Κατηγοριοποίηση ΙΙΙ

Κατηγοριοποιητές Κανόνων

Κατηγοριοποίηση με Κανόνες

Κατηγοριοποίηση των εγγραφών με βάση ένα σύνολο από κανόνες της μορφής "if…then…"

Κανόνας: (*Συνθήκη*) → *y*

όπου

Συνθήκη (Condition) είναι σύζευξη συνθηκών στα γνωρίσματα

y η ετικέτα της κλάσης

LHS: rule antecedent (πρότερο) ή condition (συνθήκη)
RHS: rule consequent (επακόλουθο ή απότοκο)

Παραδείγματα κανόνων κατηγοριοποίησης:

(Blood Type=Warm) ∧ (Lay Eggs=Yes) → Birds

(Taxable Income < 50K) ∧ (Refund=Yes) → Cheat=No

Κατηγοριοποίηση με Κανόνες

Παράδειγμα

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
human	warm	yes	no	no	mammals
python	cold	no	no	no	reptiles
salmon	cold	no	no	yes	fishes
whale	warm	yes	no	yes	mammals
frog	cold	no	no	sometimes	amphibians
komodo	cold	no	no	no	reptiles
bat	warm	yes	yes	no	mammals
pigeon	warm	no	yes	no	birds
cat	warm	yes	no	no	mammals
leopard shark	cold	yes	no	yes	fishes
turtle	cold	no	no	sometimes	reptiles
penguin	warm	no	no	sometimes	birds
porcupine	warm	yes	no	no	mammals
eel	cold	no	no	yes	fishes
salamander	cold	no	no	sometimes	amphibians
gila monster	cold	no	no	no	reptiles
platypus	warm	no	no	no	mammals
owl	warm	no	yes	no	birds
dolphin	warm	yes	no	yes	mammals
eagle	warm	no	yes	no	birds

Μοντέλο – Σύνολο Κανόνων (Rule Set)

R1: (Give Birth = no) ∧ (Can Fly = yes) → Birds
R2: (Give Birth = no) ∧ (Live in Water = yes) → Fishes
R3: (Give Birth = yes) ∧ (Blood Type = warm) → Mammals
R4: (Give Birth = no) ∧ (Can Fly = no) → Reptiles
R5: (Live in Water = sometimes) → Amphibians

Κατηγοριοποίηση με Κανόνες

Εφαρμογή Κατηγοριοποιητών με Κανόνες

Ένας κανόνας r καλύπτει (covers) ένα στιγμιότυπο (εγγραφή) αν τα γνωρίσματα του στιγμιότυπου ικανοποιούν τη συνθήκη του κανόνα

R1: (Give Birth = no) ∧ (Can Fly = yes) → Birds
R2: (Give Birth = no) ∧ (Live in Water = yes) → Fishes
R3: (Give Birth = yes) ∧ (Blood Type = warm) → Mammals
R4: (Give Birth = no) ∧ (Can Fly = no) → Reptiles
R5: (Live in Water = sometimes) → Amphibians

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
hawk	warm	no	yes	no	?
grizzly bear	warm	yes	no	no	?

Ο κανόνας R1 καλύπτει το hawk (ή αλλιώς το hawk ενεργοποιεί /πυροδοτεί (trigger) τον κανόνα) => Bird

Ο κανόνας R3 καλύπτει το grizzly bear => Mammals

Κατηγοριοποίηση με Κανόνες

Κάλυψη Κανόνα - Coverage:
Το ποσοστό των εγγραφών που ικανοποιούν το LHS του κανόνα

Ακρίβεια Κανόνα - Accuracy:
Το ποσοστό των εγγραφών που καλύπτουν και το LHS και το RHS του κανόνα

Tid	Refund	Marital Status	Taxable Income	Class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

(Status=Single) → No

Coverage = 40%, Accuracy = 50%

Κατηγοριοποίηση με Κανόνες: Εφαρμογή

R1: (Give Birth = no) ∧ (Can Fly = yes) → Birds
R2: (Give Birth = no) ∧ (Live in Water = yes) → Fishes
R3: (Give Birth = yes) ∧ (Blood Type = warm) → Mammals
R4: (Give Birth = no) ∧ (Can Fly = no) → Reptiles
R5: (Live in Water = sometimes) → Amphibians

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
lemur	warm	yes	no	no	?
turtle	cold	no	no	sometimes	?

Name	Blood Type	Give Birth	Can Fly	Live in Water	Class
dogfish	cold	yes	no	yes	?
grizzly bear	warm	yes	no	no	?

- Lemur R1
- Turtle R4, R5
- Godfish -

Κατηγοριοποίηση με Κανόνες: Εφαρμογή

Ιδιότητες Κατηγοριοποιητών Κανόνων

- Αμοιβαία αποκλειόμενοι κανόνες (Mutually exclusive rules)

  Ένας κατηγοριοποιητής περιέχει αμοιβαία αποκλειόμενους κανόνες, αν οι κανόνες είναι ανεξάρτητοι ο ένας από τον άλλο

  Κάθε εγγραφή καλύπτεται από το *πολύ έναν* κανόνα – δεν υπάρχουν δυο κανόνες που να πυροδοτούνται από την ίδια εγγραφή

- Πλήρεις κανόνες (Exhaustive rules)

  Ένας κατηγοριοποιητής έχει πλήρη κάλυψη (coverage) αν καλύπτει όλους τους πιθανούς συνδυασμούς τιμών γνωρισμάτων – υπάρχει ένας κανόνας για κάθε συνδυασμό τιμών γνωρισμάτων

  Κάθε εγγραφή καλύπτεται από *τουλάχιστον έναν* κανόνα

Κατηγοριοποίηση με Κανόνες: Εφαρμογή

- Αν οι κανόνες δεν είναι πια αμοιβαία αποκλειόμενοι
  Μια εγγραφή μπορεί να ενεργοποιήσει παραπάνω από έναν κανόνα
  Λύση (conflict resolution)
  - (1) Διάταξη του συνόλου κανόνων: αν μια εγγραφή ενεργοποιεί πολλούς κανόνες, της ανατίθεται αυτός με τη μεγαλύτερη προτεραιότητα
    - Ένα διατεταγμένο σύνολο κανόνων λέγεται και Λίστα Απόφασης (Decision list)
    - Οι κανόνες εξετάζονται με τη σειρά
    - Η διάταξη με βάση κάποιο κριτήριο, πχ. (α) με τη σειρά που παράγονται, (β) κάλυψη ή/και ακρίβεια, (γ) με το αριθμό όρων (size ordering)
  - (2) διάταξη των κλάσεων: αν μια εγγραφή ενεργοποιεί πολλούς κανόνες, της ανατίθεται η κλάση με τη μεγαλύτερη προτεραιότητα
  - (3) Χωρίς διάταξη του συνόλου κανόνων – χρήση σχήματος ψηφοφορίας, η πλειοψηφούσα κλάση + στάθμιση ψήφου με βάση την ακρίβεια του κανόνα (misclassification cost)
- Οι κανόνες δεν είναι εξαντλητικοί
  Μια εγγραφή μπορεί να μην ενεργοποιεί κάποιον κανόνα
  Χρήση default κλάσης με άδεια LHS

Κατηγοριοποίηση με Κανόνες: Κατασκευή

Κατασκευή Κατηγοριοποιητών με Κανόνες

- Άμεση Μέθοδος:
  - Εξαγωγή κανόνων απευθείας από τα δεδομένα
  - Π.χ.: RIPPER, CN2, Holte's 1R
- Έμμεση Μέθοδος:
  - Εξαγωγή κανόνων από άλλα μοντέλα κατηγοριοποιητών (πχ από δέντρα απόφασης)
  - Π.χ.: C4.5 κανόνες

Κατηγοριοποίηση με Κανόνες: Κατασκευή

Άμεσοι Μέθοδοι Κατασκευή Κατηγοριοποιητών με Κανόνες

Σειριακή Κάλυψη (sequential covering)

1. Ξεκίνα με ένα άδειο κανόνα
2. Για κάθε κατηγορία ξεχωριστά

   Repeat
   - Grow a rule using the Learn-One-Rule function
   - Remove training records covered by the rule
   - Add rule to the set

   until stopping condition

Κατηγοριοποίηση με Κανόνες: Κατασκευή

Σειριακή Κάλυψη: Παράδειγμα

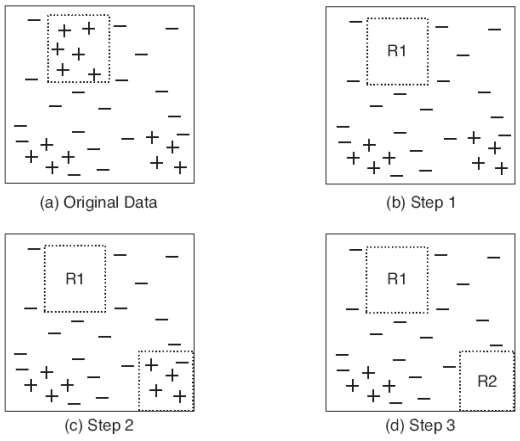


Figure 5.2. An example of the sequential covering algorithm.

- Διάταξη των κλάσεων
- Από τα περισσότερα δείγματα

Κατηγοριοποίηση με Κανόνες: Κατασκευή

Σειριακή Κάλυψη: Rule-Growing Strategy

Γενική σε ειδική

Ξεκίνα από τον κανόνα

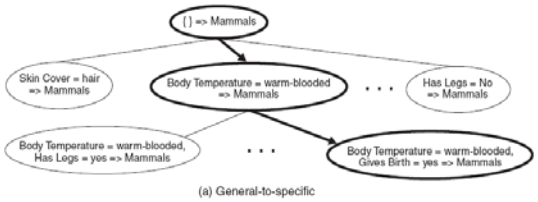
R: {} → y

Διαδοχικά πρόσθεσε νέες συζεύξεις για να βελτιωθεί η ποιότητα του κανόνα

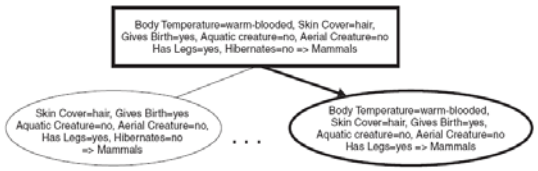
Ειδική σε γενική

Επέλεξε ένα από τα δείγματα τυχαία

Γενίκευσε αφαιρώντας συζεύξεις



(a) General-to-specific



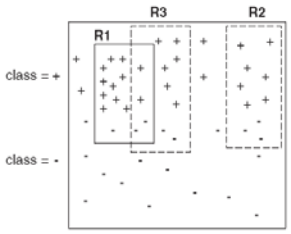
(b) Specific-to-general

Figure 5.3. General-to-specific and specific-to-general rule-growing strategies.

Κατηγοριοποίηση με Κανόνες: Κατασκευή

Σειριακή Κάλυψη: Remove Training Records

- Γιατί σβήνουμε τα θετικά δείγματα?
  - Ώστε ο επόμενος κανόνας να είναι διαφορετικός
- Σβήνουμε ή όχι και τα αρνητικά δείγματα?



Αρχικά

Ο R1 με ακρίβεια 12/15 80%, R2 7/10 – 20%, R3 8/12 70%

Επιλογή R1, σβήνουμε τα θετικά δείγματα (καλύπτονται)

τα αρνητικά?

Στη συνέχεια των R2 ή των R3?

Κατηγοριοποίηση με Κανόνες: Κατασκευή

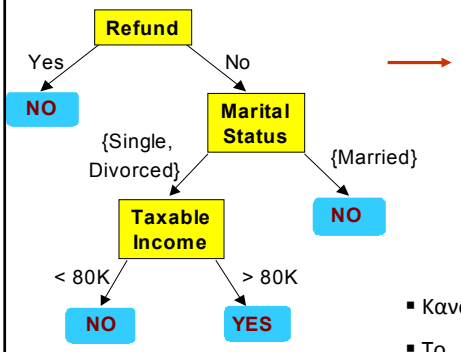
Σειριακή Κάλυψη

- Κριτήριο τερματισμού
  - Με βάση το κέρδος
  - Αν μικρό αγνόησε το κέρδος
- Κλάδεμα κανόνων (όπως και στα δέντρα απόφασης)
  Για παράδειγμα
  - Σβήσε μια από τις συζεύξεις
  - Σύγκρινε το ρυθμό σφάλματος με τα δεδομένα ελέγχου
    - Αν βελτιώνεται, σβήνεται

Κατηγοριοποίηση με Κανόνες: Κατασκευή

Έμμεση Μέθοδος: Από Δέντρα Απόφασης σε Κανόνες

Ένας κανόνας για κάθε μονοπάτι από τη ρίζα σε φύλλο
Κάθε ζευγάρι γνώρισμα-τιμή στο μονοπάτι αποτελεί ένα όρο στη σύζευξη και το φύλλο αφορά την κλάση (RHS)



Κανόνες Κατηγοριοποίησης (Classification Rules)

(Refund=Yes) ==> No

(Refund=No, Marital Status={Single,Divorced}, Taxable Income<80K) ==> No

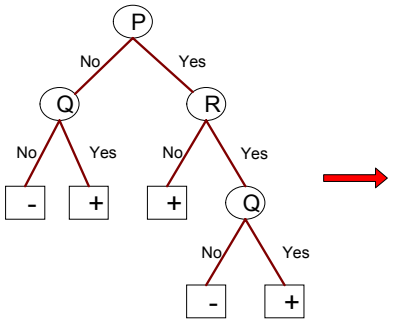
(Refund=No, Marital Status={Single,Divorced}, Taxable Income>80K) ==> Yes

(Refund=No, Marital Status={Married}) ==> No

- Κανόνες αμοιβαία αποκλειόμενοι και εξαντλητικοί
- Το σύνολο κανόνων περιέχει όση πληροφορία περιέχει και το δέντρο

Κατηγοριοποίηση με Κανόνες: Κατασκευή

Έμμεση Μέθοδος: Από Δέντρα Απόφασης σε Κανόνες

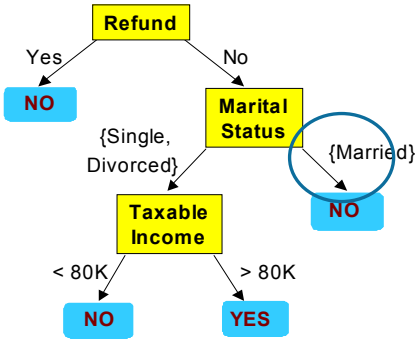


Rule Set

r1: (P=No,Q=No) ==> -
r2: (P=No,Q=Yes) ==> +
r3: (P=Yes,R=No) ==> +
r4: (P=Yes,R=Yes,Q=No) ==> -
r5: (P=Yes,R=Yes,Q=Yes) ==> +

Κατηγοριοποίηση με Κανόνες: Κατασκευή

Οι κανόνες μπορεί να απλοποιηθούν (απαλοιφή κάποιων όρων στο LHS αν δεν αλλάζει πολύ το λάθος)



Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Αρχικός Κανόνας: (Refund=No) ∧ (Status=Married) → No

Απλοποιημένος Κανόνας: (Status=Married) → No

Κατηγοριοποίηση με Κανόνες: Κατασκευή

Αν γίνει απλοποίηση (κλάδεμα):

- Οι κανόνες δεν είναι πια αμοιβαία αποκλειόμενοι
- Οι κανόνες δεν είναι πια εξαντλητικοί

Κατηγοριοποιητές Κοντινότερου Γείτονα

Κατηγοριοποιητές βασισμένοι σε Στιγμιότυπα

Μέχρι στιγμής

Κατηγοριοποίηση βασισμένη σε δύο βήματα

Βήμα 1: Induction Step – Κατασκευή Μοντέλου

Βήμα 2: Deduction Step – Εφαρμογή του μοντέλου για έλεγχο παραδειγμάτων

Eager Learners vs Lazy Learners

πχ Instance Based Classifiers (κατηγοριοποιητές βασισμένοι σε στιγμιότυπα)

Μην κατασκευάσεις μοντέλο αν δε χρειαστεί

Κατηγοριοποιητές βασισμένοι σε Στιγμιότυπα

Σύνολο Αποθηκευμένων Περιπτώσεων

Atr1	…….	AtrN	Class
			A
			B
			B
			C
			A
			C
			B

- Αποθήκευσε τις εγγραφές του συνόλου εκπαίδευσης
- Χρησιμοποίησε τις αποθηκευμένες εγγραφές για την εκτίμηση της κλάσης των νέων περιπτώσεων

Unseen Case

Atr1	…….	AtrN

Κατηγοριοποιητές βασισμένοι σε Στιγμιότυπα

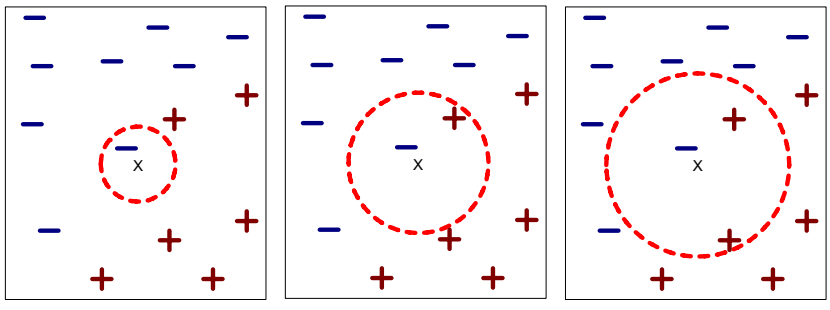
Παραδείγματα:

- Rote-learner
  - Κρατά (Memorizes) όλο το σύνολο των δεδομένων εκπαίδευσης και ταξινομεί μια εγγραφή αν ταιριάζει πλήρως με κάποιο από τα δεδομένα εκπαίδευσης

- Nearest neighbor – Κοντινότερος Γείτονας
  - Χρήση των *k* κοντινότερων "closest" σημείων (nearest neighbors) για την κατηγοριοποίηση

Κατηγοριοποιητές Κοντινότερου Γείτονα

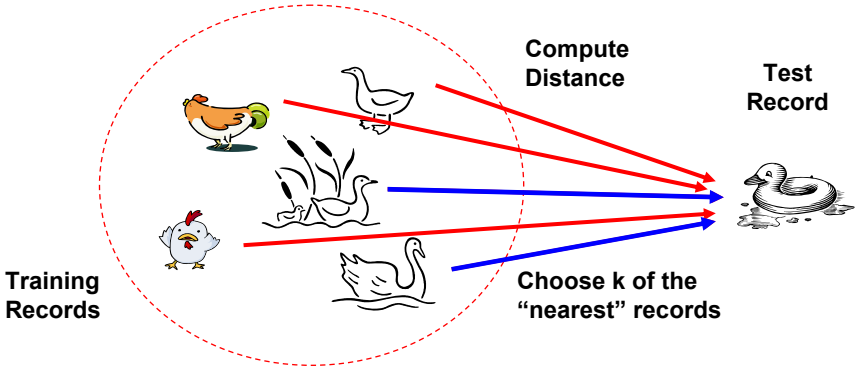
k-κοντινότεροι γείτονες μιας εγγραφής x είναι τα σημεία που έχουν την *k*-οστή μικρότερη απόσταση από το x



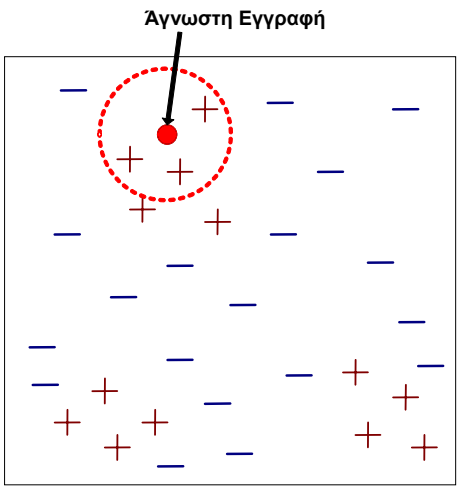
(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

Κατηγοριοποιητές Κοντινότερου Γείτονα

Basic idea: If it walks like a duck, quacks like a duck, then it's probably a duck



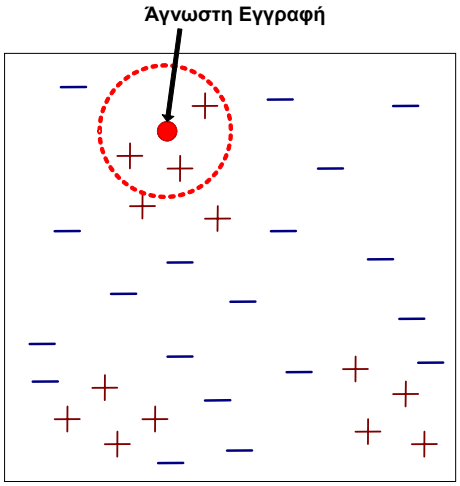
Κατηγοριοποιητές Κοντινότερου Γείτονα



Για να κατηγοριοποιηθεί μια άγνωστη εγγραφή:

- Υπολογισμός της απόστασης από τις εγγραφές του συνόλου
- Εύρεση των *k* κοντινότερων γειτόνων
- Χρήση των κλάσεων των κοντινότερων γειτόνων για τον καθορισμό της κλάσης της άγνωστης εγγραφής - π.χ., με βάση την πλειοψηφία (majority vote)

Κατηγοριοποιητές Κοντινότερου Γείτονα



Χρειάζεται

1. Το σύνολο των αποθηκευμένων εγγραφών
2. Distance Metric Μετρική απόστασης για να υπολογίσουμε την απόσταση μεταξύ εγγραφών
3. Την τιμή του *k*, δηλαδή τον αριθμό των κοντινότερων γειτόνων που πρέπει να ανακληθούν

Κατηγοριοποιητές Κοντινότερου Γείτονα

- Απόσταση μεταξύ εγγραφών:
  - Πχ ευκλείδεια απόσταση

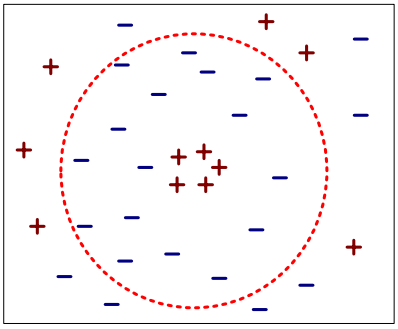
$$d(p,q) = \sqrt{\sum_i (p_i - q_i)^2}$$

- Καθορισμός τάξης
  - Απλά τη πλειοψηφική κλάση
  - Βάρος σε κάθε ψήφο με βάση την απόσταση
    - weight factor, $w = 1/d^2$

Κατηγοριοποιητές Κοντινότερου Γείτονα

Επιλογή της τιμής του *k*:

- *k* πολύ μικρό, ευαισθησία στα σημεία θορύβου
- *k* πολύ μεγάλο, η γειτονιά μπορεί να περιέχει σημεία από άλλες κλάσεις
- συχνά k = sqr(n), όπου n το μέγεθος του συνόλου εκπαίδευσης, σε εμπορικά συστήματα, συχνά, default, k=10



Κατηγοριοποιητές Κοντινότερου Γείτονα

- Θέματα Κλιμάκωσης
  - Τα γνωρίσματα ίσως πρέπει να κλιμακωθούν ώστε οι αποστάσεις να μην κυριαρχηθούν από κάποιο γνώρισμα
  - Παράδειγμα:
    - ύψος μπορεί να διαφέρει από 1.5m σε 1.8m
    - το βάρος μπορεί να διαφέρει από 90lb σε 300lb
    - το εισόδημα μπορεί να διαφέρει από $10Κσε $1M

- Δεν κατασκευάζεται μοντέλο, μεγάλο κόστος για την εφαρμογή της κατηγοριοποίησης
- Πολλές διαστάσεις (κατάρα των διαστάσεων)
- Θόρυβο (ελάττωση μέσω k-γειτόνων)

Κατηγοριοποιητές Κοντινότερου Γείτονα

- εξετάζουν και μη γραμμικές περιοχές
- το αποτέλεσμα δεν γίνεται άμεσα κατανοητό (στηρίζεται μόνο στην αρχή της τοπικότητας)

Κατηγοριοποιητές Κοντινότερου Γείτονα

2-διάστατα kd-δέντρα

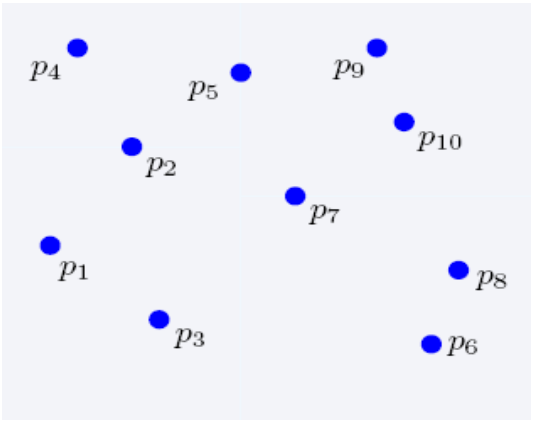
- Μια δομή για ερωτήματα διαστήματος στο R^2

Αλγόριθμος:

- Επέλεξε ή τη **x** ή τη **y** συντεταγμένη (εναλλάξ)
- Επέλεξε το διάμεσο (αυτό ορίζει μια οριζόντια ή κάθετη γραμμή)
- Αναδρομική κλήση

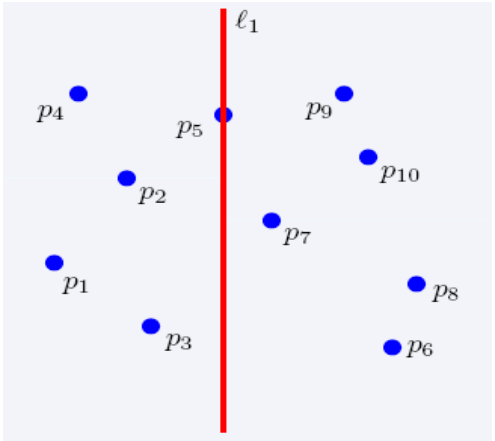
Κατηγοριοποιητές Κοντινότερου Γείτονα

2-διάστατα kd-δέντρα



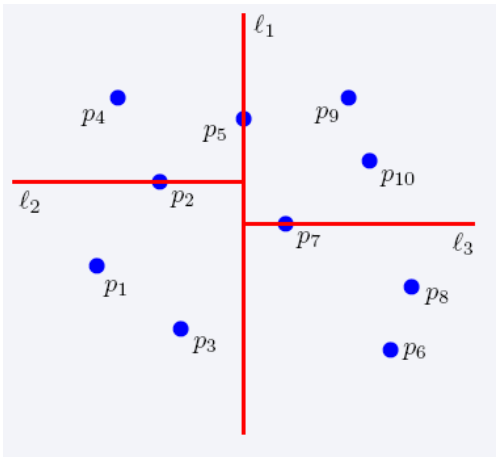
Κατηγοριοποιητές Κοντινότερου Γείτονα

2-διάστατα kd-δέντρα



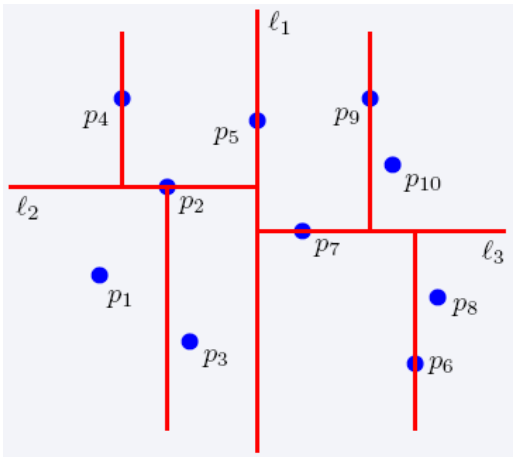
Κατηγοριοποιητές Κοντινότερου Γείτονα

2-διάστατα kd-δέντρα



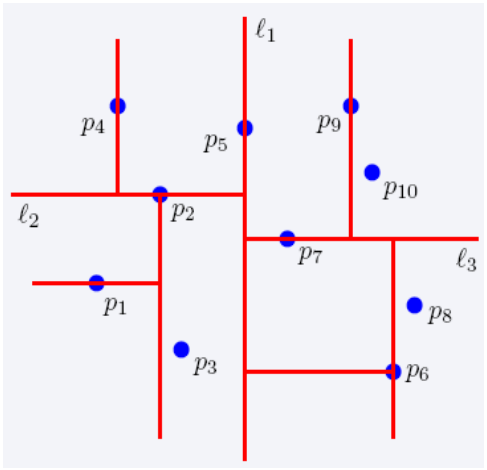
Κατηγοριοποιητές Κοντινότερου Γείτονα

2-διάστατα kd-δέντρα



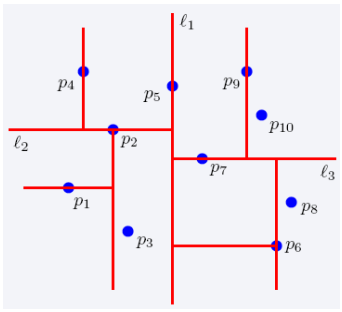
Κατηγοριοποιητές Κοντινότερου Γείτονα

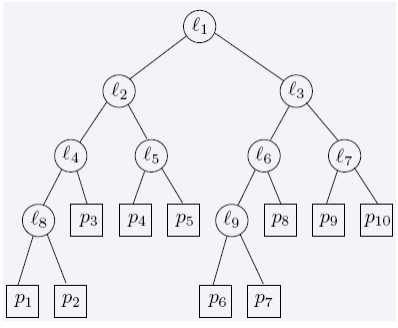
2-διάστατα kd-δέντρα



Κατηγοριοποιητές Κοντινότερου Γείτονα

2-διάστατα kd-δέντρα

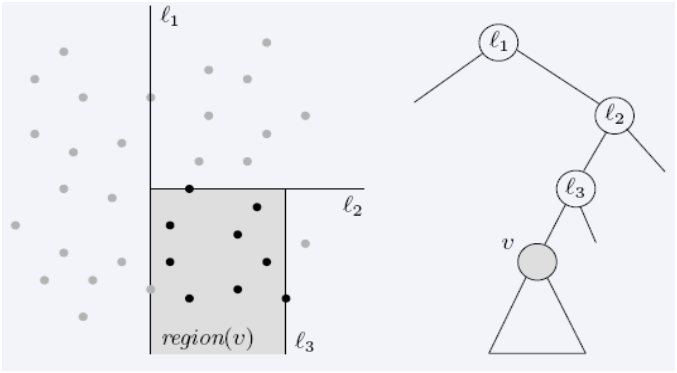




Κατηγοριοποιητές Κοντινότερου Γείτονα

2-διάστατα kd-δέντρα

Περιοχή του u – στο υποδέντρο με ρίζα το u όλα τα μαύρα σημεία



Κατηγοριοποιητές Κοντινότερου Γείτονα

2-διάστατα kd-δέντρα

- Μια δομή για ερωτήματα διαστήματος στο R^2

- Παίρνουμε ένα δυαδικό δέντρο:
  - Μέγεθος **O(n)**
  - Βάθος **O(logn)**
  - Χρόνος κατασκευής **O(nlogn)**

Επέκταση για παραπάνω από 2 διαστάσεις

- Παράδειγμα Binary Space Partitioning

Κατηγοριοποιητές Bayes

Κατηγοριοποιητές Bayes

X, Y τυχαίες μεταβλητές
Δεσμευμένη πιθανότητα (Conditional probability): **Pr(Y=y | X=x)**

Το θεώρημα του Bayes

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

Από κοινού πιθανότητα: **Pr(X=x,Y=y)**
Σχέση μεταξύ από κοινού (joint) και δεσμευμένης (conditional) πιθανότητας

$$P(Y \mid X) = \frac{P(X,Y)}{P(X)} \qquad P(X \mid Y) = \frac{P(X,Y)}{P(Y)}$$

Κατηγοριοποιητές Bayes

Το θεώρημα του Bayes: Παράδειγμα 1

- Δοθέντων ότι
  1. Αν κάποιος έχει περάσει το μάθημα "Προγραμματισμός σε C", περνάει το μάθημα "Δομές Δεδομένων" με πιθανότητα 4/5.
  2. Η εκ των προτέρων πιθανότητα κάποιος να περάσει το μάθημα "Προγραμματισμός σε C" είναι 1/3
  3. Η εκ των προτέρων πιθανότητα κάποιος να περάσει το μάθημα "Δομές Δεδομένων" είναι 2/3

- Πόσοι περνούν και τα δύο μαθήματα;
- Αν ξέρουμε ότι ένας φοιτητής έχει περάσει το μάθημα "Δομές Δεδομένων" ποια είναι η πιθανότητα να έχει περάσει το μάθημα "Προγραμματισμός σε C";

$$P(C=1 \mid \Delta=1) = \frac{P(\Delta=1 \mid C=1)P(C=1)}{P(\Delta=1)} = \frac{4/5 \times 1/3}{2/3} = 0.8$$

Κατηγοριοποιητές Bayes

Το θεώρημα του Bayes: Παράδειγμα 2

- Έστω 2 ομάδες, η Ομάδα 0 και η Ομάδα 1
  1. Η Ομάδα 0 νικά στο 65% των μεταξύ τους αγώνων
  2. Από τα παιχνίδια στα οποία νίκησε η Ομάδα 0, μόνο το 30% έγιναν στην έδρα της Ομάδας 1
  3. 75% των νικών της Ομάδας 1 γίνονται στην έδρα της

- Αν η Ομάδα 1 αναμένεται να φιλοξενήσει την Ομάδα 0 στον επόμενο αγώνα, ποια ομάδα εμφανίζεται ως πιθανότερη νικήτρια;

Κατηγοριοποιητές Bayes

Πως μπορούμε να χρησιμοποιήσουμε αυτό το θεώρημα για το πρόβλημα της κατηγοριοποίησης;

Κατηγοριοποιητές Bayes

X: σύνολο των γνωρισμάτων
Y: η μεταβλητή της κλάσης (κατηγορίας)

Y εξαρτάται από το **X** με μη ντετερμινιστικό τρόπο (***non-determininstic)***

P(*Y*|X) : Posterior probability (εκ των υστέρων)

P(*Y*): Prior probability (εκ των προτέρων)

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

...e 4.6. Training set for predicting borrowers who will default on loan payments.

X'=(Home Owner=No, Marital Status=Married, AnnualIncome=120K)

Υπολόγισε: **Pr(Yes|X'), Pr(No|X')** επέλεξε No ή Yes, ανάλογα με ποιο έχει τη μεγαλύτερη πιθανότητα

Πως θα υπολογίσουμε αυτές τις πιθανότητες;

Κατηγοριοποιητές Bayes

Φάση Εκπαίδευσης:

Εκμάθηση των εκ των υστέρων πιθανοτήτων **Pr(Y|X)** για κάθε συνδυασμό των **X** και **Y** βασισμένη στα δεδομένα εκπαίδευσης

Φάση Εφαρμογής:

Για κάθε εγγραφή ελέγχου **X'**, υπολόγισε την κλάση **Y'** που μεγιστοποιεί την εκ των υστέρων πιθανότητα **Pr(Y'|X')**

δηλαδή, την πιο πιθανή κλάση με βάση τα δεδομένα ελέγχου

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

P(X) είναι σταθερή και μπορούμε να την αγνοήσουμε (μαρτυρία - evidence)
P(Y): εκτιμάτε εύκολα από τα δεδομένα εισόδου, είναι το ποσοστό των δεδομένων εκπαίδευσης που ανήκουν στην κλάση Y (εκ των προτέρων πιθανότητα)
Pr(X|Y)?

Κατηγοριοποιητές Bayes

Υπολογισμός της εξαρτώμενης από τη κατηγορία πιθανότητας Pr(X|Y)

Υπάρχουν δύο βασικές μέθοδοι:

1. Απλοϊκός
2. Δίκτυο πεποίθησης

Θα δούμε την πρώτη μέθοδο

Οικογενειακή Κατάσταση	Αγοραστής
Διαζευγμένος	ΝΑΙ
Διαζευγμένος	ΝΑΙ
Έγγαμος	ΟΧΙ
Άγαμος	ΝΑΙ
Άγαμος	ΝΑΙ
Έγγαμος	ΟΧΙ
Διαζευγμένος	ΝΑΙ
Διαζευγμένος	ΝΑΙ
Διαζευγμένος	ΝΑΙ
Άγαμος	ΟΧΙ

Κατηγοριοποιητές Bayes

Παράδειγμα: 1 γνώρισμα (μεταβλητή) με κατηγορικές τιμές

Αν κάποιος είναι άγαμος, είναι αγοραστής ή όχι?

P(Ναι | Άγαμος) -> P(Άγαμος | Ναι) P(Ναι) = 2/7*7/10 = 0.2

P('Οχι | Άγαμος) -> P(Άγαμος | 'Οχι) P('Οχι) = 1/3 * 3/10 = 0.1

Κατηγοριοποιητές Bayes: Εκπαίδευση

Categorical attribute X_i

$Pr(Xi = x_i | Y=y)$: ποσοστό των δεδομένων εκπαίδευσης της κλάσης y που έχουν τιμή x_i στο i-οστό γνώρισμα

P(homeOwner = yes|No) = 3/7

P(MaritalStatus = Single| Yes) = 2/3

1. Τι γίνεται όταν έχουμε παραπάνω από ένα γνωρίσματα
2. Τι γίνεται όταν τα γνωρίσματα παίρνουν συνεχής τιμές

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

ure 4.6. Training set for predicting borrowers who will default on loan payments.

Κατηγοριοποιητές Bayes: Εκπαίδευση

Ηλικία	Οικογενειακή Κατάσταση	Αγοραστής
20	Διαζευγμένος	ΝΑΙ
30	Διαζευγμένος	ΝΑΙ
25	Έγγαμος	ΟΧΙ
30	Άγαμος	ΝΑΙ
40	Άγαμος	ΝΑΙ
20	Έγγαμος	ΟΧΙ
30	Διαζευγμένος	ΝΑΙ
25	Διαζευγμένος	ΝΑΙ
40	Διαζευγμένος	ΝΑΙ
20	Άγαμος	ΟΧΙ

Παράδειγμα

Αν κάποιος είναι άγαμος και 35 χρονών, είναι αγοραστής ή όχι?

Πρέπει να υπολογιστούν τα P(Ναι | Άγαμος, 35), P(Όχι | Άγαμος, 35)

Με βάση τα: P(Άγαμος, 35| Ναι) Π(Άγαμος, 35| Όχι)

Κατηγοριοποιητές Bayes

Περισσότερα από ένα γνωρίσματα (μεταβλητές)

Σύνολο $X = \{X_1,\ldots,X_d\}$ από d γνωρίσματα

Conditional independence (υπό συνθήκη ανεξαρτησία):
X είναι υπό συνθήκη ανεξάρτητο του Y, δοθέντος του Z αν:

$P(X|Y,Z) = P(X|Z)\ P(X,Y|Z) = P(X|Z)\ P(Y|Z)$

$$P(X \mid Y = y) = \prod_{i=1}^{d} P(X_i \mid Y = y)$$

Κατηγοριοποιητές Bayes

Σύνολο $X = \{X_1,\ldots,X_d\}$ από d γνωρίσματα

$$P(Y \mid X) = \frac{P(Y)\prod_{i=1}^{d} P(X_i \mid Y)}{P(X)}$$

Κατηγοριοποιητές Bayes: Εκπαίδευση

Περισσότερα από ένα γνωρίσματα (μεταβλητές)

Ηλικία	Οικογενειακή Κατάσταση	Αγοραστής
20	Διαζευγμένος	ΝΑΙ
30	Διαζευγμένος	ΝΑΙ
25	Έγγαμος	ΟΧΙ
30	Άγαμος	ΝΑΙ
40	Άγαμος	ΝΑΙ
20	Έγγαμος	ΟΧΙ
30	Διαζευγμένος	ΝΑΙ
25	Διαζευγμένος	ΝΑΙ
40	Διαζευγμένος	ΝΑΙ
20	Άγαμος	ΟΧΙ

Παράδειγμα

Αν κάποιος είναι άγαμος και 35 χρονών, είναι αγοραστής ή όχι?

Πρέπει να υπολογιστούν τα P(Ναι | Άγαμος, 35), P(Όχι | Άγαμος, 35)

Κατηγοριοποιητές Bayes

P(Ναι|Άγαμος, 35)-> P(Άγαμος,35|Ναι) * P(Ναι)=;
P(Όχι|Άγαμος, 35)-> P(Άγαμος,35| Όχι) * P(Όχι)=;

Υπόθεση: Ανεξαρτησία οικογενειακής κατάστασης και ηλικίας

P(Ναι|Άγαμος, 35)-> P(Άγαμος|Ναι) * P(35 |Ναι) * P(Ναι)=;
P(Όχι|Άγαμος, 35)-> P(Άγαμος| Όχι) * P(35 | Όχι) * P(Όχι)=;

Από το παράδειγμα μιας ιδιότητας, έχω ήδη υπολογίσει:
P(Άγαμος|Ναι) * P(Ναι) = 0.2
P(Άγαμος|Όχι) * P(Όχι) = 0.1

Παράδειγμα

Κατηγοριοποιητές Bayes

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

A: γνωρίσματα

M: mammals

N: non-mammals

$$P(A \mid M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A \mid N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A \mid M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A \mid N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Κατηγοριοποιητές Bayes

Εκτίμηση των Υπο Συνθήκη Πιθανοτήτων για Συνεχή Γνωρίσματα

Διακριτοποίηση (discretization)

Χωρίζουμε σε διαστήματα και η εκτίμηση γίνεται με βάση την αναλογία των εγγραφών εκπαίδευσης στο αντίστοιχο διάστημα

- πολλά διαστήματα -> λίγες εγγραφές εκπαίδευσης
- λίγα διαστήματα -> πιθανόν να συναθροίζουν εγγραφές που ανήκουν σε διαφορετικές κατηγορίες

Κατηγοριοποιητές Bayes

Εκτίμηση των Υπο Συνθήκη Πιθανοτήτων για Συνεχή Γνωρίσματα

Χρήση κάποιας κατανομής

Υποθέτουμε μια συγκεκριμένη μορφή κατανομής πιθανοτήτων

Συνήθως Gauss (κανονική) κατανομή

Χαρακτηρίζεται από δύο παραμέτρους

μέσο (μ)

διακύμανση (σ^2)

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

Κατηγοριοποιητές Bayes

Εκτίμηση των Υπο Συνθήκη Πιθανοτήτων για Συνεχή Γνωρίσματα

$$P(X_i = x_i \mid Y = y_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

Το μ_{ij} είναι το μέσο για όλα τα δεδομένα εκπαίδευσης της κατηγορίας (κλάσης) y_i

Όμοια εκτιμάται και η διακύμανση

Κατηγοριοποιητές Bayes

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Κανονική κατανομή

(Income, Class=No):

sample mean = 110

sample variance = 2975

$$P(Income = 120 \mid No) = \frac{1}{\sqrt{2\pi}(54.54)} e^{-\frac{(120-110)^2}{2(2975)}} = 0.0072$$

Κατηγοριοποιητές Bayes

Χρήση κανονικής κατανομής

$$\bar{x}_N = (20+30+30+\ldots 40)/7 = 30.71 \qquad s_N = 2.7$$

$$\bar{x}_O = (25+20+20)/3 = 21.67 \qquad s_O = 1.7$$

$$P(35 \leq \text{Ηλικία} \leq 35+\varepsilon|\text{'Ναι'}) = \int_{35}^{35+\varepsilon} \frac{1}{\sqrt{2\pi}s_N} e^{-\frac{(35-\bar{x}_N)^2}{2s_N^2}} \simeq \varepsilon \frac{1}{\sqrt{2\pi}s_N} e^{-\frac{(35-\bar{x}_N)^2}{2s_N^2}} = 0.11\varepsilon$$

$$P(35 \leq \text{Ηλικία} \leq 35+\varepsilon|\text{'Οχι'}) = \int_{35}^{35+\varepsilon} \frac{1}{\sqrt{2\pi}s_O} e^{-\frac{(35-\bar{x}_O)^2}{2s_O^2}} \simeq \varepsilon \frac{1}{\sqrt{2\pi}s_O} e^{-\frac{(35-\bar{x}_O)^2}{2s_O^2}} = 10^{-14}\varepsilon$$

το ε κανονικοποιείται οπότε μπορούμε να χρησιμοποιήσουμε την προηγούμενη εξίσωση

Κατηγοριοποιητές Bayes

Χρήση κανονικής κατανομής

P(Ναι|Άγαμος, 35)->
P(Άγαμος|Ναι) P(35 |Ναι) * P(Ναι)= 0.2 * 0.11ε = 0.022 ε

P('Οχι|Άγαμος, 35)->
P(Άγαμος| 'Οχι) P(35 | 'Οχι) * P('Οχι)= 0.1 * 10^{-14} ε= 10^{-15} ε

Άρα, αγοραστής

Κατηγοριοποιητές Bayes

	binary	categorical	continuous	class
Tid	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Figure 4.6. Training set for predicting borrowers who will default on loan payments.

X' = (HomeOwner = No, MaritalStatus = Married, Income=120K)
Πρέπει να υπολογιστεί Pr(Y|X'), δηλαδή Pr(Y)xPr(X'|Y)
But P(X'|Y) is
Y = No:
P(HO=No|No) x P(MS=Married|No) x P(Inc=120K|No) =
4/7 x 4/7 x 0.0072 = 0.0024
Y=Yes:
P(HO=No|Yes) x P(MS=Married|Yes) x P(Inc=120K|Yes) =
1 x 0 x 1.2x10-9 = 0

Κατηγοριοποιητές Bayes

X' = (HomeOwner = No, MaritalStatus = Married, Income=120K)
Pr(X'|Y = Yes) είναι 0!
Επειδή τα δείγματα εκπαίδευσης μπορεί να μην καλύπτουν όλες τις κατηγορίες -> Διαδικασία Διόρθωσης

$$\Pr(X_i = x_i \mid Y = y_j) = \frac{n_c + mp}{n + m}$$

n_c: ο αριθμός των εγγραφών εκπαίδευσης της κλάσης y_j που παίρνουν την τιμή x_i
n: συνολικός αριθμός εγγραφών της κλάσης y_j

m: μια παράμετρος που καλείται ισοδύναμο μέγεθος δείγματος (equivalent sample size) (ισορροπεί την εκ των υστέρων (nc/n) και την εκ των προτέρων (p) πιθανότητας

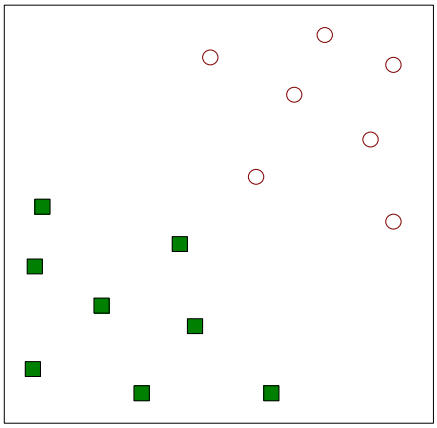
p: μια παράμετρος που καθορίζει ο χρήστης (η εκ των προτέρων πιθανότητα εμφάνισης της τιμής x_i για το γνώρισμα X_i μεταξύ των εγγραφών της κλάσης y_i)

Κατηγοριοποιητές Bayes

- Ανοχή σε μη σχετικά γνωρίσματα - Αν το X_i δεν είναι σχετικό (irrelevant), $P(X_i|Y)$ είναι σχεδόν uniform
- Πρόβλημα όταν υπάρχουν εξαρτήσεις μεταξύ των γνωρισμάτων (μεταβλητών) (correlated attributes)
- Καλή κλιμάκωση σε μεγάλο όγκο δεδομένων, μια απλή ανάγνωση των δεδομένων εκπαίδευσης
- Καλή ανοχή στο θόρυβο, γιατί τα σημεία θορύβου εξομαλύνονται
- Δεν επηρεάζονται από τιμές που λείπουν γιατί αυτές μπορούμε να τις αγνοήσουμε

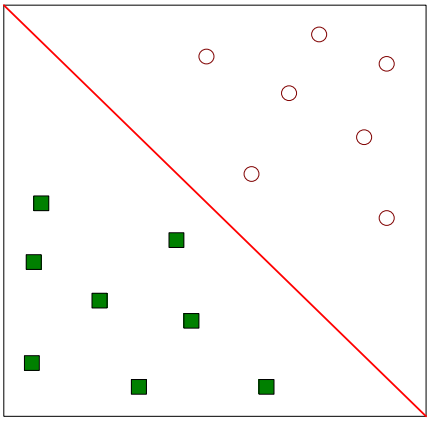
Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines)

Κατηγοριοποιητές SVM



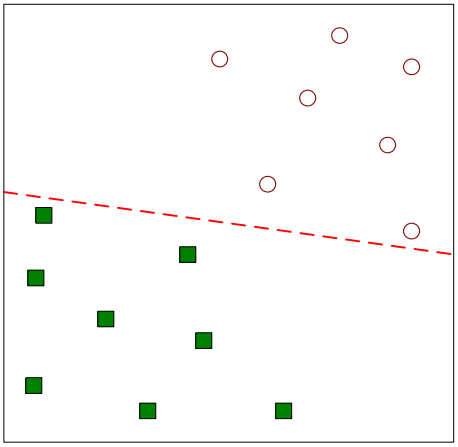
• Βρες ένα γραμμικό υπερ-επίπεδο (όριο απόφασης) που να διαχωρίζει τα δεδομένα

Κατηγοριοποιητές SVM



- Μία πιθανή λύση

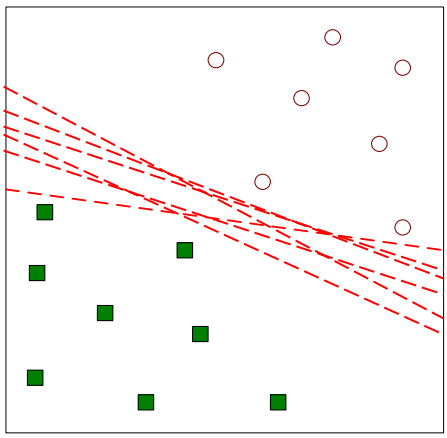
Κατηγοριοποιητές SVM



- Μια ακόμα πιθανή λύση

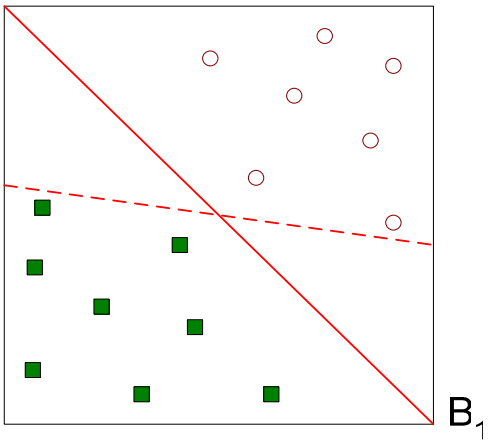
B_2

Κατηγοριοποιητές SVM



- Άλλες πιθανές λύσεις

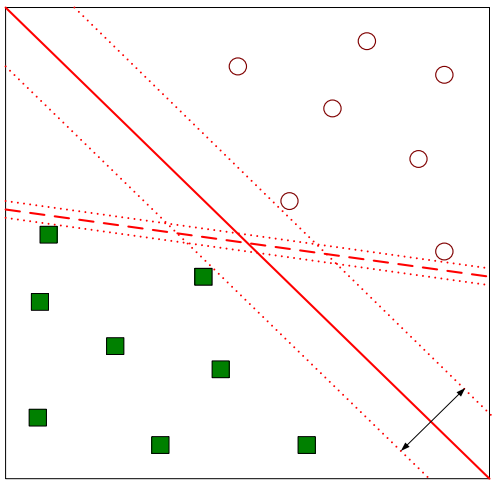
Κατηγοριοποιητές SVM



- Ποια είναι καλύτερη η B1 ή η B2?
- Πως ορίζεται το καλύτερη; Με ποιο κριτήριο;

B2

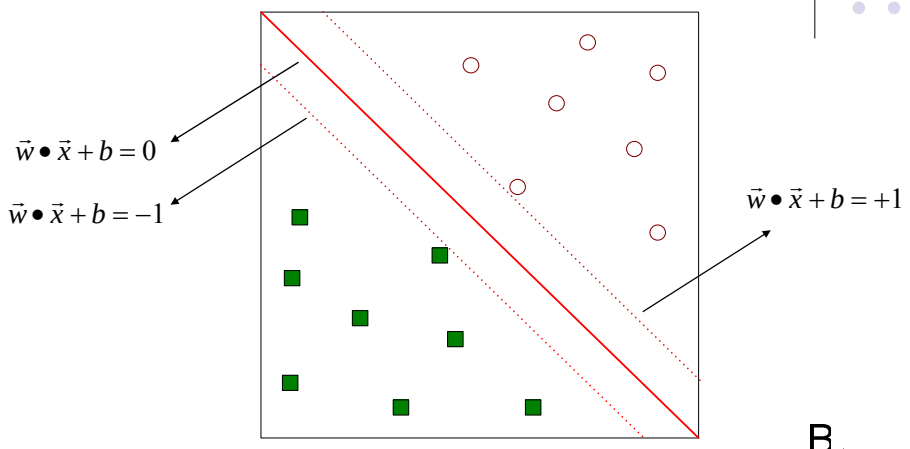
Κατηγοριοποιητές SVM



- Το υπερ-επίπεδο που μεγιστοποιεί το περιθώριο (margin) => το B1 είναι καλύτερο από το B2 (χωρητικότητα)

Γραμμικό SVM

Κατηγοριοποιητές SVM



$$f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x} + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x} + b \leq -1 \end{cases}$$

1 κλάση κύκλος

-1 κλάση τετράγωνο

$$\text{Margin} = \frac{2}{\| \vec{w} \|^2}$$

Κατηγοριοποιητές SVM

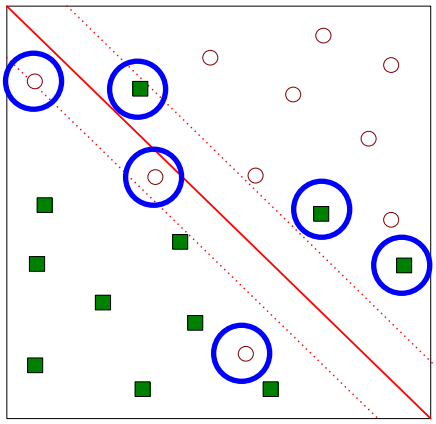
- Θέλουμε να μεγιστοποιήσουμε: $\text{Margin} = \frac{2}{\| \vec{w} \|^2}$
  - Το οποίο είναι ισοδύναμο με το να ελαχιστοποιήσουμε: $L(w) = \frac{\| \vec{w} \|^2}{2}$
  - Με βάση τους παρακάτω περιορισμούς (constraints):

$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x}_i + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x}_i + b \leq -1 \end{cases}$$

  - Ένα πρόβλημα βελτιστοποίησης περιορισμών (constrained optimization problem)
    - Αριθμητικές μέθοδοι για την επίλυση του

Κατηγοριοποιητές SVM

- Τι συμβαίνει αν το πρόβλημα δεν είναι γραμμικώς διαχωρίσιμο



Κατηγοριοποιητές SVM

- Εισαγωγή χαλαρών μεταβλητών (slack variables)
  - Ελαχιστοποίηση:

$$L(w) = \frac{\| \vec{w} \|^2}{2} + C\left(\sum_{i=1}^{N} \xi_i^k \right)$$

  - Με τους περιορισμούς:

$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x}_i + b \geq 1 - \xi_i \\ -1 & \text{if } \vec{w} \bullet \vec{x}_i + b \leq -1 + \xi_i \end{cases}$$

Κατηγοριοποιητές SVM