

Εισαγωγή στην Ταξινόμηση

Μηχανική Μάθηση

ΔΠΜΣ Επιστήμης Δεδομένων & Μηχανικής Μάθησης

Γιώργος Αλεξανδρίδης – gealexan@mail.ntua.gr

Κατηγορίες ταξινομητών

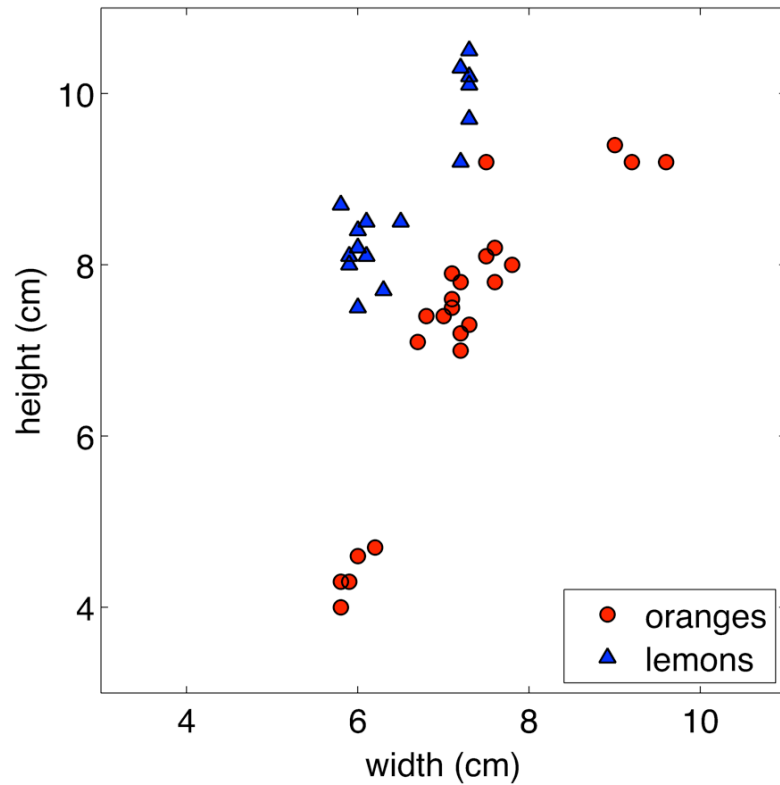
1. Διαμεριστικοί (divisive)

- Προσπαθούν να *τεμαχίσουν* το χώρο των δεδομένων σε *μη-επικαλυπτόμενες* υπο-περιοχές εντός των οποίων βρίσκονται δεδομένα μιας κλάσης
- Ταξινομητές πλησιέστερων γειτόνων, δέντρα αποφάσεων, (πολυεπίπεδα) νευρωνικά δίκτυα πρόσθιας τροφοδότησης, μηχανές διανυσμάτων υποστήριξης, ...

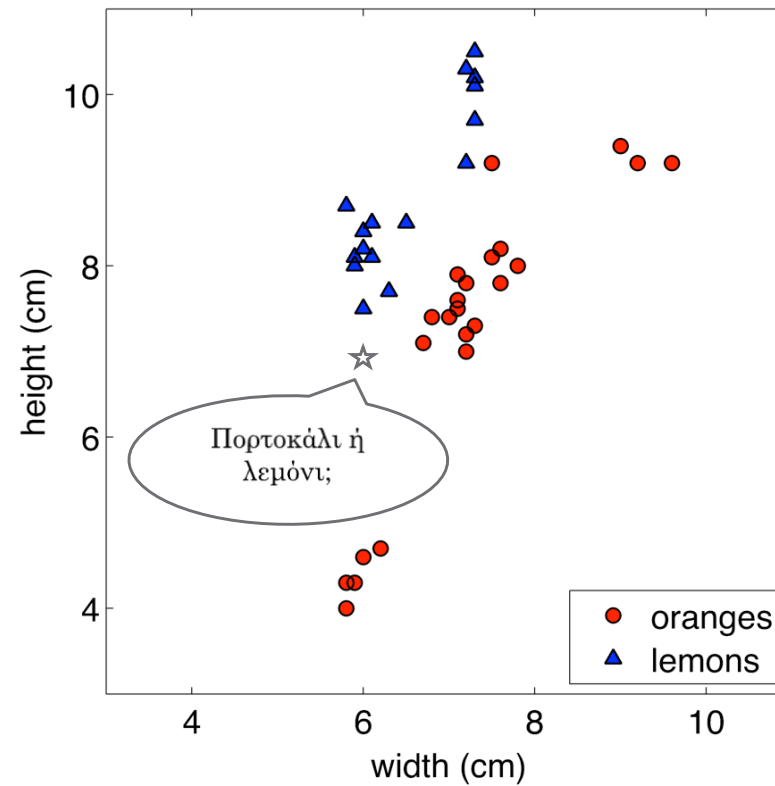
2. Παραγωγικοί (generative)

- Στατιστική θεώρηση των δεδομένων
- Προσπαθούν να μάθουν την *υποκείμενη κατανομή* (underlying distribution) που παράγει τα δεδομένα
- Αφελείς μπεϋζιανοί ταξινομητές, γκαουσιανά μοντέλα μίξης, κρυφά μαρκοβιανά μοντέλα,...

Πρόβλημα ταξινόμησης: Πορτοκάλια ή Λεμόνια;



Ταξινόμηση και Επαγωγή (Induction)



Εκμάθηση Μέσω Παραδειγμάτων

Instance-based Learning

Εκμάθηση μέσω Παραδειγμάτων

- **Μη-παραμετρικά μοντέλα** (*non-parametric models*)
- Πρόκειται για απλά μοντέλα τα οποία χρησιμοποιούνται για την προσέγγιση προβλημάτων συνεχών τιμών (παλινδρόμηση) ή διακριτών τιμών (ταξινόμηση)
 - Ταξινομητές ***k*-πλησιέστερων γειτόνων** (*k-nearest neighbors classifier – kNN*), **δίκτυα ακτινικών συναρτήσεων βάσης** (*radial basis function – RBF*), ..
- Η διαδικασία της μάθησης είναι ισοδύναμη με την αποθήκευση των **δεδομένων** που χρησιμοποιούνται για την **εκπαίδευση** του μοντέλου (*training data*)
- Τα νέα στιγμιότυπα ταξινομούνται χρησιμοποιώντας «ομοειδή» στιγμιότυπα από το σύνολο εκπαίδευσης
 - Τα νέα στιγμιότυπα αναφέρονται επίσης και ως **στιγμιότυπα δοκιμής** (*test instances*) ή **δεδομένα δοκιμής** (*test data*)
- Κωδικοποιούν λογικές **υποκείμενες υποθέσεις** (*underlying assumptions*)
 - Οι κλάσεις (έξοδος) μεταβάλλονται «ομαλά» (*smooth*) συναρτήσει της εισόδου
 - Τα δεδομένα καταλαμβάνουν έναν υπο-χώρο του αρχικού χώρου μεγάλων διαστάσεων

Ταξινομητής Πλησιέστερου Γείτονα

- Έστω ότι το σύνολο T των δειγμάτων δεδομένων εκπαίδευσης ανήκουν στον Ευκλείδειο χώρο ($x \in \mathbb{R}^d$)

- **Λειτουργία**

- Η ετικέτα που λαμβάνει κάθε νέο δείγμα είναι η ετικέτα του πλησιέστερου δείγματος δεδομένων εκπαίδευσης
- Συνήθως χρησιμοποιείται η Ευκλείδεια απόσταση

$$\|x^{(a)} - x^{(b)}\|_2 = \sqrt{\sum_{j=1}^d (x_j^{(a)} - x_j^{(b)})^2}$$

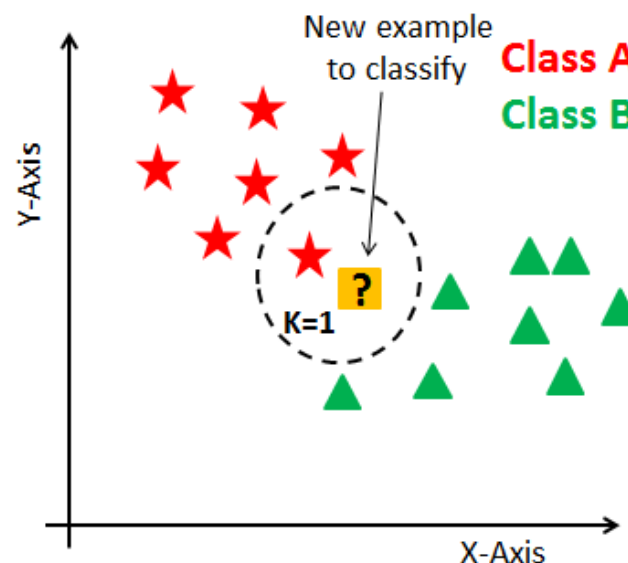
- **Αλγόριθμος**

1. Βρες το παράδειγμα $(x^*; t^*)$ από τα δεδομένα εκπαίδευσης το οποίο είναι «εγγύτερα» στο νέο στιγμιότυπο x

$$x^* = \underset{x^{(i)} \in T}{\operatorname{argmin}} \operatorname{distance}(x^{(i)}, x)$$

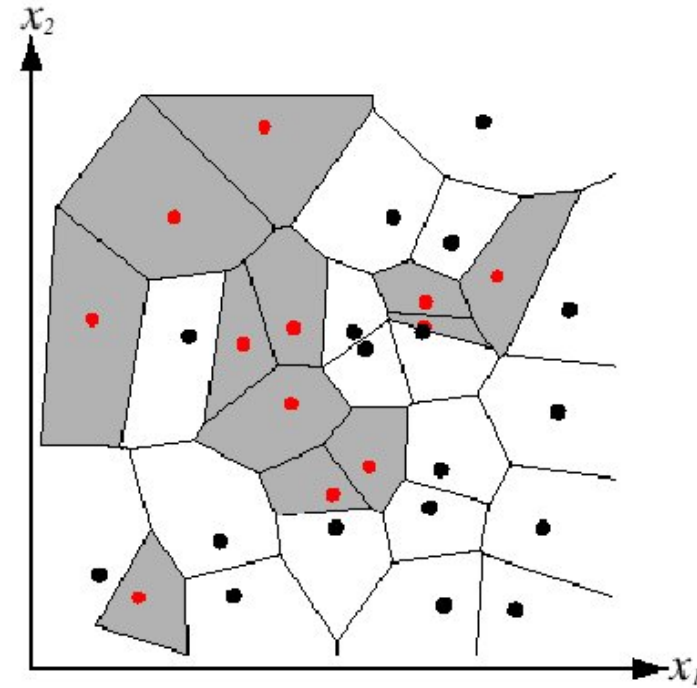
1. Ανάθεσε στο x την ετικέτα t^* : $(x; t^*)$

- Στην πραγματικότητα, δεν χρειάζεται να υπολογίσουμε την ρίζα
 - Γιατί;

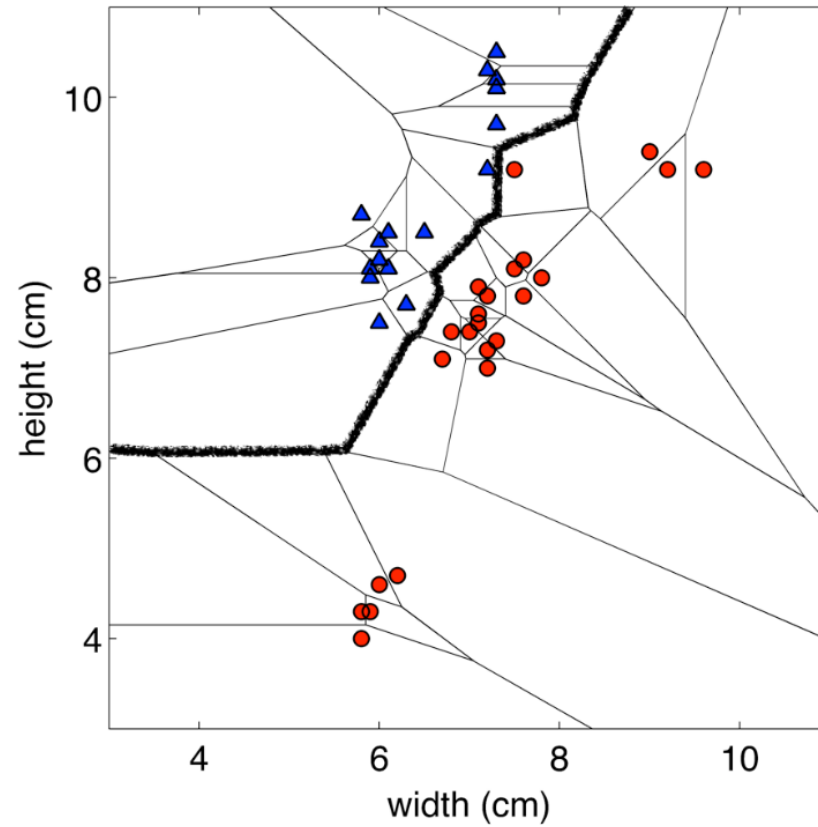


Όρια απόφασης

- Ο ταξινομητής πλησιέστερου γείτονα δεν υπολογίζει άμεσα όρια απόφασης μεταξύ των κλάσεων
 - Ωστόσο, αυτά προκύπτουν έμμεσα
- Όρια απόφασης: Διαγράμματα Voronoi
 - Ο χώρος των δεδομένων χωρίζεται σε διαφορετικές περιοχές ανάλογα με την ετικέτα (κλάση) τους
 - Τα ευθύγραμμα τμήματα αποτελούν τις μεσοκαθέτους μεταξύ δύο «γειτονικών» στιγμιοτύπων των δεδομένων εκπαίδευσης

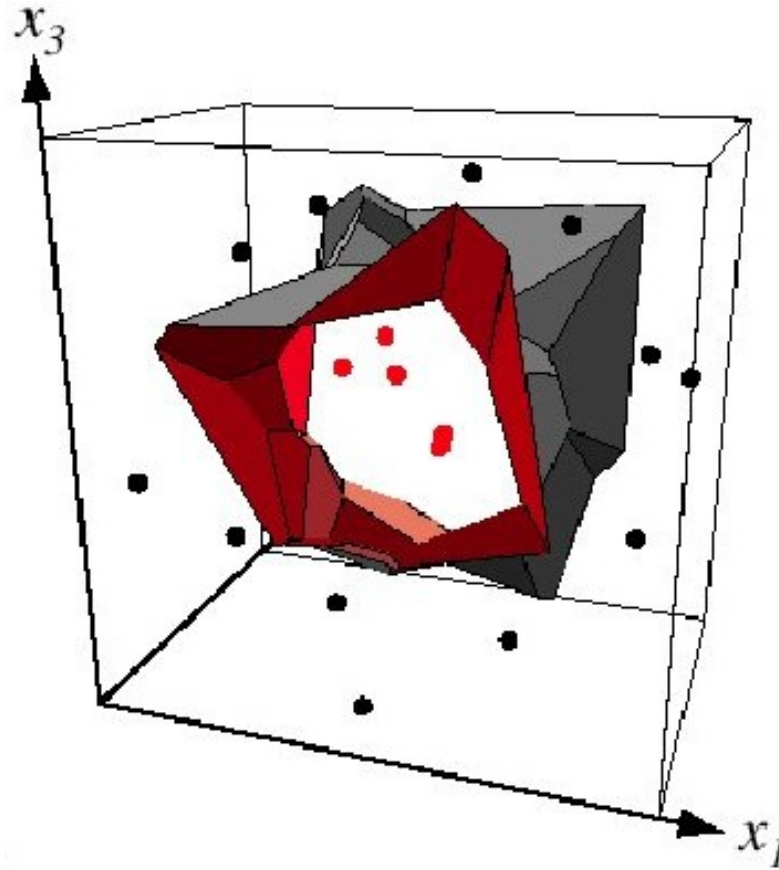


Πορτοκάλια ή Λεμόνια: Όριο απόφασης



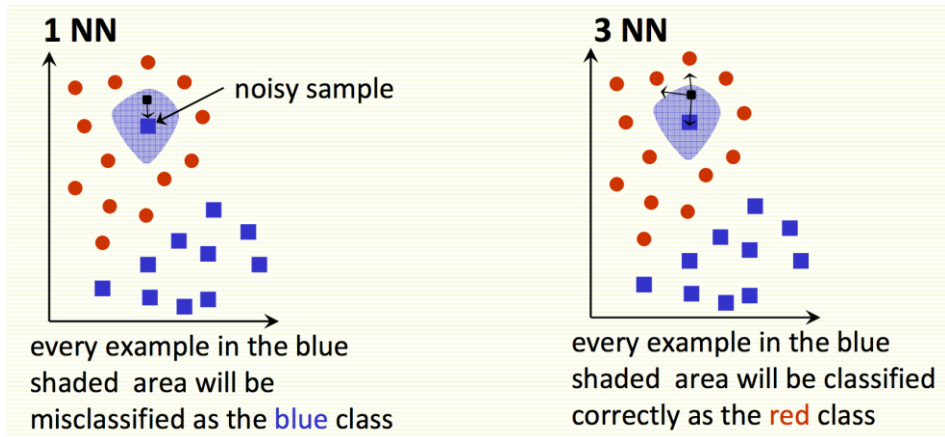
Σύνθετο όριο απόφασης (όχι ευθύγραμμο τμήμα)

Όριο απόφασης στις 3 διαστάσεις



k -πλησιέστεροι γείτονες

- Ο ταξινομητής πλησιέστερου γείτονα είναι «ευαίσθητος» στην ύπαρξη θορύβου στα δεδομένα
- **Λύση**
 - «Ομαλοποίηση» της ταξινόμησης μέσω «ψηφοφορίας» των k πλησιέστερων γειτόνων
- **Αλγόριθμος**
 1. Βρες τα k πλησιέστερα παράδειγμα $(\mathbf{x}^{(i)}; t^{(i)})$ από τα δεδομένα εκπαίδευσης τα οποία είναι «εγγύτερα» στο νέο στιγμιότυπο $\mathbf{x}$
 2. Ανάθεσε στο $\mathbf{x}$ την ετικέτα y :
 - $y = \underset{t^{(z)}}{\operatorname{argmax}} \sum_{r=1}^k \delta(t^{(z)}, t^{(r)})$



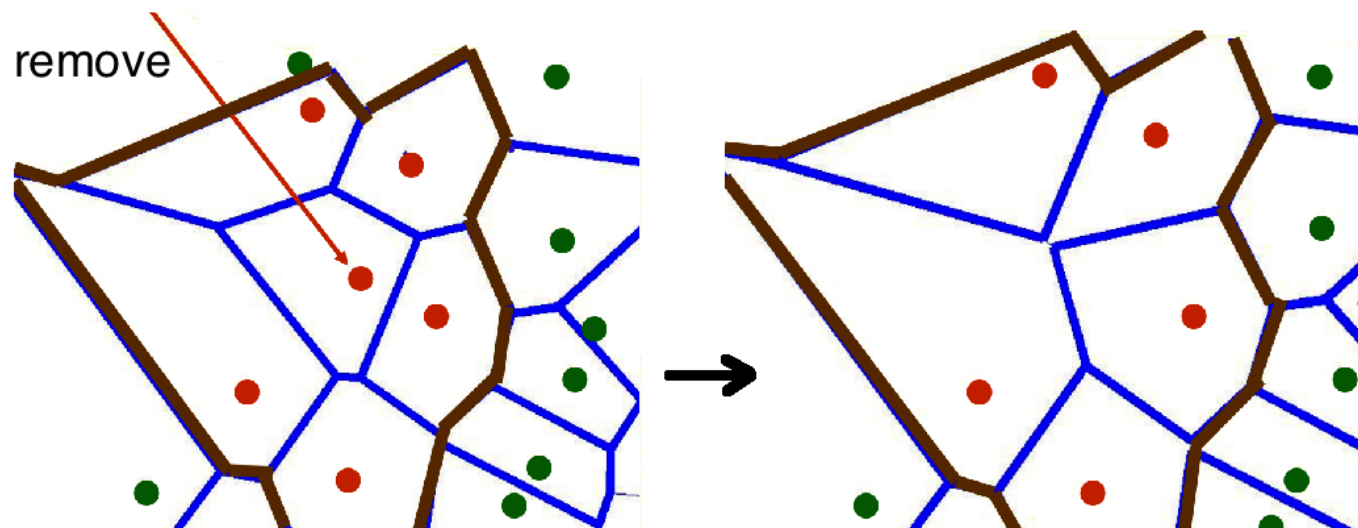
Προσδιορισμός k

- **Πως** βρίσκουμε την **κατάλληλη** τιμή για το k ;
 - Όσο μεγαλύτερο είναι, τόσο βελτιώνεται η απόδοση του ταξινομητή
 - Ωστόσο αν είναι αρκετά μεγάλο, τότε το εύρος της «γειτονιάς» μεγαλώνει, συμπεριλαμβάνοντας δείγματα τα οποία μπορεί να βρίσκονται πολύ μακριά από το νέο στιγμιότυπο
- Εύρεση k μέσω τεχνικών **διασταυρούμενης επικύρωσης** (*cross-validation*)
 - Θα μιλήσουμε σε επόμενες διαλέξεις και στο εργαστήριο για αυτές
- Εμπειρικός κανόνας
 - $k < \sqrt{N}$, όπου N το πλήθος των παραδειγμάτων εκπαίδευσης

k -πλησιέστεροι γείτονες: ζητήματα πολυπλοκότητας

- **Υψηλή πολυπλοκότητα** κατά τη διαδικασία ελέγχου
 - Για να βρούμε έναν πλησιέστερο γείτονα πρέπει να υπολογίσουμε την απόσταση από όλα τα δεδομένα εκπαίδευσης
 - Λύσεις
 - Χρησιμοποίηση υποσυνόλου των διαστάσεων, χρήση αποδοτικών δομών δεδομένων (πχ kd-trees), υπολογισμός προσεγγιστικής απόστασης, αφαίρεση πλεονασματικών δεδομένων, ...
- **Υψηλές απαιτήσεις αποθήκευσης**
 - Πρέπει να αποθηκεύσουμε στη μνήμη όλα τα δεδομένα εκπαίδευσης
 - Λύσεις
 - Αφαίρεση πλεονασματικών δεδομένων
- **Δεδομένα πολλών διαστάσεων** («Κατάρα διαστατικότητας»)
 - Το πλήθος των απαιτούμενων δεδομένων εκπαίδευσης αυξάνει εκθετικά όσο αυξάνουν οι διαστάσεις
 - Επίσης αυξάνει και το υπολογιστικό κόστος
 - Λύσεις
 - Εφαρμογή τεχνικών μείωσης διαστάσεων και επιλογής χαρακτηριστικών

Αφαίρεση πλεονασματικών δεδομένων



Αν όλοι οι γείτονες έχουν την ίδια κλάση, μπορούμε να αφαιρέσουμε το δείγμα δεδομένων

Εφαρμογή Ταξινόμησης Πλησιέστερων Γειτόνων

- Σε ποιο μέρος τραβήχτηκε η ακόλουθη φωτογραφία;
 - James Hays, Alexei A. Efros. im2gps: estimating geographic information from a single image. CVPR'08. Project page: <http://graphics.cs.cmu.edu/projects/im2gps/>



Που τραβήχθηκε η ακόλουθη φωτογραφία;

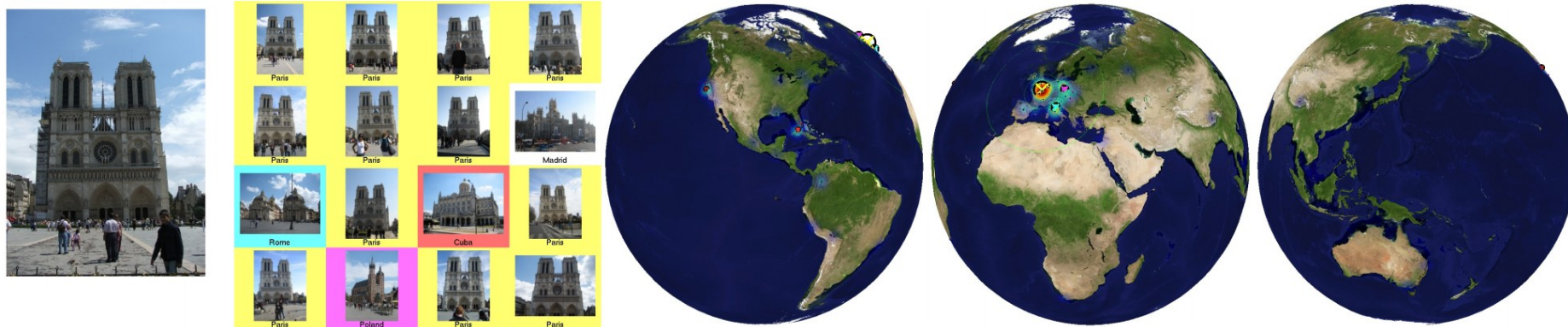
- **Δεδομένα εκπαίδευσης**

- 6 εκ. εικόνες από το Flickr που περιέχουν μεταδεδομένα τοποθεσίας
- Αρκετά πυκνή δειγματοληψία σε όλη την υφήλιο

- Αναπαράσταση κάθε φωτογραφίας με συγκεκριμένα, περιγραφικά χαρακτηριστικά

- Πρόβλεψη τοποθεσίας για νέες φωτογραφίες μέσω ταξινόμησης k -πλησιέστερων γειτόνων

- Οι ερευνητές βρήκαν ότι το βέλτιστο k ήταν 120



k -πλησιέστεροι γείτονες: Συμπεράσματα

- Σχηματίζουν περίπλοκα όρια απόφασης, τα οποία προσαρμόζονται στην πυκνότητα των δεδομένων εκπαίδευσης
- Σε περιπτώσεις που τα δεδομένα εκπαίδευσης είναι πολλά, η μέθοδος των k -πλησιέστερων γειτόνων λειτουργεί ικανοποιητικά
- **Ζητήματα**
 1. Ευαισθησία στο θόρυβο
 2. Ευαισθησία στο εύρος τιμών των χαρακτηριστικών των δεδομένων
 3. Η έννοια της απόστασης μεταξύ στιγμιοτύπων δεδομένων χάνει τη σημασία της όσο οι διαστάσεις μεγαλώνουν
 4. Γραμμική υπολογιστική πολυπλοκότητα συναρτήσει του πλήθους των δεδομένων εκπαίδευσης (ψευδο-πολυωνυμικός αλγόριθμος)

Αφελής Μπεϋζιανός Ταξινομητής

Naïve Bayesian Classifier

Αφελείς Μπεϋζιανοί Ταξινομητές

- **Naïve Bayesian Classifiers (NBC)**

- Οικογένεια *πιθανοτικών ταξινομητών* οι οποίοι βασίζονται στο **Θεώρημα του Bayes**
- Υποθέτουν ισχυρή ανεξαρτησία μεταξύ των χαρακτηριστικών των δεδομένων
 - Εξου και ο χαρακτηρισμός *αφελείς* (naïve)

- **Thomas Bayes (1701-1761)**

- Βρετανός στατιστικολόγος, φιλόσοφος και ιερωμένος της Πρεσβυτεριανής Εκκλησίας

- **Θεώρημα Bayes ή Κανόνας Bayes**

- Έστω $A_1, A_2, \dots, A_n$ **αμοιβαίως αποκλειόμενα** (ανεξάρτητα) ενδεχόμενα που καθορίζουν δειγματικό χώρο S . Έστω B ένα οποιοδήποτε ενδεχόμενο του χώρου, τέτοιο ώστε $P(B) > 0$. Τότε

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^n P(B|A_j)P(A_j)}$$



Θεώρημα Bayes: Απλή μορφή

- $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$, όπου A, B ενδεχόμενα με $P(B) > 0$
- Το θεώρημα του Bayes
 - επιτρέπει την *ενημέρωση* της πιθανότητας εμφάνισης ενός ενδεχομένου, **ενσωματώνοντας νέα πληροφορία**
 - Ενσωματώνει τις *εκ των προτέρων πιθανότητες* (prior probabilities) για να δημιουργήσει εκ των *υστέρων πιθανότητες* (posterior probabilities)
- $P(A|B)$: *εκ των υστέρων πιθανότητα* εμφάνισης ενδεχομένου A , δεδομένου ότι το ενδεχόμενο B έχει συμβεί
- $P(A)$: *εκ των προτέρων πιθανότητα* εμφάνισης ενδεχομένου A , πριν την πραγματοποίηση νέας παρατήρησης
- $P(B|A)$: *πιθανοφάνεια* (likelihood) ενδεχομένου A
 - πιθανότητα να συμβεί το B ενώ το A έχει ήδη συμβεί
- $P(B)$: πιθανότητα εμφάνισης ενδεχομένου B
 - Καλείται και «*απόδειξη*» (evidence)

Θεώρημα Bayes: Παράδειγμα

Έστω ότι έχουμε κατάστημα ηλεκτρολογικού υλικού και προμηθευόμαστε λαμπτήρες από *τρεις* κατασκευαστές: τον **A**, τον **B** και τον **C**. Πιο συγκεκριμένα ο **A** μας προμηθεύει το *80%* των λαμπτήρων που πουλάμε, ο **B** το *15%* και ο **C** το υπόλοιπο *5%*. Επίσης, οι κατασκευαστές μας έχουν ενημερώσει ο **μεν A** ότι το *4%* των λαμπτήρων του είναι ελλαττωματικό, ο **B** το *6%* και ο **C** το *9%*. Δεδομένου ότι ένας πελάτης μας επιστρέφει **έναν λαμπτήρα** πίσω ως **ελλαττωματικό**, **ποια είναι η πιθανότητα να έχει κατασκευαστεί από τον A;**

Λύση

$$P(A) = 0,8, P(B) = 0,15, P(C) = 0,05$$

$$P(E|A) = 0,04, P(E|B) = 0,06, P(E|C) = 0,09$$

Εφαρμογή θεωρήματος Bayes

$$P(A|E) = \frac{P(E|A)P(A)}{P(E|A)P(A) + P(E|B)P(B) + P(E|C)P(C)} = \frac{0,04 \cdot 0,8}{0,04 \cdot 0,8 + 0,06 \cdot 0,15 + 0,09 \cdot 0,05} \approx 0,7033$$

Άρα η πιθανότητα ο ελλαττωματικός λαμπτήρας να έχει κατασκευαστεί από τον A είναι **περίπου 70,33%**

Αφελής Μπεϋζιανός Ταξινομητής: Μοντέλο

- **Αφελής υπόθεση** (naïve assumption)

- Όλα τα n χαρακτηριστικά του νέου δείγματος δεδομένων $\mathbf{x}$ είναι ανεξάρτητα μεταξύ τους
- $\mathbf{x} = \{x_1, x_2, \dots, x_n\}$
- $P(x_i | x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n, C_k) = P(x_i | C_k)$, C_k : Κλάση δείγματος
- $P(\mathbf{x} | C_k) = \prod_{i=1}^n P(x_i | C_k)$

- **Εφαρμογή Θεωρήματος Bayes**

- $P(C_k | \mathbf{x}) = \frac{P(C_k)P(\mathbf{x} | C_k)}{P(\mathbf{x})} \propto P(C_k)P(\mathbf{x} | C_k) \Rightarrow P(C_k | \mathbf{x}) \propto P(C_k) \prod_{i=1}^n P(x_i | C_k)$
- $P(\mathbf{x})$: Πιθανότητα εμφάνισης συγκεκριμένου δείγματος δεδομένων, σταθερά
 - «Απόδειξη» (evidence)

- **Ταξινόμηση** νέου δείγματος δεδομένων $\mathbf{x}$

- $\hat{y} = \operatorname{argmax}_{k \in 1, \dots, K} P(C_k) \prod_{i=1}^n P(x_i | C_k)$

Αφελής Μπεϋζιανός Ταξινομητής: Εκπαίδευση Μοντέλου

- **Ταξινομητής:** $\hat{y} = \underset{k \in 1, \dots, K}{\operatorname{argmax}} P(C_k) \prod_{i=1}^n P(x_i | C_k)$
- **Δεδομένα εκπαίδευσης** $\{X, y\}$
 - m στιγμιότυπα (instances) n χαρακτηριστικών το κάθε ένα, μαζί με την αντίστοιχη ετικέτα τους
- **Παράμετροι μοντέλου**
 1. $P(C_k)$: Υπολογισμός κατανομής κλάσεων στο σύνολο δεδομένων εκπαίδευσης
 2. $P(x_i | C_k)$: Υπολογισμός πιθανότητας εμφάνισης κάθε χαρακτηριστικού σε κάθε κλάση, στο σύνολο δεδομένων εκπαίδευσης
- Με την προσθήκη νέων δεδομένων εκπαίδευσης, οι τιμές των παραμέτρων του μοντέλου *ενημερώνονται*

Αφελής Μπεϋζιανός Ταξινομητής: Παράδειγμα

Παρατηρήσεις βροχόπτωσης

Θερμοκρασία	Υγρασία	Βροχή
Κρύο	Υψηλή	Ναι
Κρύο	Χαμηλή	Όχι
Μέση	Χαμηλή	Ναι
Μέση	Μέτρια	Όχι
Ζέστη	Μέτρια	Όχι
Ζέστη	Υψηλή	Όχι

Αν σήμερα η *Θερμοκρασία* είναι **Μέση**
και η *Υγρασία* **Υψηλή**, θα βρέξει ή όχι;

Λύση

- $P(B) = \frac{2}{6}, P(OB) = \frac{4}{6}$
- $P(M|B) = \frac{1}{2}, P(M|OB) = \frac{1}{4}$
- $P(Y|B) = \frac{1}{2}, P(Y|OB) = \frac{1}{4}$
- $P(B|M, Y) \propto P(B) P(M|B)P(Y|B) = \frac{2}{6} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{12}$
- $P(OB|M, Y) \propto P(OB) P(M|OB)P(Y|OB) = \frac{4}{6} \cdot \frac{1}{4} \cdot \frac{1}{4} = \frac{1}{24}$
- Επειδή $P(B|M, Y) > P(OB|M, Y)$ ο αφελής μπεϋζιανός ταξινομητής προβλέπει ότι θα βρέξει σήμερα

Συμπεράσματα

- **Περιορισμοί**

- Ελλιπή δεδομένα εκπαίδευσης
 - Αν η τιμή κάποιου χαρακτηριστικού *δεν υπάρχει καθόλου* στα δεδομένα εκπαίδευσης, τότε *δεν μπορεί* να προσδιοριστεί σε ποια κλάση ανήκει το δείγμα.
- Συνεχή χαρακτηριστικά
 - Πρέπει να βρεθεί η *υποκείμενη κατανομή* τους (πχ κανονική, Laplace, Poisson)
- Μη-ανεξαρτησία μεταξύ των χαρακτηριστικών των δεδομένων
 - Εφαρμογή *τεχνικών απεικόνισης* των χαρακτηριστικών σε νέο χώρο, όπου η ανεξαρτησία τους είναι διασφαλισμένη ως ένα βαθμό (π.χ. ανάλυση κυρίων συνιστωσών)

- **Εφαρμογές**

- Φίλτρα ενοχλητικής αλληλογραφίας
- Εκτιμητές ρίσκου για χορήγηση δανείων/έκδοση πιστωτικών καρτών
- ...

Βιβλιογραφία

1. Ταξινομητές Πλησιέστερων Γειτόνων (Εκμάθηση μέσω Παραδειγμάτων)
 - M. Kubat – Εισαγωγή στη Μηχανική Μάθηση
 - Κεφάλαιο 3
2. Αφελείς Μπεϋζιανοί Ταξινομητές
 - M. Kubat – Εισαγωγή στη Μηχανική Μάθηση
 - Κεφάλαιο 2 – Ενότητες 2.1-2.4