Αναγνώριση Προτύπων

K-NN / Πιθανοτικοί ταξινομητές

Ανδρέας Λ. Συμεωνίδης

Αν. Καθηγητής

Τμήμα Ηλεκτρολόγων Μηχ/κών
&
Μηχ/κών Υπολογιστών, Α.Π.Θ.

Email: asymeon@eng.auth.gr

issel
Intelligent Systems
& Software Engineering
Labgroup

ARISTOTLE
UNIVERSITY OF
THESSALONIKI

Διάρθρωση διάλεξης

- K-Nearest Neighbor
- Πιθανοτικοί ταξινομητές

K-Nearest Neighbor

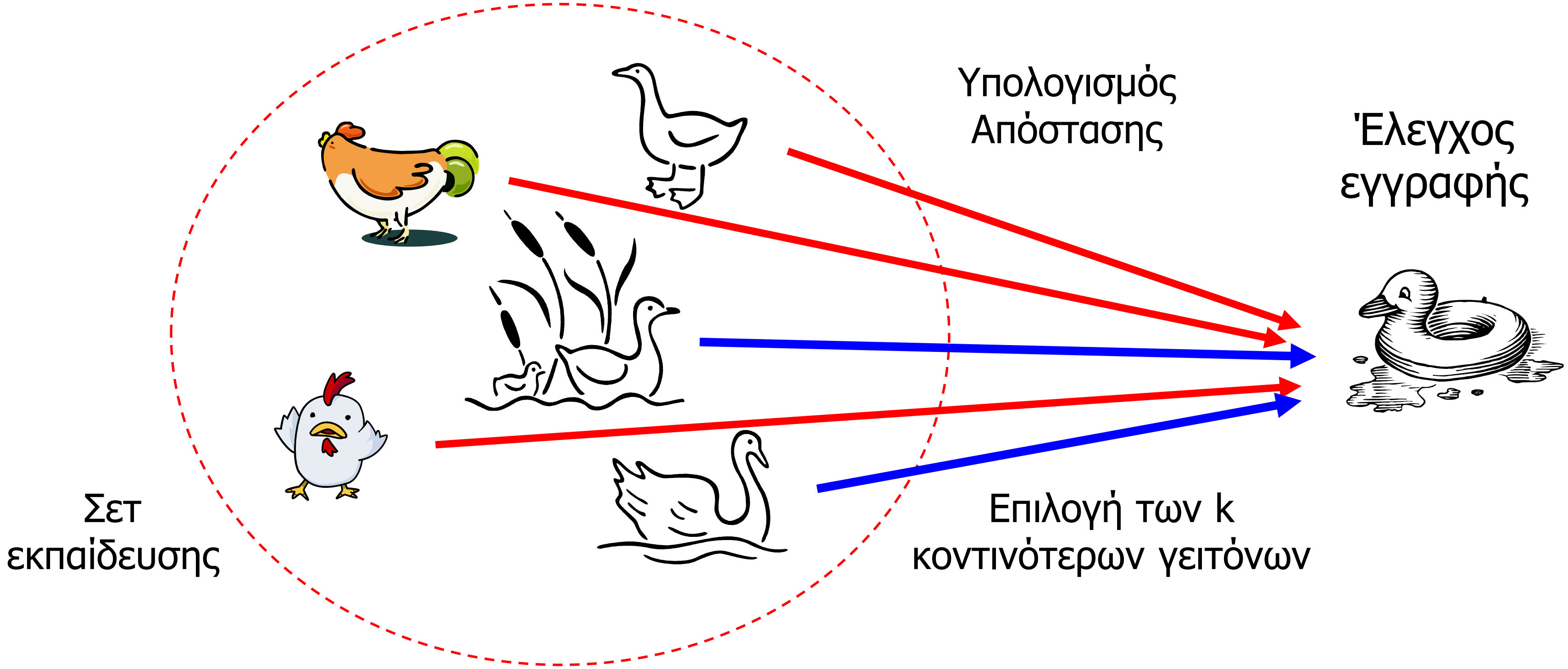
- Instanced-based classifiers
- Voronoi diagrams
- Ταξινόμηση κοντινότερου γείτονα

Ταξινομητές βασισμένοι στις εγγραφές

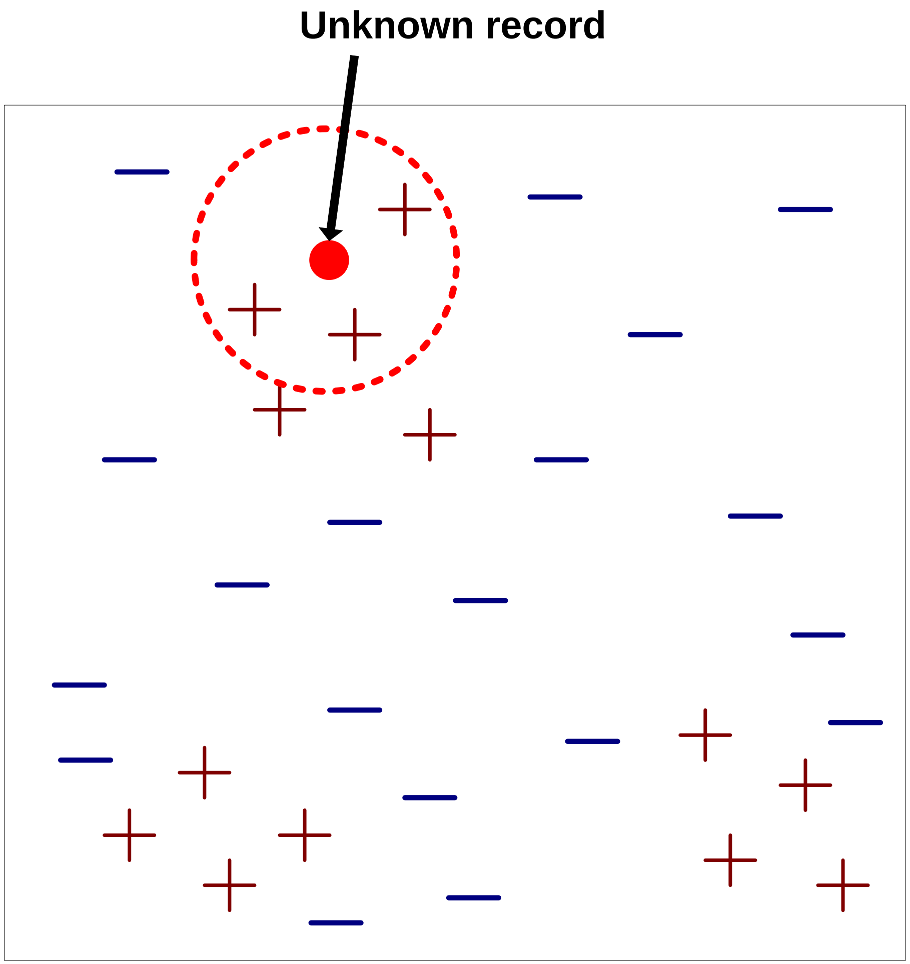
- Παραδείγματα:
  - Rote-learner
    - Περνά στη μνήμη και «θυμάται» όλο το σετ εκπαίδευσης.
    - Κάνει ταξινόμηση μόνο εάν οι τιμές των μεταβλητών μιας νέας εγγραφής ταιριάζουν απόλυτα με κάποια από το σετ εκπαίδευσης
  - Nearest neighbor
    - Χρησιμοποιεί τα k “κοντινότερα” σημεία (nearest neighbors) για να κάνει ταξινόμηση

Nearest Neighbor ταξινομητές

- Η βασική ιδέα:
  - Εάν περπατά σαν πάπια και ακούγεται σαν πάπια, τότε είναι μάλλον πάπια



Nearest-Neighbor ταξινομητής



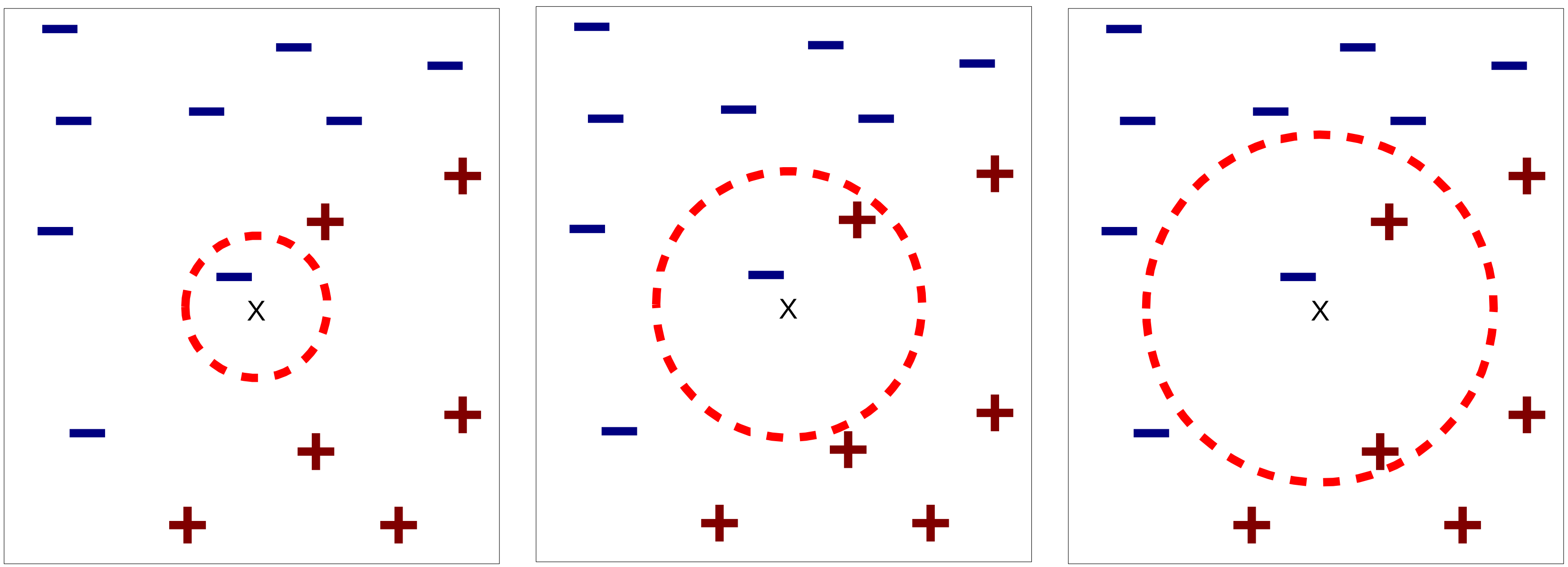
Απαιτεί 3 πράγματα:

1. Το σετ εκπαίδευσης
2. Μια μετρική απόστασης ώστε να υπολογίζει την απόσταση ανάμεσα στα σημεία
3. Την τιμή του *k*, του αριθμού των κοντινότερων γειτόνων

Για την ταξινόμηση ενός νέου σημείου:

- Υπολογίστε την απόστασή του από τα σημεία του σετ εκπαίδευσης
- Αναγνωρίστε τους *k* κοντινότερους γείτονές του
- Επιλέξτε την κλάση του σημείου βάσει των κλάσεων των κοντινότερων γειτόνων

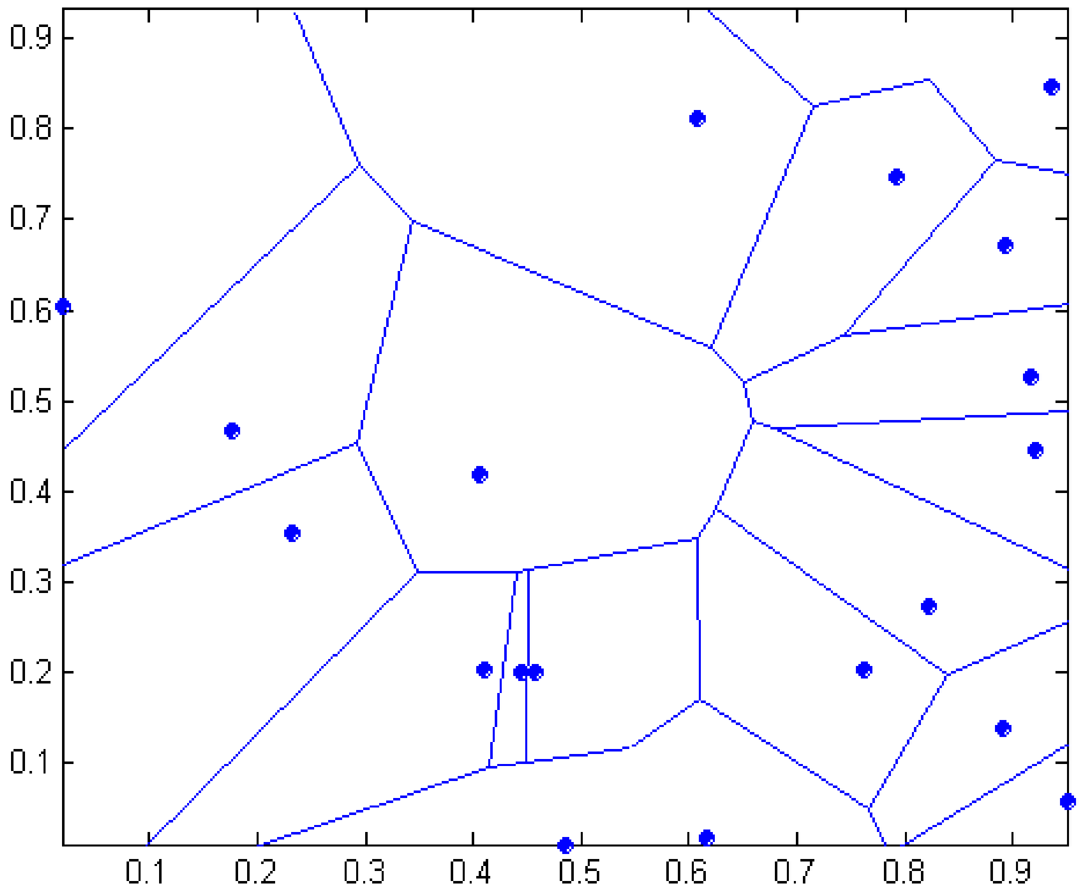
Καθορισμός του Nearest Neighbor



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

1 nearest-neighbor

Διάγραμμα Voronoi



Nearest Neighbor ταξινόμηση

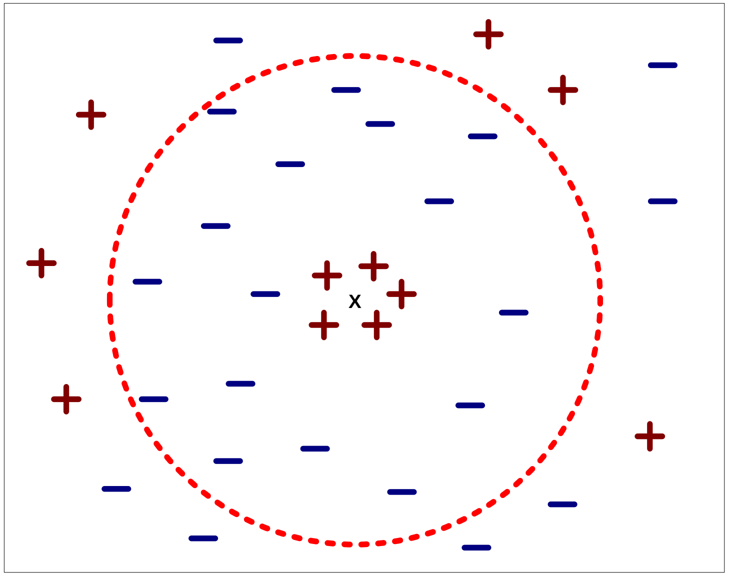
- Υπολογισμός της απόστασης ανάμεσα σε 2 σημεία:
  - Ευκλείδεια απόσταση

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

- Καθορισμός της κλάσης από τη λίστα των k γειτόνων
  - Επιλογή της κλάσης της πλειοψηφίας
  - Προσθήκη βάρους ανάλογα με την απόσταση
    - π.χ. συντελεστής βάρους, $w = 1/d^2$

Nearest Neighbor ταξινόμηση

- Επιλογή της τιμής του k:
  - Εάν το k είναι πολύ μικρό, ευαίσθητος στον θόρυβο
  - Εάν το k είναι πολύ μεγάλο, η γειτονιά μπορεί να περιέχει σημεία από άλλες κλάσεις



Nearest Neighbor ταξινόμηση

- Θέματα κλιμάκωσης
  - Οι μεταβλητές πρέπει να είναι κανονικοποιημένες
- Πρόβλημα της Ευκλείδειας απόστασης:
  - Μεγάλη διαστατικότητα των δεδομένων (curse of dimensionality)

1 1 1 1 1 1 1 1 1 1 1 0		1 0 0 0 0 0 0 0 0 0 0 0
	vs	
0 1 1 1 1 1 1 1 1 1 1 1		0 0 0 0 0 0 0 0 0 0 0 1
d = 1.4142		d = 1.4142

Λύση: Κανονικοποίηση των διανυσμάτων σε μήκος μονάδας (unit length)

Πιθανοτικοί Ταξινομητές

- Θεώρημα Bayes
- Ανεξαρτησία μεταβλητών
- Εκτίμηση πιθανότητας

Bayes Classifier

- Μια πιθανοτική υποδομή επίλυσης προβλημάτων ταξινόμησης
- Δεσμευμένη πιθανότητα:

$$P(C \mid A) = \frac{P(A,C)}{P(A)}$$

$$P(A \mid C) = \frac{P(A,C)}{P(C)}$$

- Θεώρημα Bayes:

$$P(C \mid A) = \frac{P(A \mid C)P(C)}{P(A)}$$

Παράδεγμα του θεωρήματος Bayes

- Δεδομένα:
  - Ένας γιατρός γνωρίζει ότι η μηνιγγίτιδα προκαλεί πόνο στον λαιμό 50% των περιπτώσεων
  - Η πιθανότητα να έχει ένας ασθενής μηνιγγίτιδα είναι 1/50,000
  - Η πιθανότητα να έχει ένας ασθενής πόνο στον λαιμό είναι 1/20
- Ποια η πιθανότητα ένας ασθενής με πόνο στον λαιμό να έχει μηνιγγίτιδα;

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Πιθανοτικοί (Bayesian) ταξινομητές

- Θεωρήστε κάθε μεταβλητή, καθώς και τη μεταβλητή-κλάση ως τυχαίες μεταβλητές
- Δεδομένης μιας εγγραφής με μεταβλητές $(A_1, A_2,\dots,A_n)$
  - Στόχος είναι να προβλεφθεί η κλάση C
  - Συγκεκριμένα, επιθυμούμε την εύρεση της τιμής του C που μεγιστοποιεί την πιθανότητα $P(C| A_1, A_2,\dots,A_n)$
- Μπορούμε να εκτιμήσουμε την $P(C| A_1, A_2,\dots,A_n)$ απευθείας από τα δεδομένα;

Bayesian ταξινομητές

- Προσέγγιση:
  - Υπολογισμός της μεθύστερης (posterior) πιθανότητας $P(C \mid A_1, A_2, \ldots, A_n)$ για όλες τις τιμές του C με βάση το θεώρημα Bayes:

$$\boxed{P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}}$$

  - Επιλογή της τιμής του C που μεγιστοποιεί την:
    $P(C \mid A_1, A_2, \ldots, A_n)$
  - Ισοδύναμο με την επιλογή της τιμής του C που μεγιστοποιεί την
    $P(A_1, A_2, \ldots, A_n \mid C)\, P(C)$
- Πώς γίνεται η εκτίμηση της $P(A_1, A_2, \ldots, A_n \mid C)$;

Απλοϊκός ταξινομητής Bayes

- Θεωρούμε ανεξαρτησία ανάμεσα στις μεταβλητές A_i δεδομένης της κλάσης:
  - $P(A_1, A_2, \dots, A_n | C) = P(A_1 | C_j)\, P(A_2 | C_j) \dots P(A_n | C_j)$
  - Μπορούμε να εκτιμήσουμε τα $P(A_i | C_j)$ για όλες τις A_i και την C_j.
  - Η νέα εγγραφή ταξινομείται ως C_j εάν η πιθανότητα $P(C_j) \prod P(A_i | C_j)$ είναι μέγιστη.

Υπολογισμός πιθανοτήτων από τα δεδομένα

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**

- Κλάση: $P(C) = N_c/N$
  - π.χ., *P(No) = 7/10,*
    P(Yes) = 3/10
- Για διακριτές μεταβλητές:
  $P(A_i \mid C_k) = |A_{ik}|/ N_c$
  - Όπου $|A_{ik}|$ είναι ο αριθμός των στιγμιοτύπων που έχουν την τιμή A_i μιας μεταβλητής και ανήκουν στην κλάση C_k
  - Παραδείγματα:
    P(Status=Married|No) = 4/7
    P(Refund=Yes|Yes)=0

Υπολογισμός πιθανοτήτων από τα δεδομένα

- Για συνεχείς μεταβλητές:
  - Διακριτοποίηση του εύρους σε τμήματα
    - Μια ordinal μεταβλητή ανά τμήμα
    - Παραβιάζει την υπόθεση ανεξαρτησίας
  - Δυαδικός χωρισμός (split): (A < v) or (A > v)
    - Επιλογή μόνο ενός από τα δυο τμήματα ως μια νέα μεταβλητή
  - Εκτίμηση πυκνότητας πιθανότητας:
    - Υποθέστε ότι η μεταβλητή ακολουθεί κανονική κατανομή
    - Χρησιμοποιήστε τα δεδομένα για τον υπολογισμό των παραμέτρων της κατανομής (μέση τιμή, τυπική απόκλιση)
    - Με γνωστή την κατανομή πιθανότητας, χρήση της για τον υπολογιμό της δεσμευμένης πιθανότητας $P(A_i|c)$

Υπολογισμός πιθανοτήτων από τα δεδομένα

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

- Κανονική κατανομή:

$$P(A_i \mid c_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i-\mu_{ij})^2}{2\sigma_{ij}^2}}$$

  - Μια για κάθε ζεύγος (A_i,c_j)
- Για (Income, Class=No):
  - Εάν Class=No
    - sample mean = 110
    - sample variance = 2975

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

Παράδειγμα ενός απλοϊκού ταξινομητή Bayes

Δεδομένης μιας εγγραφής ελέγχου:

$$X = (\text{Refund} = \text{No}, \text{Married}, \text{Income} = 120\text{K})$$

naive Bayes Classifier:

P(Refund=Yes|No) = 3/7
P(Refund=No|No) = 4/7
P(Refund=Yes|Yes) = 0
P(Refund=No|Yes) = 1
P(Marital Status=Single|No) = 2/7
P(Marital Status=Divorced|No)=1/7
P(Marital Status=Married|No) = 4/7
P(Marital Status=Single|Yes) = 2/7
P(Marital Status=Divorced|Yes)=1/7
P(Marital Status=Married|Yes) = 0

For taxable income:
If class=No: sample mean=110
sample variance=2975
If class=Yes: sample mean=90
sample variance=25

- P(X|Class=No) = P(Refund=No|Class=No) $\times$ P(Married| Class=No) $\times$ P(Income=120K| Class=No) = 4/7 $\times$ 4/7 $\times$ 0.0072 = 0.0024
- P(X|Class=Yes) = P(Refund=No| Class=Yes) $\times$ P(Married| Class=Yes) $\times$ P(Income=120K| Class=Yes) = 1 $\times$ 0 $\times$ 1.2 $\times$ 10^{-9} = 0

Ισχύει ότι: P(X|No)P(No) > P(X|Yes)P(Yes)

Κατά συνέπεια : P(No|X) > P(Yes|X)
=> Class = No

Απλοϊκός ταξινομητής Bayes

- Εάν μια από τις δεσμευμένες πιθανότητες είναι 0, τότε μηδενίζεται ολόκληρη η έκφραση
- Εκτίμηση πιθανότητας:

$$\text{Original}: P(A_i \mid C) = \frac{N_{ic}}{N_c}$$

$$\text{Laplace}: P(A_i \mid C) = \frac{N_{ic} + 1}{N_c + c}$$

$$\text{m - estimate}: P(A_i \mid C) = \frac{N_{ic} + mp}{N_c + m}$$

c: ο αριθμός των κλάσεων

p: η πιθανότητα

m: παράμετρος

Παράδειγμα ενός απλοϊκού ταξινομητή 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

A: Attributes

M: mammals

N: non-mammals

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

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

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

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

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

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

=> Mammals

Απλοϊκός Bayes - Σύνοψη

- Εύρωστος σε σχέση με μεμονωμένα σημεία θορύβου
- Διαχείριση των τιμών που λείπουν, αγνοώντας τις αντίστοιχες εγγραφές κατά τον υπολογισμό της εκτίμησης πιθανοτήτων
- Εύρωστος σε μη σχετικές (με το πρόβλημα) μεταβλητές
- Η υπόθεση ανεξαρτησίας μπορεί να μην ισχύει για κάποιες μεταβλητές
  - Χρήση άλλων τεχνικών όπως Bayesian Belief Networks (BBN)

Bayesian Belief Networks

- Κοινή κατανομή πιθανότητας
- Ανεξαρτησία υπό συνθήκη
- Δημιουργία Bayesian δικτύων
- Μοντέλα μειγμάτων (mixture models)

Η υπόθεση ανεξαρτησίας…

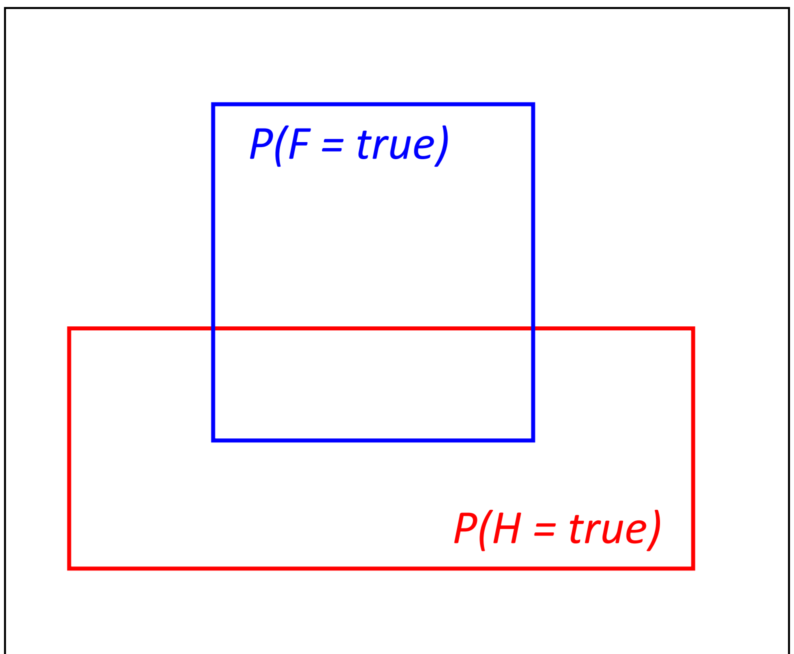
- … κάνει τον υπολογισμό δυνατό
- … όταν ικανοποιείται δημιουργεί βέλτιστους ταξινομητές
- … σπάνια ικανοποιείται στην πράξη, καθώς τα χαρακτηριστικά (μεταβλητές) συχνά συσχετίζονται.
- Προσπάθεια υπερκερασμού του προβλήματος:
  - Δένδρα απόφασης, τα οποία συμπεραίνουν για ένα χαρακτηριστικό τη φορά, ξεκινώντας από τα πιο σημαντικά
  - Bayesian δίκτυα, τα οποία συνδυάζουν Bayesian λογική με αιτιατές σχέσεις ανάμεσα στα χαρακτηριστικά

Η κοινή κατανομή πιθανότητας

P(F = true, H = true):

"η πιθανότητα του *F = true* **και** *H = true*"

- Στην περίπτωση που *P(F = true | H = true)*



$$P(H=true|F=true)$$

$$=\frac{\text{Area of "H and F" region}}{\text{Area of "F" region}}$$

$$=\frac{\mathrm{P(H = true, F = true)}}{\mathrm{P(F = true)}}$$

Γενικά, *P(X|Y)=P(X,Y)/P(Y)*

Η κοινή κατανομή πιθανότητας

- Οι κοινές πιθανότητες μπορούν να είναι ανάμεσα σε οποιοδήποτε αριθμό χαρακτηριστικών

  π.χ. *P(A = true, B = true, C = true)*
- Για κάθε συνδυασμό τιμών, πρέπει να πούμε πόσο πιθανός είναι ο συνδυασμός
- Το άθροισμα όλων των συνδυασμών είναι 1
- Ο υπολογισμός όλων των συνδυασμών είναι επίπονος:

  για *k* Boolean τυχαία χαρακτηριστικά, χρειάζεται ένας πίνακας μεγέθους 2^k
- Γι’ αυτό και ορίσαμε την έννοια της **ΑΝΕΞΑΡΤΗΣΙΑΣ**

A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

Άθροισμα 1

Ανεξαρτησία (Υπενθύμιση)

Οι μεταβλητές *A* και *B* είναι ανεξάρτητες αν ισχύει οποιοδήποτε από τα παρακάτω:

- $P(A,B) = P(A)\,P(B)$
- $P(A \mid B) = P(A)$
- $P(B \mid A) = P(B)$

Αυτό σημαίνει ότι το να γνωρίζω το αποτέλεσμα του *A* δε μου λέει τίποτα για το αποτέλεσμα του *B*

Ανεξαρτησία υπό συνθήκη

Οι μεταβλητές *A* και *B* είναι ανεξάρτητες υπό συνθήκη δεδομένου του *C* αν ισχύει οποιοδήποτε από τα παρακάτω:

- $P(A, B \mid C) = P(A \mid C)\, P(B \mid C)$
- $P(A \mid B, C) = P(A \mid C)$
- $P(B \mid A, C) = P(B \mid C)$

Γνώση του *C* δηλώνει τα πάντα για το *B*. Δε γνωρίζω τίποτα για το *A* (είτε γιατί το *A* δεν επηρεάζει το *B* είτε γιατί η γνώση του *C* δίνει όλη την πληροφορία που θα έδινε το *A*)

Bayesian δίκτυα (δίκτυα πιθανοτήτων)

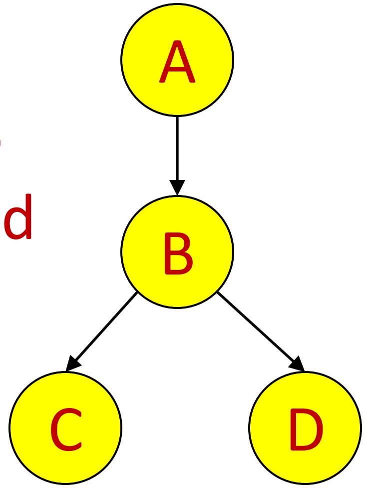
Βασικά χαρακτηριστικά τους:

1. Κωδικοποιούν τις σχέσεις ανάμεσα στις υπό συνθήκη ανεξάρτητες μεταβλητές που υπάρχουν στον γράφο
2. Είναι μια συμπαγής αναπαράσταση της κοινής κατανομής πιθανότητας στο σύνολο των μεταβλητών

Ένα Bayesian Δίκτυο

Αποτελείται από:

1. Έναν κατευθυνόμενο ακυκλικό γράφο (Directed Acyclic Graph – DAG)



2. Ένα σετ από πίνακες για κάθε κόμβο του γράφου

A	P(A)
false	0.6
true	0.4

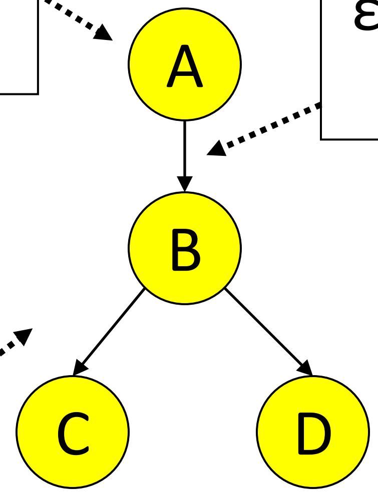
A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

Κατευθυνόμενος ακυκλικός γράφος

Κάθε κόμβος στον γράφο είναι μια τυχαία μεταβλητή



Ένας κόμβος *X* είναι γονέας ενός κόμβου *Y* εάν υπάρχει ακμή από τον *X* στον *Y*

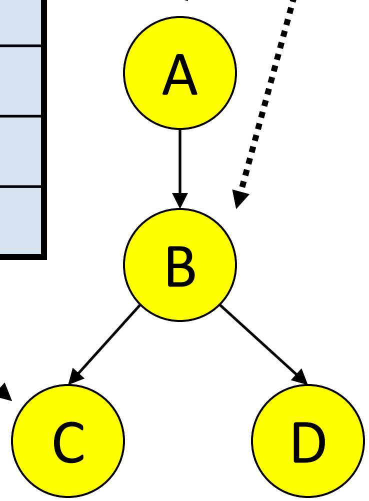
Άτυπα, ένα βέλος από τον *X* στον *Y* σημαίνει ότι ο *X* έχει άμεση επίδραση στον *Y*

Ένας πίνακας για κάθε κόμβο

A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1



B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

- Κάθε κόμβος X_i έχει δεσμευμένη πιθανότητα P(X_i |Parents(X_i)), η οποία ποσοτικοποιεί την επίδραση των γονέων στον κόμβο (Συνθήκη Markov)
- Οι παράμετροι του δικτύου είναι οι πιθανότητες στους πίνακες δεσμευμένης πιθανότητας (CPTs)

Ένας πίνακας για κάθε κόμβο (συν.)

Η κατανομή της δεσμευμένης πιθανότητας του C δεδομένου του B

B	**C**	**P(C|B)**
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

Για δεδομένο συνδυασμό τιμών των γονέων (B στην περίπτωση αυτή), οι τιμές του P(C=true | B) και P(C=false | B) πρέπει να αθροίζονται στο 1

π.χ. P(C=true | B=false) + P(C=false |B=false)=1

Εάν έχετε Boolean μεταβλητή με k Boolean γονείς, ο πίνακας περιέχει 2^{k+1} πιθανότητες (αλλά χρειάζεται να αποθηκευτούν οι 2^{k})

Η κοινή κατανομή πιθανότητας

Βάσει της συνθήκης Markov μπορούμε να υπολογίσουμε την κοινή κατανομή πιθανότητας στο σύνολο των μεταβλητών X_1, …, X_n του δικτύου ως εξής:

$$P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i \mid Parents(X_i))$$

Όπου Parents(X_i) δηλώνουν τις τιμές των γονέων του κόμβου X_i ως προς τον γράφο

Bayesian δίκτυο - Παράδειγμα

Έστω ότι το δίκτυο με τις τιμές πιθανοτήτων που ορίστηκαν παραπάνω και θέλουμε να υπολογίσουμε το:

Αυτό καθορίζεται από τον γράφο

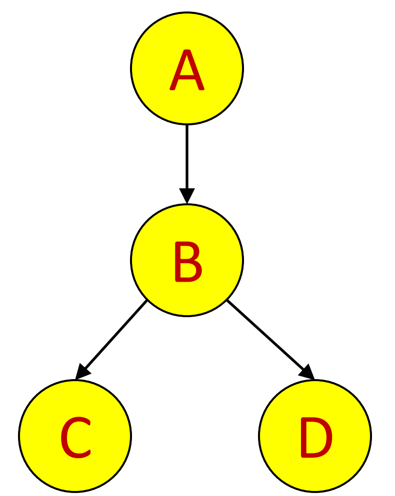
P(A = true, B = true, C = true, D = true)

= P(A = true) * P(B = true | A = true) *

P(C = true | B = true) P(D = true | B = true)

= (0.4)*(0.3)*(0.1)*(0.95)

Οι τιμές αυτές προκύπτουν από τους CDTs



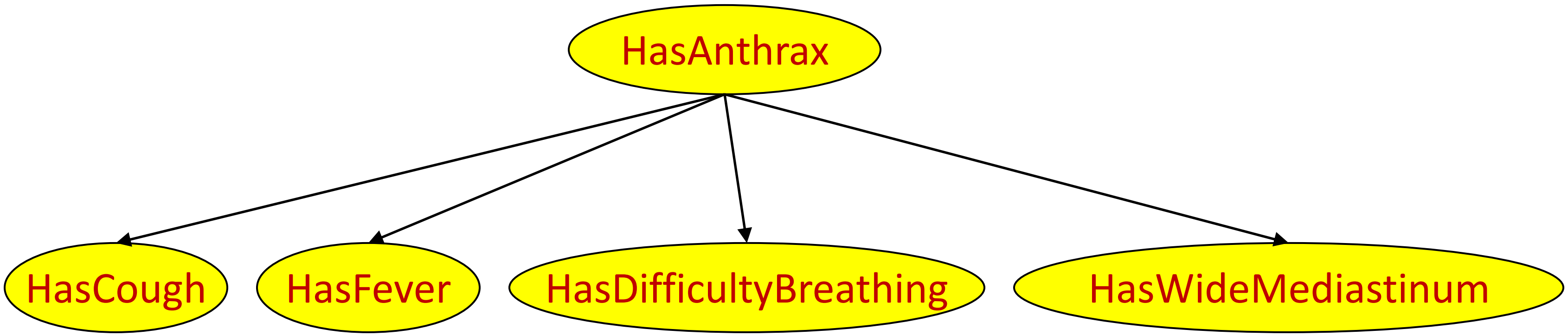
Χρήση Bayesian δικτύων για επαγωγή γνώσης

- Η χρήση ενός Bayesian δικτύου για τον υπολογισμό πιθανοτήτων
- Σε γενικές γραμμές, η επαγωγή εμπλέκει ερωτήματα της μορφής: $P(X \mid E)$

E = Οι γνωστές μεταβλητές

X = Οι μεταβλητές των ερωτημάτων

Επαγωγή γνώσης – Παράδειγμα



- Παράδειγμα ερωτήματος:

  P(*HasAnthrax = true* | *HasFever = true*, *HasCough = true*)

- Σημείωση: Παρότι τα *HasDifficultyBreathing* και *HasWideMediastinum* βρίσκονται στο δίκτυο, δε τους δίνονται τιμές στο ερώτημα (δηλ. δεν χρησιμοποιούνται ούτε ως μεταβλητές ερωτήματος ούτε ως γνωστές μεταβλητές) - θεωρούνται ως μη παρατηρημένες

Πηγές

- Introduction to Data Mining, Tan, Steinbach, Kumar.