aligned}&-(1-\beta)\mathcal{L}_{T,i} - (1-\beta)^2 \sum_{t=1}^{T} l_{t,i}^2 \le \log N - (1-\beta)\mathcal{L}_T\\ \Longrightarrow\ &\mathcal{L}_T \le \mathcal{L}_{T,i} + \frac{\log N}{(1-\beta)} + (1-\beta)\sum_{t=1}^{T} l_{t,i}^2.\end{aligned}$$

Clearly, for any $i \in [N]$, $\sum_{t=1}^{T} l_{t,i}^2 \le T$. Thus, for all $i \in [N]$,

$$\mathcal{L}_T \le \mathcal{L}_{T,i} + \frac{\log N}{(1-\beta)} + (1-\beta)T,$$

in particular, $\mathcal{L}_T \le \mathcal{L}_T^{\min} + \frac{\log N}{(1-\beta)} + (1-\beta)T$. Differentiating with respect to β and setting the result to zero gives $\frac{\log N}{(1-\beta)^2} - T = 0$, as in the case of the RWM algorithm. Thus, for $\beta = \max\{1/2, 1 - \sqrt{(\log N)/T}\}$, $\mathcal{L}_T \le \mathcal{L}_T^{\min} + 2\sqrt{T\log N}$, that is $R_T \le 2\sqrt{T\log N}$.

8.9 General inequality

(a) We analyze the function $f(x) = \log(1-x) + x + \frac{x^2}{\alpha}$ and show that it is positive for $x \in [0, 1 - \frac{\alpha}{2}]$. First note that $f(0) = 0$, then note that

$$\begin{aligned}f'(x) \ge 0 &\iff -1 + 1 - x\frac{2}{\alpha}x(1-x) \ge 0\\ &\iff \frac{2}{\alpha}x(1-x) \ge x\\ &\iff \frac{2}{\alpha}(1-x) \ge 1 \quad (\text{for } x \ge 0)\\ &\iff x \le 1 - \frac{\alpha}{2}.\end{aligned}$$

Thus, the derivative is only increasing for $x \in [1, 1 - \frac{\alpha}{2}]$, which implies that the function is positive for the same interval.

In order to apply the inequality, inequality the valid range of β is $[\frac{2}{\alpha}, 1)$.

(b) The bound follows directly using the same steps as in the original proof, but with the general inequality. The optimal choice of β is $\max\{\frac{\alpha}{2}, 1-\sqrt{\alpha(\log N)/T}\}$, which gives

$$R_T \leq \sqrt{\frac{\log(N)T}{\alpha}} + \sqrt{\alpha \log(N)T}\,.$$

(c) Setting α close to 2 forces β close to 1, which results in an algorithm that downweights experts in a very conservative fashion. From the bound in part (b) we see that $\alpha = 1$, as is used in the chapter, is the optimal choice.

8.10 On-line to batch — non-convex loss

(a) We use the following series of inequalities:

$$\begin{aligned}
&\min_{i\in[T]} (R(h_i) + 2c_\delta(T-i+1)) \\
&\quad\leq \frac{1}{T}\sum_{i=1}^{T}(R(h_i) + 2c_\delta(T-i+1)) \\
&\quad= \frac{1}{T}\sum_{i=1}^{T} R(h_{i-1}) + \frac{2}{T}\sum_{i=0}^{T-1}\sqrt{\frac{1}{2(T-i)}\log\frac{T(T+1)}{\delta}} \\
&\quad< \frac{1}{T}\sum_{i=1}^{T} R(h_{i-1}) + \frac{2}{T}\sum_{i=0}^{T-1}\sqrt{\frac{1}{2(T-i)}\log\Big(\frac{(T+1)}{\delta}\Big)^2} \\
&\quad= \frac{1}{T}\sum_{i=1}^{T} R(h_{i-1}) + \frac{2}{T}\sum_{i=0}^{T-1}\sqrt{\frac{1}{(T-i)}\log\frac{(T+1)}{\delta}} \\
&\quad\leq \frac{1}{T}\sum_{i=1}^{T} R(h_{i-1}) + 4\sqrt{\frac{1}{T}\log\frac{T+1}{\delta}}\,.
\end{aligned}$$

The first inequality follows, since the minimum is always less than or equal to the average and the final inequality follows from $\sum_{i=0}^{T-1}\sqrt{1/(T-i)} = \sum_{i=1}^{T}\sqrt{1/i} \leq 2\sqrt{T}$.

(b) Coupling the inequality of part (a) with the high probability statement of lemma 8.14 to bound $\frac{1}{T}\sum_{i=1}^{T} R(h_i)$ shows the desired bound.

(c) The square-root terms in part (b) can be bounded further by $6\sqrt{\frac{1}{T}\log\frac{2(T+1)}{\delta}}$.
Now, note that for two events A and B that each occur with probability at least $1-\delta$,

$$\begin{aligned}
&\mathbb{P}[\neg A \cup \neg B] \leq \mathbb{P}[\neg A] + \mathbb{P}[\neg B] \leq 2\delta \\
&\iff \mathbb{P}[A \wedge B] \geq 1 - 2\delta\,.
\end{aligned}$$

Thus, the probability that both bounds in (b) and (c) hold simultaneously is at least $1-2\delta$; substituting δ with $\delta/2$ everywhere completes the bound.

8.11 On-line to batch — kernel Perceptron margin bound

(a) Let $\mathbf{u} \in \mathbb{H}$ with $\|\mathbf{u}\| = 1$. Observe that for any $t \in [T]$, we can write

$$1 - \left(1 - \frac{y_t(\mathbf{u}\cdot\Phi(x_t))}{\rho}\right)_+ \leq \frac{y_t(\mathbf{u}\cdot\Phi(x_t))}{\rho}.$$

Let $\mathcal{I}$ be the set of $t \in [T]$ at which kernel Perceptron makes an update, and let M_T be the total number of updates made, then, summing up the previous inequalities over all such ts and using the Cauchy-Schwarz inequality yields

$$M_T - \sum_{t\in\mathcal{I}}\left(1 - \frac{y_t(\mathbf{u}\cdot\Phi(x_t))}{\rho}\right) \leq \sum_{t\in\mathcal{I}}\frac{y_t(\mathbf{u}\cdot\Phi(x_t))}{\rho} \leq \frac{\|\sum_{t\in\mathcal{I}} y_t\Phi(x_t)\|}{\rho}.$$

In view of the proof for separable case of the perceptron algorithm (theorem 8.8), the right-hand side can be bounded by $\frac{\sqrt{\sum_{t\in\mathcal{I}} K(x_t,x_t)}}{\rho}$. Thus, for any $\mathbf{u} \in \mathbb{H}$ with $\|\mathbf{u}\| = 1$,

$$M_T \leq \sum_{t\in\mathcal{I}} \left(1 - \frac{y_t(\mathbf{u}\cdot\Phi(x_t))}{\rho}\right)_+ + \frac{\sqrt{\sum_{t\in\mathcal{I}} K(x_t,x_t)}}{\rho}.$$

(b) Plugging in the result from part (a) into (8.31) gives the desired bound. This bound is with respect to expected error of best in class, while corollary 6.13 is with respect to empirical error of selected hypothesis.

Chapter 9

9.5 Decision trees. A binary decision tree with n nodes has exactly $n+1$ leaves. Each node can be labeled with an integer from $\{1,\ldots,N\}$ indicating which dimension is queried to make a binary split and each leaf can be labeled with ±1 to indicate the classification made at that leaf. Fix an ordering of the nodes and leaves and consider all possible labelings of this sequence. There can be no more than $(N+2)^{2n+1}$ distinct binary trees and, thus, the VC-dimension of this finite set of hypotheses can be no larger than $(2n+1)\log(N+2) = O(n\log N)$.

Chapter 11

11.1 Pseudo-dimension and monotonic functions

If for some $m > 0$, there exists $(t_1,\ldots,t_m)$ and a set of points $(x_1,\ldots,x_m)$ that $\mathcal{H}$ shatters, then $\phi\circ\mathcal{H}$ can also shatter it. To see that, note that if for some $h\in\mathcal{H}$,

$$h(x_i) \geq t_i\,,$$

then by the monotonic property of ϕ,

$$\phi(h(x_i)) \geq \phi(t_i)\,.$$

A similar argument holds for the case $h(x_i) < t_i$. Thus, $\phi\circ\mathcal{H}$ can shatter the set of points $(x_1,\ldots,x_m)$ with thresholds $(\phi(t_1),\ldots,\phi(t_m))$, and this proves that $\text{Pdim}(\phi\circ\mathcal{H}) \geq \text{Pdim}(\mathcal{H})$. Since ϕ is strictly monotonic, it is invertible, and a similar argument with ϕ^{-1} can be used to show $\text{Pdim}(\mathcal{H}) \geq \text{Pdim}(\phi\circ\mathcal{H})$.

11.2 Pseudo-dimension of linear functions

From equation (11.3) we have that

$$\text{Pdim}(\mathcal{H}) = \text{VCdim}\Big(\big\{(\mathbf{x},t)\mapsto 1_{(\mathbf{w}^\top\mathbf{x}-t)>0} \colon h\in\mathcal{H}\big\}\Big)\,.$$

Note that $1_{(\mathbf{w}^\top\mathbf{x}-t)>0} = \frac{1+\text{sgn}(\mathbf{w}^\top\mathbf{x}-t)}{2}$ is a linear separator with *fixed* offset. It is easy to show that the VC-dimension of such a hypothesis class is d (as oppose to $d+1$ in the case of linear separators with a free offset parameter).

11.3 Linear regression

(a) In order for the matrix $\mathbf{X}\mathbf{X}^\top$ to be invertible, we need the number of (linearly independent) examples to outnumber the number of features used to represent each example.

(b) For any $\mathbf{v}\in\mathbb{R}^m$ we can choose $\mathbf{w} = (\mathbf{X}\mathbf{X}^\top)^\dagger\mathbf{X}\mathbf{y} + (\mathbf{I} - (\mathbf{X}^\dagger)^\top\mathbf{X}^\top)\mathbf{v}$. To see this, observe that

$$\begin{aligned}\mathbf{X}^\top(\mathbf{I} - (\mathbf{X}^\dagger)^\top\mathbf{X}^\top) &= \mathbf{X}^\top - \mathbf{X}^\top(\mathbf{X}^\dagger)^\top\mathbf{X}^\top\\ &= \mathbf{X}^\top - (\mathbf{X}\mathbf{X}^\dagger\mathbf{X})^\top\\ &= \mathbf{X}^\top - \mathbf{X}^\top = \mathbf{0}\,,\end{aligned}$$

and hence, we have $\mathbf{X}^\top \mathbf{w} = \mathbf{X}^\dagger \mathbf{X}\mathbf{y}$.

11.4 Perturbed kernels

(a) Using the closed form solutions $\boldsymbol{\alpha} = (\mathbf{K}+\lambda\mathbf{I})^{-1}\mathbf{y}$ and the fact $\mathbf{M'}^{-1}-\mathbf{M}^{-1} = -\mathbf{M'}^{-1}(\mathbf{M'}-\mathbf{M})\mathbf{M}^{-1}$ (this can be verified by simply expanding the right-hand side), we have

$$\begin{aligned}\boldsymbol{\alpha}' - \boldsymbol{\alpha} &= \big((\mathbf{K}' + \lambda\mathbf{I})^{-1} - (\mathbf{K} + \lambda\mathbf{I})^{-1}\big)\mathbf{y}\\ &= \big((\mathbf{K}' + \lambda\mathbf{I})^{-1}(\mathbf{K}' + \lambda\mathbf{I} - \mathbf{K} - \lambda\mathbf{I})(\mathbf{K} + \lambda\mathbf{I})^{-1}\big)\mathbf{y}\\ &= \big((\mathbf{K}' + \lambda\mathbf{I})^{-1}(\mathbf{K}' - \mathbf{K})(\mathbf{K} + \lambda\mathbf{I})^{-1}\big)\mathbf{y}\,.\end{aligned}$$

(b) Using the fact that for any vector $\mathbf{v}$ and matrix $\mathbf{A}$, $\|\mathbf{A}\mathbf{v}\|^2 = \mathbf{v}^\top\mathbf{A}^\top\mathbf{A}\mathbf{v} \le \|\mathbf{v}\|^2\|\mathbf{A}^\top\mathbf{A}\| = \|\mathbf{v}\|^2\|\mathbf{A}\|^2$ we have

$$\|\boldsymbol{\alpha}' - \boldsymbol{\alpha}\| \le \|(\mathbf{K}' + \lambda\mathbf{I})^{-1}\|\|\mathbf{K}' - \mathbf{K}\|\|(\mathbf{K} + \lambda\mathbf{I})^{-1}\|\|\mathbf{y}\|\,.$$

Since $|y| \le M$, we have $\|\mathbf{y}\| \le \sqrt{m}M$ and we can use the observation $\|(\mathbf{K} + \lambda\mathbf{I})^{-1}\| = \lambda_{\max}((\mathbf{K} + \lambda\mathbf{I})^{-1}) = \lambda_{\min}(\mathbf{K} + \lambda\mathbf{I})^{-1} \le 1/\lambda$, where $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ are the maximum and minimum eigenvalues of A, respectively.

11.5 Huber loss

The primal function can be written as:

$$\begin{aligned}\min_{\mathbf{w},b}\quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^m (L_c(\xi_i) + L_c(\xi_i'))\\ \text{s.t.}\quad & \mathbf{w}\cdot\Phi(\mathbf{x}_i) + b - y_i \le \xi_i,\ \forall i \in [m]\\ & y_i - \mathbf{w}\cdot\Phi(\mathbf{x}_i) - b \le \xi_i',\ \forall i \in [m]\\ & \xi_i, \xi_i' \ge 0,\ \forall i \in [m]\end{aligned}$$

The Lagrangian is written as follows,

$$\begin{aligned}L(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\xi}', \boldsymbol{\alpha}, \boldsymbol{\alpha}', \boldsymbol{\beta}, \boldsymbol{\beta}') &= \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^m (L_c(\xi_i) + L_c(\xi_i'))\\ &+ \sum_{i=1}^m \Big(\alpha_i(\mathbf{w}\cdot\Phi(\mathbf{x}_i) + b - y_i - \xi_i) + \alpha_i'(y_i - \mathbf{w}\cdot\Phi(\mathbf{x}_i) - b - \xi_i')\Big)\\ &- \sum_{i=1}^m (\beta_i\xi_i + \beta_i'\xi_i')\,,\end{aligned}$$

and the associated KKT conditions are:

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} + \sum_{i=1}^m (\alpha_i - \alpha_i')\Phi(\mathbf{x}_i) = 0\,,$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^m (\alpha_i - \alpha_i') = 0\,,$$

$$\frac{\partial L}{\partial \xi_i} = 0 \iff \begin{cases} C\xi_i = \alpha_i + \beta_i, & \text{if } \xi_i \le c\\ Cc = \alpha_i + \beta_i, & \text{otherwise}\end{cases}\,, \tag{E.134}$$

$$\frac{\partial L}{\partial \xi_i'} = 0 \iff \begin{cases} C\xi_i = \alpha_i' + \beta_i', & \text{if } \xi_i \le c\\ Cc = \alpha_i' + \beta_i', & \text{otherwise}\end{cases}\,, \tag{E.135}$$

$$\beta_i\xi_i = 0, \forall i \in [m]\,, \tag{E.136}$$

$$\beta_i'\xi_i' = 0, \forall i \in [m]\,. \tag{E.137}$$

The first two conditions are the same as in SVR the standard ϵ-insensitive loss, and using them to simplify the Lagrangian gives several familiar terms. The novel conditions involve ξ and ξ'. Collecting all terms in the Lagrangian that depend on ξ_i, we have the following equality (regardless of which condition holds in (E.134))

$$CL_c(\xi_i) - \xi_i(\alpha_i + \beta_i) = -\frac{(\alpha_i + \beta_i)^2}{2C}\,. \tag{E.138}$$

If it is the case that $\xi_i > 0$, then $\beta_i = 0$ by condition (E.136). In the case $\xi_i = 0$, then $\alpha_i + \beta_i = 0$ by the first case in condition (E.134). However, this also implies $\alpha_i = \beta_i = 0$, since both dual variables are constrained to be positive. Thus, in either case, we can simplify (E.138) to $\frac{-\alpha^2}{2C}$. Similar arguments can be used to simplify the ξ_i' terms, resulting in the final dual problem:

$$\begin{aligned}\max_{\boldsymbol{\alpha}} \quad & \mathbf{y}(\boldsymbol{\alpha}' - \boldsymbol{\alpha}) - \frac{1}{2}\Big((\boldsymbol{\alpha}' - \boldsymbol{\alpha})^\top \mathbf{K}(\boldsymbol{\alpha}' - \boldsymbol{\alpha}) + \frac{1}{C}\boldsymbol{\alpha}^\top \mathbf{1} + \frac{1}{C}\boldsymbol{\alpha}'^\top \mathbf{1}\Big) \\ \text{s.t.} \quad & (\boldsymbol{\alpha}' - \boldsymbol{\alpha})^\top \mathbf{1} = 0 \\ & 0 \leq \alpha_i, \alpha_i' \leq cC,\ \forall i \in [m]\,.\end{aligned}$$

11.8 Optimal kernel matrix

(a) Using the closed-form solution for the inner maximization problem $\boldsymbol{\alpha} = (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}$, simplifies the joint optimization to a simpler minimization:

$$\min_{\mathbf{K}\succeq 0} \mathbf{y}^\top(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y}\,, \quad \text{s.t.} \quad \|\mathbf{K}\|_2 \leq 1\,.$$

Note that for any invertible matrix $\mathbf{A}$, $\mathbf{y}^\top\mathbf{A}^{-1}\mathbf{y} \geq \|\mathbf{y}\|^2\lambda_{\min}(\mathbf{A}^{-1}) = \|\mathbf{y}\|^2\lambda_{\max}(\mathbf{A})^{-1}$. Thus, it is easy to see that $\min_{\mathbf{K}\succeq 0} \mathbf{y}^\top(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \geq \frac{\|y\|^2}{1+\lambda}$ since $\|\mathbf{K}\|_2 = \lambda_{\max}(\mathbf{K}) \leq 1$. We now show $\mathbf{K} = \frac{1}{\|\mathbf{y}\|^2}\mathbf{y}\mathbf{y}^\top$ achieves this lower bound. First, note that $(\frac{1}{\|\mathbf{y}\|^2}\mathbf{y}\mathbf{y}^\top + \lambda\mathbf{I})\mathbf{y} = (1+\lambda)\mathbf{y}$, so $\mathbf{y}$ is an eigenvector of the matrix with eigenvalue $(1+\lambda)$. Since the matrix is invertible, it can be shown that $\mathbf{y}$ is also an eigenvector of $(\frac{1}{\|\mathbf{y}\|^2}\mathbf{y}\mathbf{y}^\top + \lambda\mathbf{I})^{-1}$ with eigenvalue $\frac{1}{1+\lambda}$ (for example, consider the eigen decomposition of the matrix).

(b) The kernel matrix alone is not useful for classifying future unseen points x, which requires computing $\sum_{i=1}^m K(x_i, x)$ and needs access to an underlying kernel *function* that in consistent with the kernel matrix. Finding such a kernel function may be difficult in general, and furthermore the choice of function may not be unique.

11.9 Leave-one-out error

(a) Note that the hypothesis h_{S_i} will make zero error on the ith point of S_i' and is defined as the minimizer with respect to the remainder of the points. Thus, h_{S_i} is also the minimizer with respect to the set S_i'.

(b) Using part (a) and the definition of the KRR hypothesis with respect to the dual variables we have $h_{S_i}(x_i) = h_{S_i'}(x_i) = \boldsymbol{\alpha}_{S_i'}\mathbf{K}\epsilon_i$, where $\boldsymbol{\alpha}_{S_i'}$ is the optimal set of dual variable for KRR trained with S_i'. Noting that the closed-form solution is $\boldsymbol{\alpha} = (K + \lambda\mathbf{I})^{-1}\mathbf{y}_i$ proves the equality.

(c) Using part (b) we can write

$$\begin{aligned}h_{S_i}(x_i) - y_i &= \mathbf{y}_i^\top(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{K}\boldsymbol{e}_i - y_i \\ &= (\mathbf{y} - y_i\boldsymbol{e}_i + h_{S_i}(x_i)\boldsymbol{e}_i)^\top(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{K}\boldsymbol{e}_i - y_i \\ &= h_S(x_i) - y_i + (h_{S_i}(x_i) - y_i)\boldsymbol{e}_i^\top(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{K}\boldsymbol{e}_i\,,\end{aligned}$$

which implies $h_{S_i}(x_i) - y_i = (h_S(x_i) - y_i)/(\boldsymbol{e}_i^\top(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{K}\boldsymbol{e}_i)$. Thus, we can write

$$\widehat{R}_{\text{LOO}}(\mathcal{A}) = \frac{1}{m}\sum_{i=1}^m \Big(\frac{h_S(x_i) - y_i}{\boldsymbol{e}_i^\top(\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{K}\boldsymbol{e}_i}\Big)^2\,.$$

(d) In this case the two losses differ only by the factor $\frac{1}{\gamma^2}$. Thus, if $\gamma = \sqrt{m}$, the two performance measures coincide.