Κατανομές Πιθανότητας

Τυχαίες Μεταβλητές

- Μια τυχαία μεταβλητή x παίρνει τιμές σε ένα συγκεκριμένο σύνολο με διαφορετική πιθανότητα
- Παράδειγμα: στη ρίψη ενός ζαριού, το αποτέλεσμα είναι τυχαίο και υπάρχουν έξι πιθανά αποτελέσματα/τιμές με πιθανότητα η κάθε μια ίση με 1/6 (ομοιόμορφη κατανομή)
- Παράδειγμα: αν πραγματοποιήσουμε μια δημοσκόπηση σε ανθρώπους για τις προτιμήσεις τους σε ένα συγκεκριμένο ζήτημα, το ποσοστό του δείγματος που θα απαντήσουν Ναι σε μια ερώτηση αποτελεί μια τυχαία μεταβλητή (το ποσοστό θα είναι κάθε φορά ελαφρώς διαφορετικό να εκτελέσουμε τη δημοσκόπηση πολλές φορές)
- Η πιθανότητα μπορεί να οριστεί σαν τη συχνότητα που αναμένουμε να συμβούν τα διαφορετικά αποτελέσματα αν επαναλάβουμε πολλές φορές ένα πείραμα (“frequentist” view)

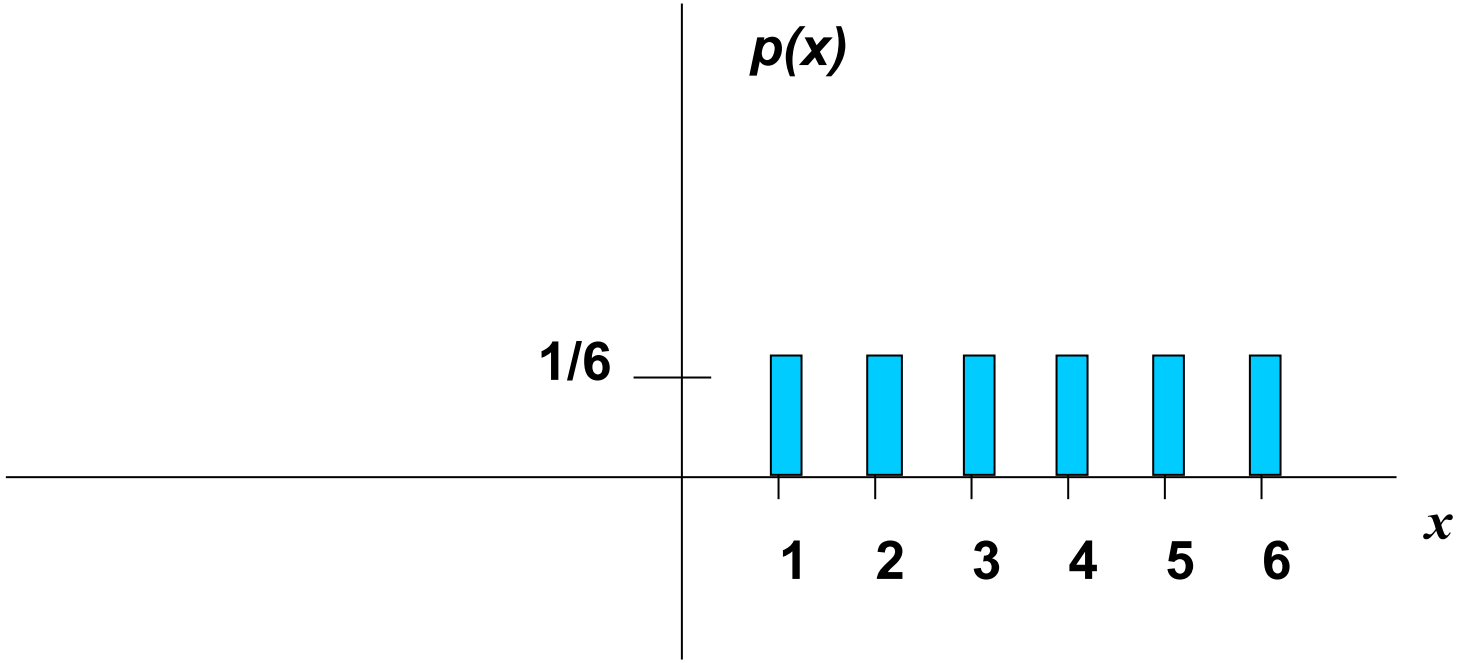
Τυχαίες Μεταβλητές

- Οι διακριτές τυχαίες μεταβλητές έχουν ένα συγκεκριμένο μετρήσιμο αριθμό διαφορετικών αποτελεσμάτων
  - Παραδείγματα: ζάρι, μετρητές, κ.λπ.
- Οι συνεχείς τυχαίες μεταβλητές έχουν ένα άπειρο αριθμό τιμών που μπορούν να πάρουν
  - Παραδείγματα: πίεση αίματος, βάρος, ταχύτητα ενός αυτοκινήτου, οποιαδήποτε μεταβλητή που παίρνει πραγματικές τιμές.

Συναρτήσεις

- Μια συνάρτηση πιθανότητας αντιστοιχεί κάθε πιθανή τιμή μιας τυχαίας μεταβλητής x σε αριθμητικές τιμές πιθανοτήτων p(x)
- Η p(x) είναι ένας αριθμός μεταξύ 0 & 1
- Η περιοχή κάτω από τη συνάρτηση πιθανότητας είναι πάντα 1

Ρίψη Ζαριού



$$\sum_{\text{all } x} P(x) = 1$$

CDF

- Cumulative Distribution Function

$$F_X(x) = \sum_{x_k \le x} P_X(x_k)$$

x	$P(x \le A)$
1	$P(x \le 1) = 1/6$
2	$P(x \le 2) = 2/6$
3	$P(x \le 3) = 3/6$
4	$P(x \le 4) = 4/6$
5	$P(x \le 5) = 5/6$
6	$P(x \le 6) = 6/6$

Διωνυμική Κατανομή

- Binomial Distribution
  - n δειγματοληψίες από μια κατανομή Bernoulli
  - Η τυχαία διωνυμική κατανομή X απεικονίζει το πλήθος των φορών στις οποίες το πείραμα είναι επιτυχές

$$\Pr(X = x) = p_\theta(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

  - E[X] = np, Var(X) = np(1-p)

Poisson

- Poisson Distribution
  - Μπορεί να εξαχθεί από τη διωνυμική κατανομή
  - Η αναμενόμενη τιμή είναι λ=np
  - Θεωρούμε ότι n → ∞
  - Μια διωνυμική κατανομή μετατρέπεται σε Poisson κατανομή

$$\Pr(X = x) = p_\theta(x) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

  - E[X] = λ, Var(X) = λ

Συνεχείς Μεταβλητές

- Μια συνάρτηση πιθανότητας που συνοδεύει μια συνεχή τυχαία μεταβλητή, είναι μια συνεχής μαθηματική συνάρτηση η οποία ολοκληρώνει στο 1
- Οι πιθανότητες που σχετίζονται με συνεχείς συναρτήσεις είναι οι περιοχές κάτω από την καμπύλη της συνάρτησης (ολοκλήρωμα).
- Οι πιθανότητες δίνονται σε μια περιοχή τιμών παρά σε συγκεκριμένες τιμές (π.χ. η πιθανότητα να πάρουμε βαθμό στο μάθημα από 5 μέχρι 10 είναι 30%).

Παράδειγμα

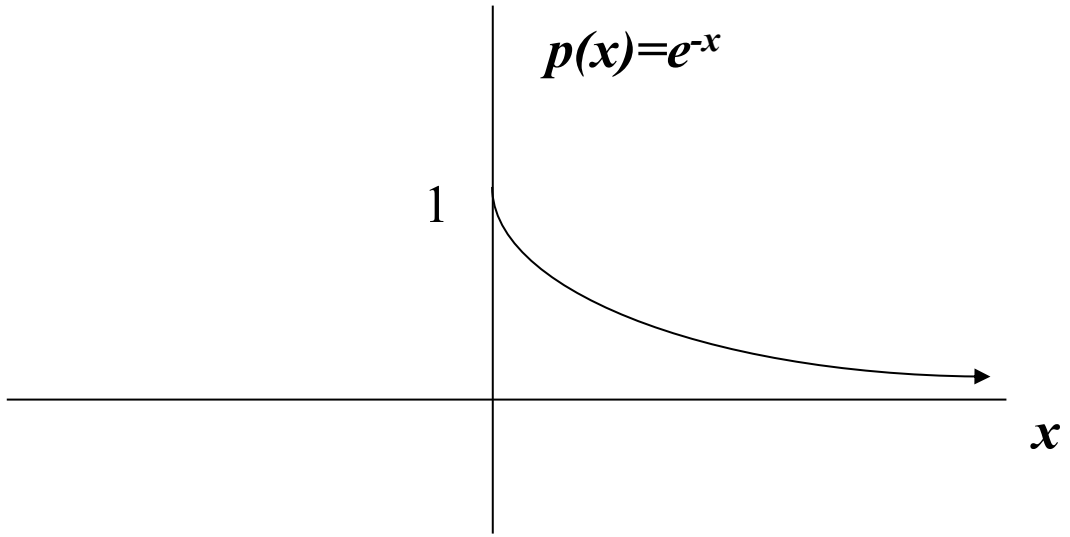
- Εκθετική συνάρτηση που ολοκληρώνει στο 1

$$f(x) = e^{-x}$$

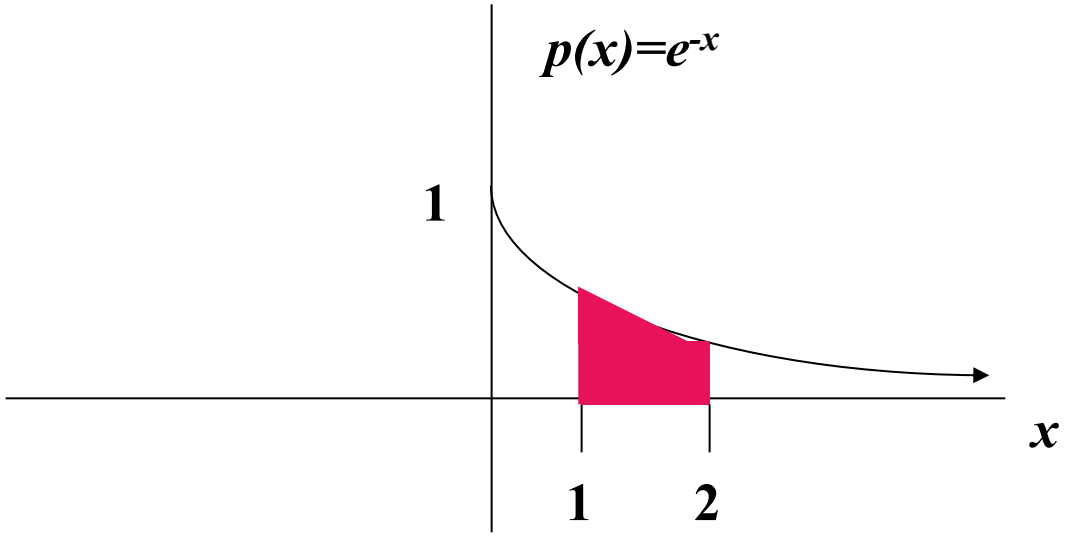
$$\int_0^{+\infty} e^{-x} = -e^{-x} \Big|_0^{+\infty} = 0 + 1 = 1$$

PDF

- Η συνάρτηση πυκνότητας πιθανότητας (probability density function - pdf) απεικονίζει ότι την πιθανότητα η μεταβλητή x να είναι μια οποιαδήποτε συγκεκριμένη τιμή (π.χ. 1.9976)



Παράδειγμα



$$P(1 \le x \le 2) = \int_1^2 e^{-x} = -e^{-x} \Big|_1^2 = -e^{-2} - -e^{-1} = -.135 + .368 = .23$$

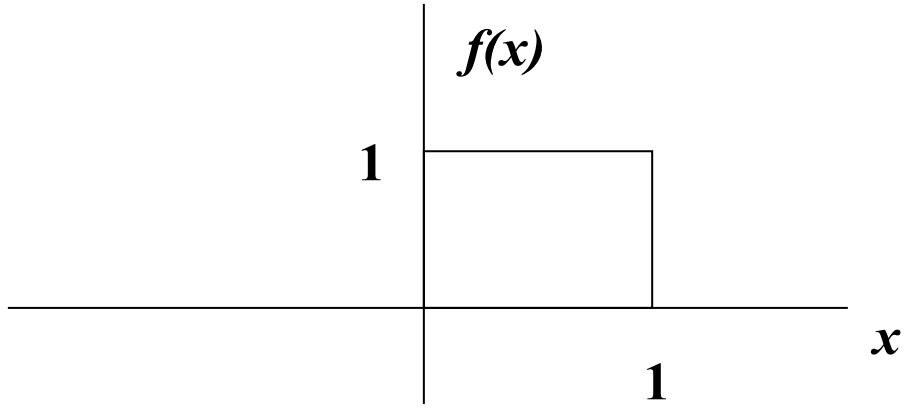
CDF

- CDF=P(x≤A)

$$\int_0^A e^{-x} = -e^{-x} \Big|_0^A = -e^{-A} - -e^{0} = -e^{-A} + 1 = 1 - e^{-A}$$

Ομοιόμορφη Κατανομή

- Ομοιόμορφη κατανομή (Uniform distribution): όλες οι τιμές έχουν την ίδια πιθανότητα
- f(x)= 1, for 1 ≥ x ≥ 0



- Παρατηρούμε ότι πρόκειται για κατανομή πιθανότητας αφού ολοκληρώνει στο 1 – η περιοχή κάτω από τη συνάρτηση είναι 1

$$\int_0^1 1 = x \Big|_0^1 = 1 - 0 = 1$$

Κανονική Κατανομή

- X~N(μ,σ)

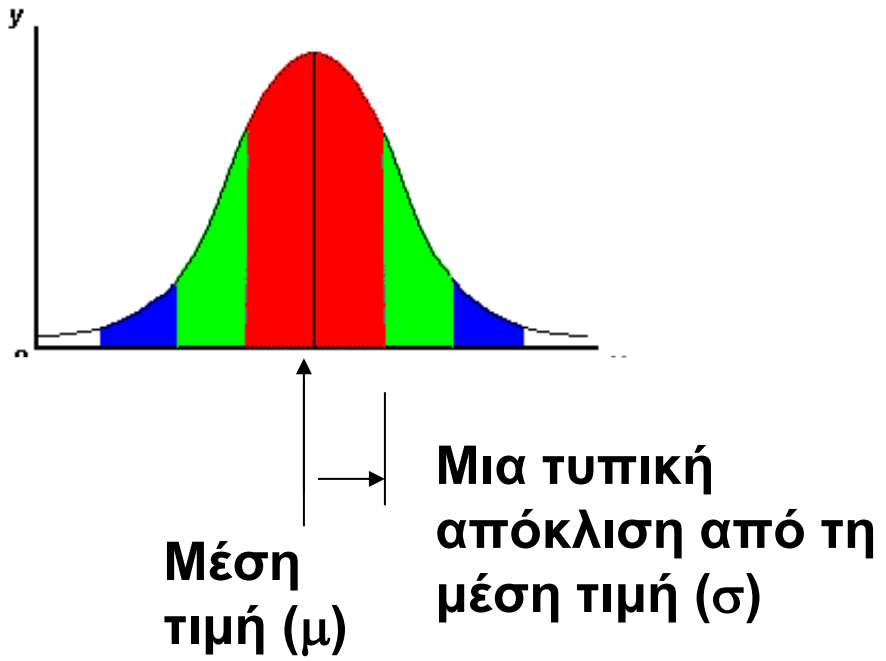
$$p_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$$\Pr(a \le X \le b) = \int_a^b p_\theta(x)dx = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}dx$$

- $E[X] = \mu$, $Var(X) = \sigma^2$

Μετρικές

- Όλες οι κατανομές πιθανότητες χαρακτηρίζονται από τη μέση τιμή / αναμενόμενη τιμή και τη διακύμανσή τους / τυπική απόκλισή τους



Μέση Τιμή

- Αν καταλάβουμε την συνάρτηση πυκνότητας πιθανότητας μιας κατανομής, τότε μπορούμε να πάρουμε αποφάσεις όσον αφορά στο πως περιμένουμε να συμπεριφερθεί η τυχαία μεταβλητή στο μέσο όρο και σε μακροπρόθεσμο ορίζοντα (“frequentist” theory of probability).

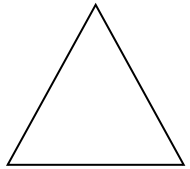
- Η αναμενόμενη τιμή είναι η βεβαρημένη μέση τιμή (μ) μιας τυχαίας μεταβλητής

$$\mathrm{E}[X] = \sum_{i=1}^{k} x_i\, p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$

$$\mathrm{E}[X] = \int_{\mathbb{R}} x f(x)\, dx$$

Παράδειγμα

x	10	11	12	13	14
P(x)	.4	.2	.2	.1	.1



$$\sum_{i=1}^{5} x_i\, p(x) = 10(.4) + 11(.2) + 12(.2) + 13(.1) + 14(.1) = 11.3$$

Πράξεις

- Αν c είναι ένας σταθερός αριθμός και X, Y είναι τυχαίες μεταβλητές τότε ισχύουν τα ακόλουθα:
- $E(c) = c$
- $E(cX) = cE(X)$
- $E(c + X) = c + E(X)$
- $E(X+Y) = E(X) + E(Y)$

Διακύμανση/Απόκλιση

“Η μέση τετραγωνική απόσταση από τη μέση τιμή”

$$\sigma^2 = Var(x) = E[(x-\mu)^2] = \sum_{\text{all x}} (x_i - \mu)^2 \, p(x_i)$$

Διακύμανση/Απόκλιση

Διακριτή μεταβλητή

$$Var(X) = \sigma^2 = \sum_{\text{all x}} (x_i - \mu)^2 \, p(x_i)$$

Συνεχής μεταβλητή

$$Var(X) = \sigma^2 = \int_{-\infty}^{\infty} (x_i - \mu)^2 \, p(x_i) dx$$

Διακύμανση/Απόκλιση

- Η διακύμανση του δείγματος δίνεται από:

$$\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{n-1} = \sum_{i=1}^{N}(x_i - \bar{x})^2 \left(\frac{1}{n-1}\right)$$

Υπό συνθήκη

- Αν A και B είναι συμβάντα με Pr(A) > 0, η πιθανότητα υπό συνθήκη του B δεδομένου του A είναι

$$\Pr(B \mid A) = \frac{\Pr(AB)}{\Pr(A)}$$

- Παράδειγμα: ένα τεστ

	Women	Men
Success	200	1800
Failure	1800	200

	Women		Men	
	Drug I	Drug II	Drug I	Drug II
Success	200	10	19	1000
Failure	1800	190	1	1000

Κανόνας Bayes

- Δοσμένων δύο μεταβλητών A και B με Pr(A) > 0, έχουμε:

$$\Pr(B \mid A) = \frac{\Pr(AB)}{\Pr(A)} = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A)}$$

- Παράδειγμα:

Pr(R) = 0.8

Pr(W)= 0.8

R: It is a rainy day

W: The grass is wet

Pr(R|W) = ?

Pr(W|R)	R	¬R
W	0.7	0.4
¬W	0.3	0.6

Μάθηση υπό Επίβλεψη

Παραδείγματα

- Ασθενείς που έρχονται για θεραπεία και για τους οποίους παρακολουθούμε π.χ. 17 μεταβλητές (blood pressure, age, etc)
- Απόφαση: να θα εισαχθούν σε μονάδα εντατικής θεραπείας
- Προτεραιότητα με βάση τις ανάγκες τους
- Πρόβλεψη: υψηλού κινδύνου ασθενείς – χαμηλού κινδύνου ασθενείς

Παραδείγματα

- Μια μεγάλη τράπεζα λαμβάνει χιλιάδες αιτήσεις για νέες πιστωτικές κάρτες
- Κάθε αίτηση περιέχει πληροφορίες σχετικά με τον αιτούντα, την ηλικία, την οικογενειακή κατάσταση, τον ετήσιο μισθό, τις εκκρεμείς οφειλές, την πιστοληπτική ικανότητα κ.λπ.
- Πρόβλημα: να αποφασίσουμε εάν μια αίτηση πρέπει να εγκριθεί ή να ταξινομήσουμε τις αιτήσεις σε δύο κατηγορίες, εγκεκριμένες και μη εγκεκριμένες

Εστίαση

- Ένας υπολογιστής δεν έχει «εμπειρίες».
- Ένα σύστημα υπολογιστή μαθαίνει από δεδομένα, τα οποία αντιπροσωπεύουν κάποιες «προηγούμενες εμπειρίες» ενός τομέα εφαρμογής.
- Η εστίασή μας: **να μάθουμε μια συνάρτηση στόχο που μπορεί να χρησιμοποιηθεί για την πρόβλεψη τιμών ενός διακριτού χαρακτηριστικού κλάσης**, π.χ. έγκριση ή μη εγκεκριμένη και υψηλού ή χαμηλού κινδύνου.
- Η εργασία ονομάζεται συνήθως: Εποπτευόμενη μάθηση, ταξινόμηση ή επαγωγική μάθηση.

Δεδομένα

- Δεδομένα: Ένα σύνολο εγγραφών δεδομένων (ονομάζονται επίσης παραδείγματα, στιγμιότυπα ή περιπτώσεις) που περιγράφονται από
- k ιδιότητες/μεταβλητές: A_1, A_2, … A_k.
- Κλάση: Κάθε παράδειγμα επισημαίνεται με μια προκαθορισμένη κλάση.
- Στόχος: Να μάθουν ένα μοντέλο κατηγοριοποίησης από τα δεδομένα που μπορούν να χρησιμοποιηθούν για την πρόβλεψη των κατηγοριών νέων (μελλοντικών ή δοκιμαστικών) περιπτώσεων/περιπτώσεων.

Παράδειγμα

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	**No**
2	young	false	false	good	**No**
3	young	true	false	good	**Yes**
4	young	true	true	fair	**Yes**
5	young	false	false	fair	**No**
6	middle	false	false	fair	**No**
7	middle	false	false	good	**No**
8	middle	true	true	good	**Yes**
9	middle	false	true	excellent	**Yes**
10	middle	false	true	excellent	**Yes**
11	old	false	true	excellent	**Yes**
12	old	false	true	good	**Yes**
13	old	true	false	good	**Yes**
14	old	true	false	excellent	**Yes**
15	old	false	false	fair	**No**

Μάθηση

- Να μάθουμε ένα μοντέλο κατηγοριοποίησης από τα δεδομένα
- Παράδειγμα: Χρήση του μοντέλου για να κατηγοριοποιήσουμε τις μελλοντικές αιτήσεις δανείου σε

  Ναι (εγκρίθηκε) ή

  Όχι (δεν εγκρίθηκε)

  Ποια είναι η τάξη για την παρακάτω περίπτωση/περίπτωση;

Age	Has_Job	Own_house	Credit-Rating	Class
young	false	false	good	?

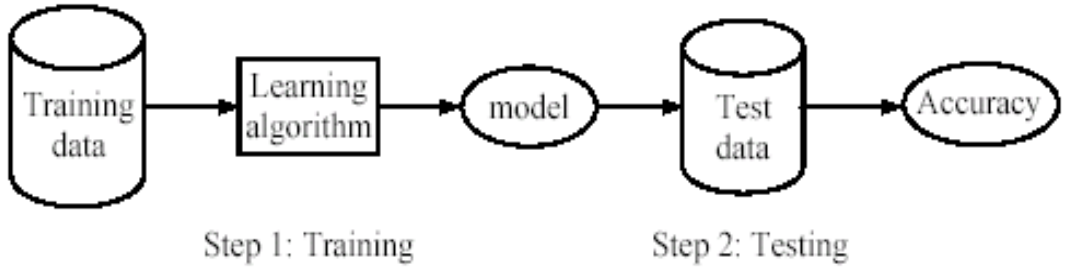
Είδη Μάθησης

- Εποπτευόμενη μάθηση: η κατηγοριοποίηση θεωρείται ως εποπτευόμενη μάθηση από παραδείγματα.
  - Επίβλεψη: Τα δεδομένα (παρατηρήσεις, μετρήσεις κ.λπ.) επισημαίνονται με προκαθορισμένες κλάσεις. Είναι σαν ένας «δάσκαλος» να δίνει τα μαθήματα (επίβλεψη).
  - Τα δεδομένα δοκιμής ταξινομούνται επίσης σε αυτές τις κατηγορίες.
- Εκμάθηση χωρίς επίβλεψη (ομαδοποίηση)
  - Οι ετικέτες κλάσεων των δεδομένων είναι άγνωστες
  - Με ένα σύνολο δεδομένων, η εργασία είναι να διαπιστωθεί η ύπαρξη κλάσεων ή συστάδων στα δεδομένα

Βήματα

- Μάθηση (εκπαίδευση): Να μάθουμε ένα μοντέλο χρησιμοποιώντας τα δεδομένα εκπαίδευσης
- Δοκιμή/Αποτίμηση: Δοκιμάζουμε το μοντέλο χρησιμοποιώντας δεδομένα δοκιμής που δεν έχουν υιοθετηθεί κατά τη διαδικασία της εκπαίδευσης για να αξιολογήσουμε την ακρίβειά του

$$Accuracy = \frac{\text{Number of correct classifications}}{\text{Total number of test cases}},$$



Πρόβλημα

Δεδομένων

- ένος σύνολο δεδομένων D,
- μιας εργασία T, και
- ένα μέτρο απόδοσης M,

ένα σύστημα υπολογιστή λέγεται ότι μαθαίνει από το D για να εκτελέσει την εργασία T εάν μετά την εκμάθηση η απόδοση του συστήματος στο T βελτιωθεί όπως μετράται από τον M.

Με άλλα λόγια, το μοντέλο εκμάθησης βοηθά το σύστημα να αποδώσει καλύτερα το T σε σύγκριση με την απουσία μάθησης.

Υποθέσεις

- **Υπόθεση**: Η κατανομή των παραδειγμάτων εκπαίδευσης είναι πανομοιότυπη με την κατανομή των παραδειγμάτων δοκιμής (συμπεριλαμβανομένων μελλοντικών παραδειγμάτων που δεν έχουμε δει).
- Στην πράξη, αυτή η υπόθεση συχνά παραβιάζεται σε κάποιο βαθμό.
- Οι σοβαρές παραβιάσεις θα οδηγήσουν σαφώς σε χαμηλή ακρίβεια κατηγοριοποίησης.
- Για να επιτευχθεί καλή ακρίβεια στα δεδομένα της δοκιμής, τα παραδείγματα εκπαίδευσης πρέπει να είναι επαρκώς αντιπροσωπευτικά των δεδομένων δοκιμής.

Κατηγοριοποίηση

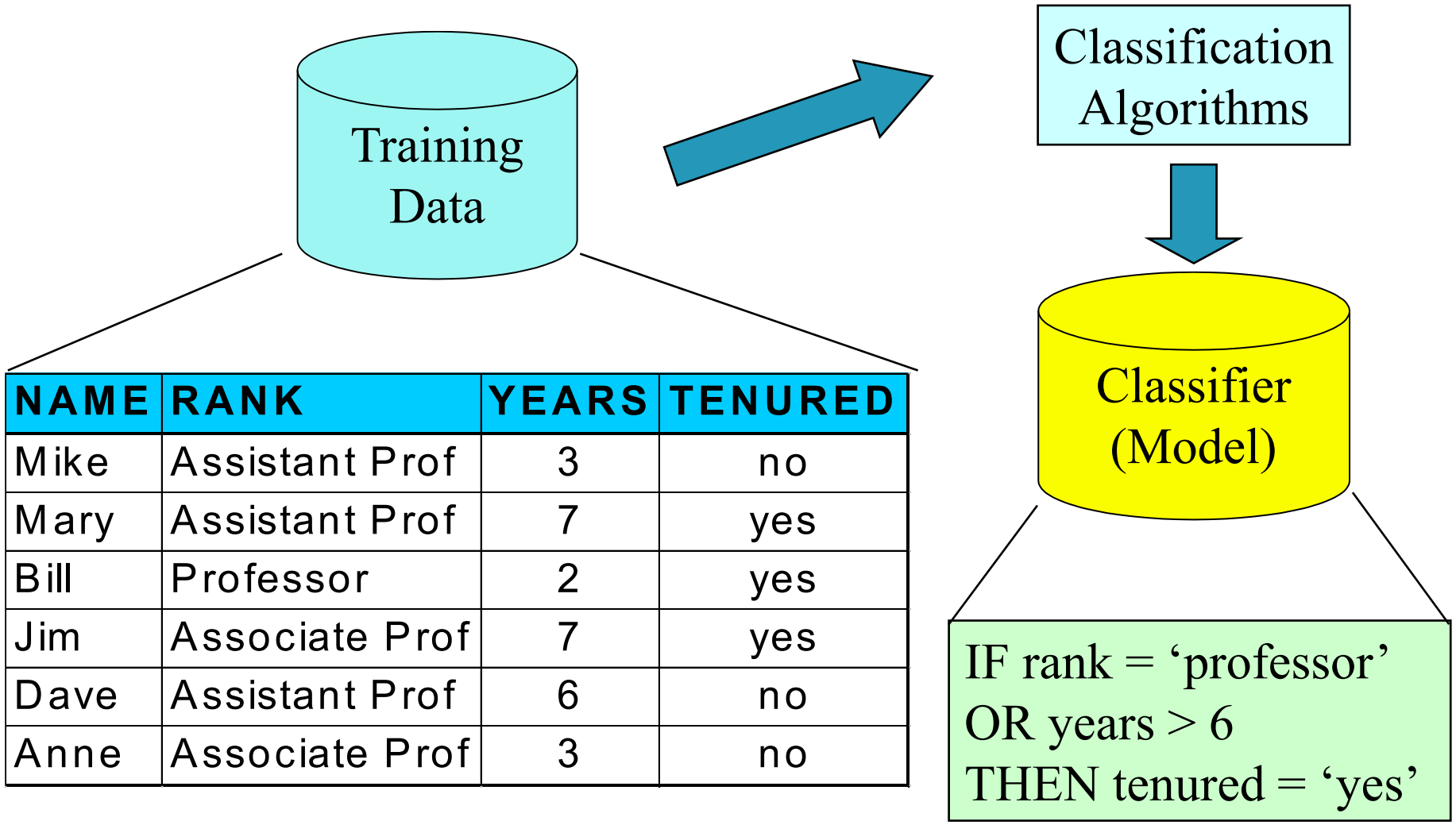
Σύγκριση

- Κατηγοριοποίηση
  - προβλέπει ετικέτες κατηγοριών (διακριτές ή ονομαστικές)
  - ταξινομεί δεδομένα (κατασκευάζει ένα μοντέλο) με βάση το σύνολο εκπαίδευσης και τις τιμές (ετικέτες κλάσεων) σε ένα χαρακτηριστικό κατηγοριοποίησης και το χρησιμοποιεί για την κατηγοριοποίηση νέων δεδομένων
- Αριθμητική Πρόβλεψη
  - μοντελοποιεί συναρτήσεις συνεχούς τιμής, δηλ. προβλέπει άγνωστες τιμές ή τιμές που λείπουν
- Τυπικές εφαρμογές
  - Έγκριση πίστωσης/δανείου
  - Ιατρική διάγνωση: εάν ένας όγκος είναι καρκινικός
  - Ανίχνευση απάτης: εάν μια συναλλαγή είναι δόλια
  - Κατηγοριοποίηση ιστοσελίδων: σε ποια κατηγορία ανήκει

Βήματα

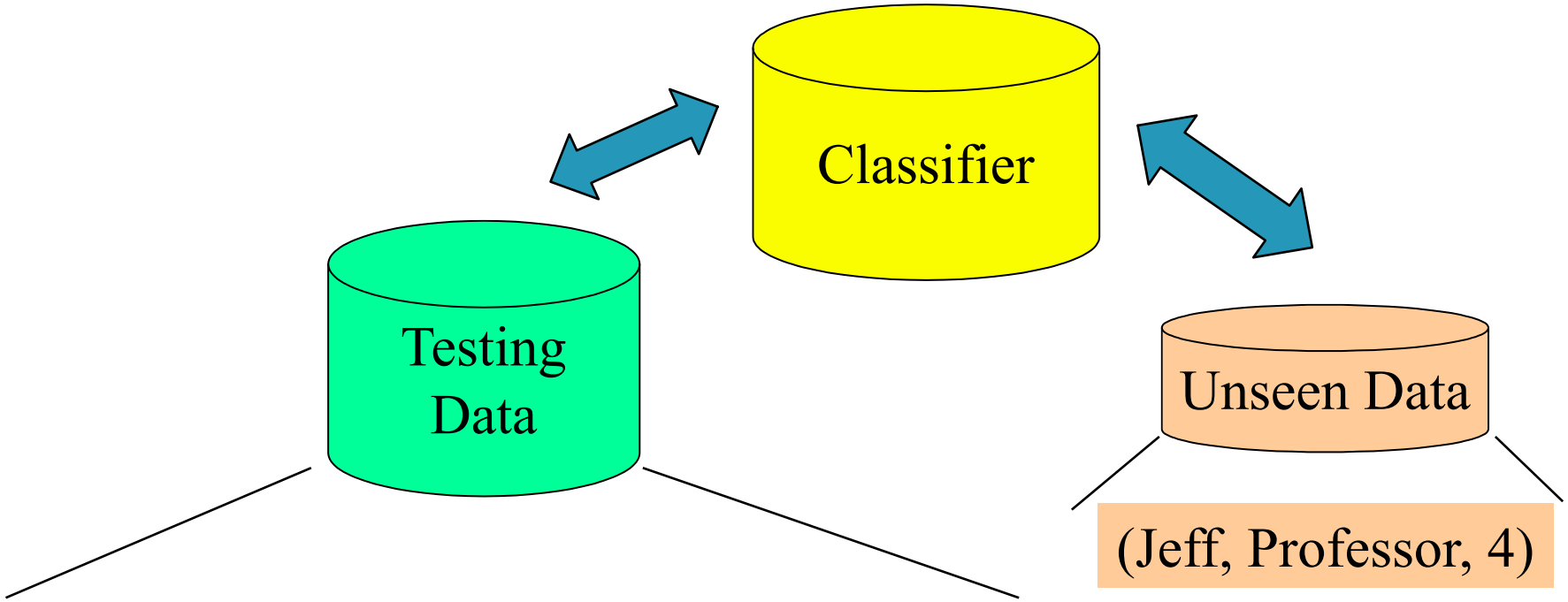
- Κατασκευή μοντέλου: περιγράφει ένα σύνολο προκαθορισμένων κλάσεων
  - Κάθε πλειάδα/δείγμα θεωρείται ότι ανήκει σε μια προκαθορισμένη κλάση, όπως καθορίζεται από το χαρακτηριστικό class label
  - Το σετ των πλειάδων που χρησιμοποιούνται για την κατασκευή μοντέλων είναι σετ εκπαίδευσης
  - Το μοντέλο αναπαρίσταται ως κανόνες κατηγοριοποίησης, δέντρα αποφάσεων ή μαθηματικούς τύπους
- Χρήση μοντέλου: για ταξινόμηση μελλοντικών ή άγνωστων αντικειμένων
  - Εκτίμηση της ακρίβειας του μοντέλου
    - Η γνωστή ετικέτα του δείγματος δοκιμής συγκρίνεται με το ταξινομημένο αποτέλεσμα από το μοντέλο
    - Το ποσοστό ακρίβειας είναι το ποσοστό των δειγμάτων του συνόλου δοκιμής που ταξινομούνται σωστά από το μοντέλο
    - Το σετ δοκιμής είναι ανεξάρτητο από το σετ εκπαίδευσης (διαφορετικά υπερπροσαρμογή - overfitting)
  - Εάν η ακρίβεια είναι αποδεκτή, χρησιμοποιούμε το μοντέλο για να κατηγοριοποιήσουμε νέα δεδομένα
- Σημείωση: Εάν το σύνολο δοκιμής χρησιμοποιείται για την επιλογή μοντέλων, ονομάζεται σύνολο επικύρωσης (δοκιμής).

Κατασκευή



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

Χρήση



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

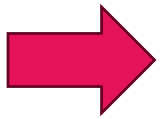
Δ έ ν δ ρ α
Α π ο φ ά σ ε ω ν

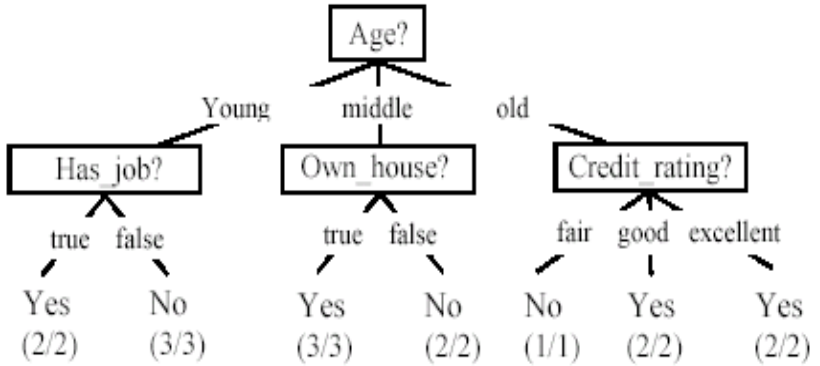
Εισαγωγή

- Η εκμάθηση δέντρων αποφάσεων (decision trees) είναι μια από τις πιο ευρέως χρησιμοποιούμενες τεχνικές κατηγοριοποίησης.
- Η ακρίβεια ταξινόμησης του είναι ανταγωνιστική με άλλες μεθόδους και είναι πολύ αποτελεσματική.
- Το μοντέλο κατηγοριοποίησης είναι ένα δέντρο, που ονομάζεται δέντρο αποφάσεων.
- Ο αλγόριθμος C4.5 του Ross Quinlan είναι ίσως το πιο γνωστό σύστημα.

Παράδειγμα

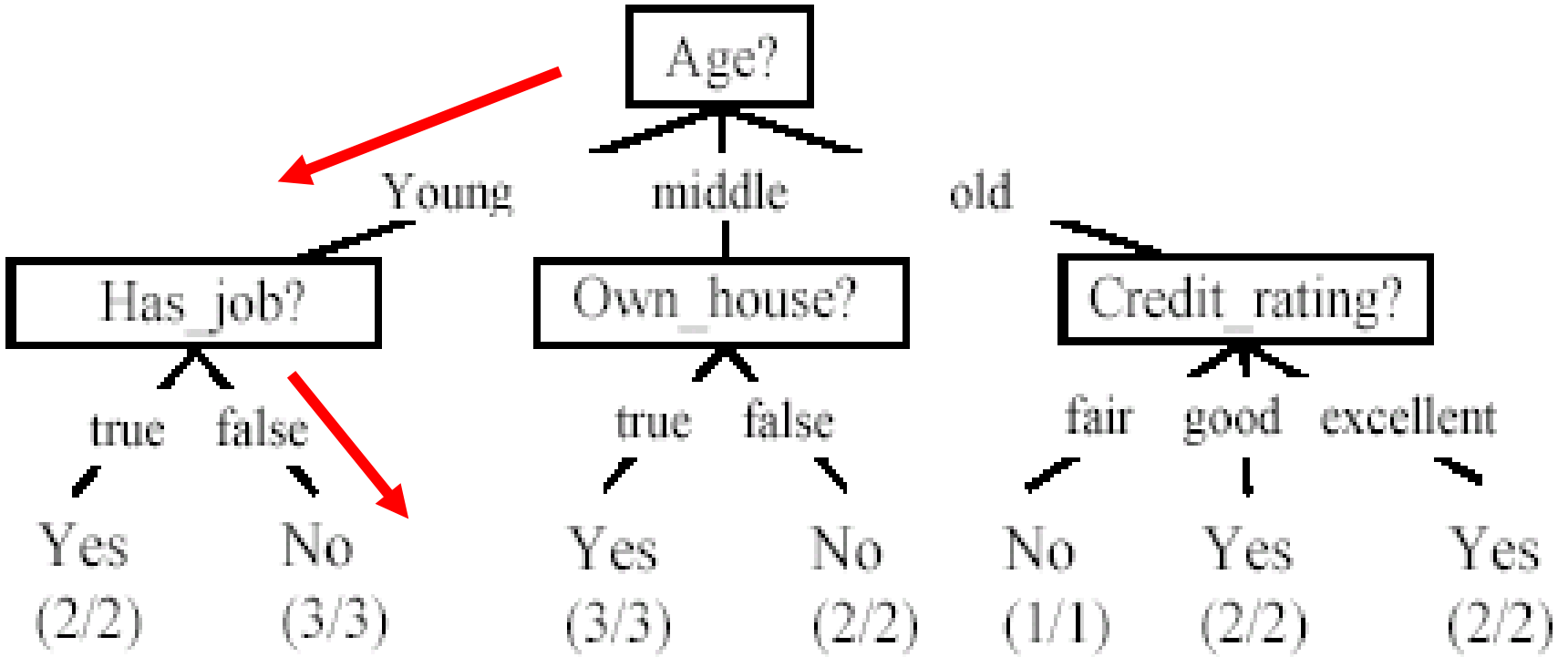
ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	**No**
2	young	false	false	good	**No**
3	young	true	false	good	**Yes**
4	young	true	true	fair	**Yes**
5	young	false	false	fair	**No**
6	middle	false	false	fair	**No**
7	middle	false	false	good	**No**
8	middle	true	true	good	**Yes**
9	middle	false	true	excellent	**Yes**
10	middle	false	true	excellent	**Yes**
11	old	false	true	excellent	**Yes**
12	old	false	true	good	**Yes**
13	old	true	false	good	**Yes**
14	old	true	false	excellent	**Yes**
15	old	false	false	fair	**No**





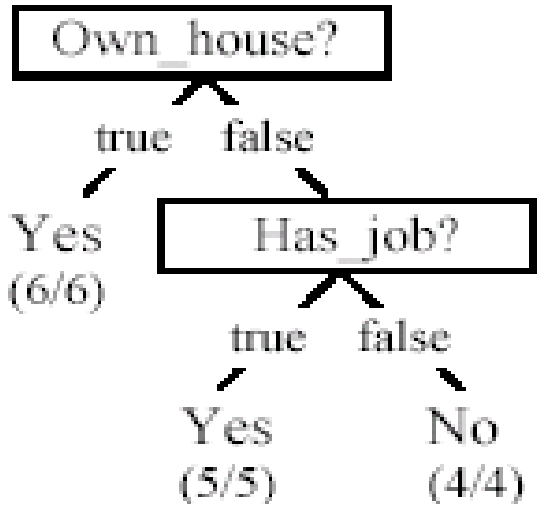
Χρήση

Age	Has_Job	Own_house	Credit-Rating	Class
young	false	false	good	?



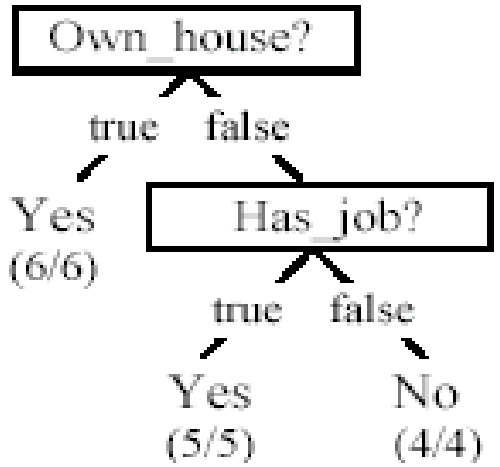
Μοναδικότητα

- Θέλουμε το μικρότερο δέντρο και ακριβές δέντρο.
  - Εύκολη κατανόηση και καλύτερη απόδοση.
- Η εύρεση του καλύτερου δέντρου είναι NP-hard.
- Όλοι οι τρέχοντες αλγόριθμοι δημιουργίας δέντρων είναι ευρετικοί αλγόριθμοι



Κανόνες

- Ένα δέντρο αποφάσεων μπορεί να μετατραπεί σε ένα σύνολο κανόνων
- Κάθε διαδρομή από τη ρίζα σε ένα φύλλο είναι ένας κανόνας



Own_house = true → Class =Yes [sup=6/15, conf=6/6]
Own_house = false, Has_job = true → Class = Yes [sup=5/15, conf=5/5]
Own_house = false, Has_job = false → Class = No [sup=4/15, conf=4/4]

Πρόβλημα

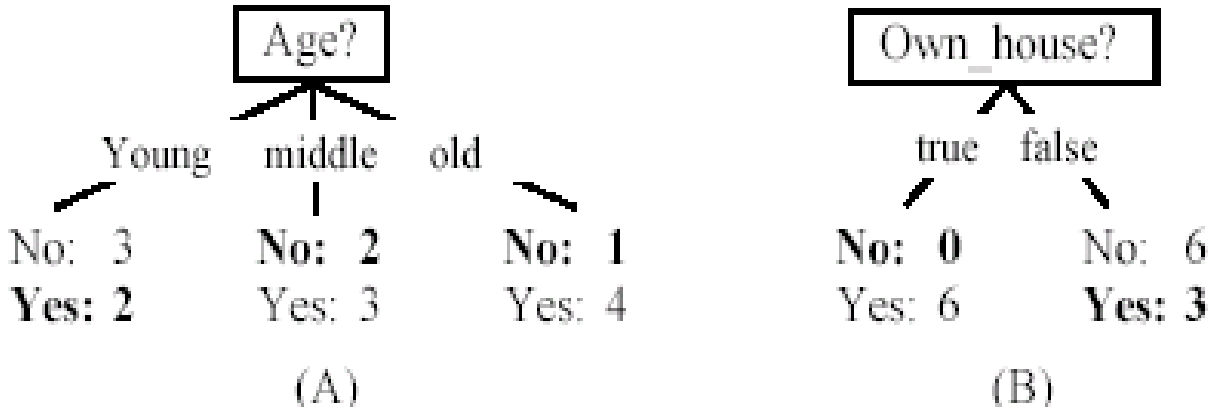
- Βασικός αλγόριθμος (ένας άπληστος αλγόριθμος διαίρει και βασίλευε)
  - Ας υποθέσουμε ότι τα χαρακτηριστικά είναι categorical (οι αλγόριθμοι μπορούν να χειριστούν και τα συνεχή χαρακτηριστικά)
  - Το δέντρο κατασκευάζεται με αναδρομικό τρόπο από πάνω προς τα κάτω
  - Στην αρχή, όλα τα παραδείγματα εκπαίδευσης βρίσκονται/αντιστοιχούν στη ρίζα
  - Τα παραδείγματα χωρίζονται αναδρομικά με βάση τα επιλεγμένα χαρακτηριστικά
  - Τα χαρακτηριστικά επιλέγονται με βάση μια συνάρτηση (π.χ. κέρδος πληροφορίας – information gain)
- Προϋποθέσεις διακοπής της κατάτμησης
  - Όλα τα παραδείγματα για έναν δεδομένο κόμβο ανήκουν στην ίδια κλάση
  - Δεν υπάρχουν εναπομείναντα χαρακτηριστικά για περαιτέρω κατάτμηση – η πλειοψηφική τάξη είναι το φύλλο
  - Δεν έχουν μείνει παραδείγματα

Purity

- Το κλειδί για τη δημιουργία ενός δέντρου αποφάσεων - ποιο χαρακτηριστικό να επιλέξετε για να διακλαδώσετε.
- Ο στόχος είναι να μειωθεί όσο το δυνατόν περισσότερο η αβεβαιότητα στα δεδομένα.
- **Ένα υποσύνολο δεδομένων είναι καθαρό/αγνό εάν όλα τα στιγμιότυπα ανήκουν στην ίδια κλάση.**
- Η ευρετική στο C4.5 είναι να επιλέξετε το χαρακτηριστικό με το μέγιστο κέρδος πληροφοριών (Information Gain) ή αναλογία κέρδους (Gain ratio) με βάση τη θεωρία πληροφοριών.

Παράδειγμα

- Το Β φαίνεται καλύτερο



Θεωρία Πληροφορίας

- Η θεωρία πληροφοριών παρέχει μια μαθηματική βάση για τη μέτρηση του περιεχομένου των πληροφοριών.
- Για να κατανοήσετε την έννοια της πληροφορίας, σκεφτείτε ότι παρέχει την απάντηση σε μια ερώτηση, για παράδειγμα, εάν ένα νόμισμα βγει ‘κεφαλή’.
- Εάν κάποιος έχει ήδη μια καλή εικασία για την απάντηση, τότε η πραγματική απάντηση είναι λιγότερο κατατοπιστική.
- Εάν κάποιος γνωρίζει ήδη ότι το κέρμα είναι στημένο έτσι ώστε να έχει κεφαλές με πιθανότητα 0.99, τότε ένα μήνυμα (προηγμένες πληροφορίες) σχετικά με το πραγματικό αποτέλεσμα μιας ρίψης αξίζει λιγότερο από ότι θα ήταν για ένα τίμιο νόμισμα (50-50).
- Για ένα δίκαιο (τίμιο) νόμισμα, δεν έχετε πληροφορίες και είστε διατεθειμένοι να πληρώσετε περισσότερα (ας πούμε σε $) για προηγμένες πληροφορίες – όσο λιγότερα ξέρετε, τόσο πιο πολύτιμες είναι οι πληροφορίες.
- Η θεωρία πληροφοριών χρησιμοποιεί την ίδια διαίσθηση, αλλά αντί να μετράει την αξία για πληροφορίες σε δολάρια, μετρά το περιεχόμενο πληροφοριών σε bit.
- Ένα bit πληροφορίας αρκεί για να απαντήσει κανείς σε μια ερώτηση ναι/όχι για την οποία δεν έχει ιδέα, όπως η ανατροπή ενός νομίσματος

Εντροπία

- Ο τύπος της εντροπίας,

$$entropy(D) = -\sum_{j=1}^{|C|} \Pr(c_j) \log_2 \Pr(c_j)$$

$$\sum_{j=1}^{|C|} \Pr(c_j) = 1,$$

- Το $\Pr(c_j)$ είναι η πιθανότητα της κλάσης c_j στο σύνολο δεδομένων D
- Χρησιμοποιούμε την εντροπία ως μέτρο αγνότητας/καθαρότητας ή διαταραχής του συνόλου δεδομένων D (ή ένα μέτρο πληροφοριών σε ένα δέντρο απόφασης)

Παράδειγμα

- Καθώς τα δεδομένα γίνονται όλο και πιο καθαρά, η τιμή της εντροπίας γίνεται όλο και μικρότερη. Αυτό μας είναι χρήσιμο!

1. The data set D has 50% positive examples (Pr(*positive*) = 0.5) and 50% negative examples (Pr(*negative*) = 0.5).

$$entropy(D) = -0.5 \times \log_2 0.5 - 0.5 \times \log_2 0.5 = 1$$

2. The data set D has 20% positive examples (Pr(*positive*) = 0.2) and 80% negative examples (Pr(*negative*) = 0.8).

$$entropy(D) = -0.2 \times \log_2 0.2 - 0.8 \times \log_2 0.8 = 0.722$$

3. The data set D has 100% positive examples (Pr(*positive*) = 1) and no negative examples, (Pr(*negative*) = 0).

$$entropy(D) = -1 \times \log_2 1 - 0 \times \log_2 0 = 0$$

Εντροπία

- Δίνοντας ένα σύνολο παραδειγμάτων D, υπολογίζουμε πρώτα την εντροπία του:

$$entropy(D) = -\sum_{j=1}^{|C|} \Pr(c_j) \log_2 \Pr(c_j)$$

- Αν έχουμε το χαρακτηριστικό A_i, με τις v τιμές, σαν τη ρίζα του δέντρου απόφασης, αυτό θα χωρίσει το D σε v υποσύνολα D_1, D_2, …, D_v .
- Η αναμενόμενη εντροπία εάν το A_i χρησιμοποιείται ως τρέχουσα ρίζα είναι:

$$entropy_{A_i}(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times entropy(D_j)$$

Information Gain

- Οι πληροφορίες που αποκτώνται επιλέγοντας το χαρακτηριστικό A_i σε διακλάδωση ή κατάτμηση των δεδομένων είναι

$$gain(D, A_i) = entropy(D) - entropy_{A_i}(D)$$

- Επιλέγουμε το χαρακτηριστικό με το υψηλότερο κέρδος για να διαχωρίσουμε το τρέχον δέντρο.

Information Gain

- Επιλέγουμε το χαρακτηριστικό με το υψηλότερο κέρδος πληροφοριών
- Έστω p_i η πιθανότητα ότι μια αυθαίρετη πλειάδα στο D ανήκει στην κατηγορία C_i, που υπολογίζεται από |Ci, D|/|D|
- Αναμενόμενες πληροφορίες (εντροπία) που απαιτούνται για την ταξινόμηση μιας πλειάδας στο D:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

- Απαιτούνται πληροφορίες (μετά τη χρήση του A για τον διαχωρισμό του D σε v τμήματα) για την ταξινόμηση του D:

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$$

- Πληροφορίες που λαμβάνονται με διακλάδωση στο χαρακτηριστικό A

$$Gain(A) = Info(D) - Info_A(D)$$

Παράδειγμα

- Το Own_house είναι η καλύτερη επιλογή για τη χρήση του ως ρίζα

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	No
2	young	false	false	excellent	No
3	young	true	false	good	Yes
4	young	true	true	good	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

$$entropy(D) = \frac{6}{15} \times \log_2 \frac{6}{15} + \frac{9}{15} \times \log_2 \frac{9}{15} = 0.971$$

$$entropy_{Own\_house}(D) = \frac{6}{15} \times entropy(D_1) + \frac{9}{15} \times entropy(D_2)$$

$$= \frac{6}{15} \times 0 + \frac{9}{15} \times 0.918$$

$$= 0.551$$

$$entropy_{Age}(D) = \frac{5}{15} \times entropy(D_1) + \frac{5}{15} \times entropy(D_2) + \frac{5}{15} \times entropy(D_3)$$

$$= \frac{5}{15} \times 0.971 + \frac{5}{15} \times 0.971 + \frac{5}{15} \times 0.722$$

$$= 0.888$$

Age	Yes	No	entropy(Di)
young	2	3	0.971
middle	3	2	0.971
old	4	1	0.722

$gain(D, \text{Age}) = 0.971 - 0.888 = 0.083$

$gain(D, \text{Own\_house}) = 0.971 - 0.551 = 0.420$

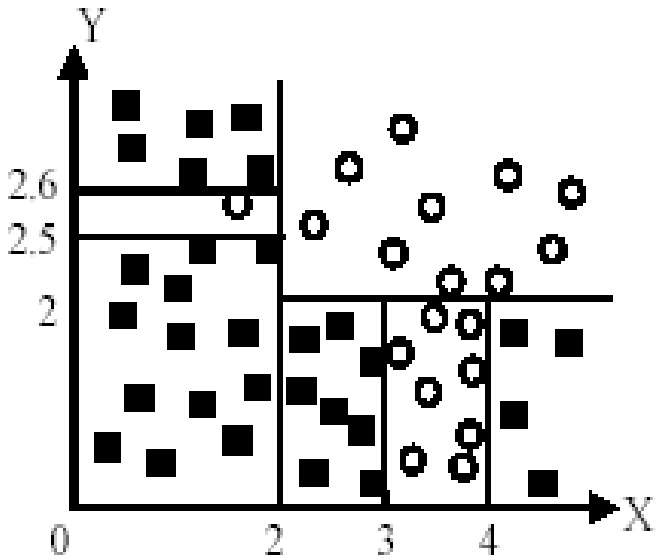
$gain(D, \text{Has\_Job}) = 0.971 - 0.647 = 0.324$

$gain(D, \text{Credit\_Rating}) = 0.971 - 0.608 = 0.363$

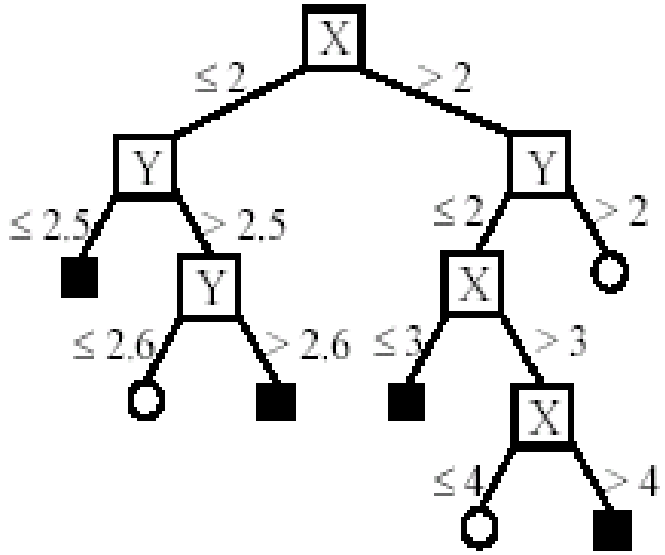
Συνεχείς Ιδιότητες

- Έστω το χαρακτηριστικό A ένα χαρακτηριστικό συνεχούς τιμής
- Πρέπει να προσδιορίσουμε το καλύτερο σημείο διαχωρισμού για το A
- Ταξινομούμε την τιμή A με αύξουσα σειρά
- Τυπικά, το μέσο σημείο μεταξύ κάθε ζεύγους γειτονικών τιμών θεωρείται ως πιθανό σημείο διαίρεσης$(a_i+a_{i+1})/2$ είναι το μέσο μεταξύ των τιμών του a_i και του a_{i+1}
- Το σημείο με την ελάχιστη αναμενόμενη απαίτηση πληροφοριών για το A επιλέγεται ως διαχωριστικό σημείο για το A
- Διαχωρισμός:
  - Το D1 είναι το σύνολο των πλειάδων στο D που ικανοποιούν το σημείο διαίρεσης A ≤ split_point και το D2 είναι το σύνολο των πλειάδων στο D που ικανοποιούν το σημείο διαίρεσης A > split_point

Συνεχείς Ιδιότητες

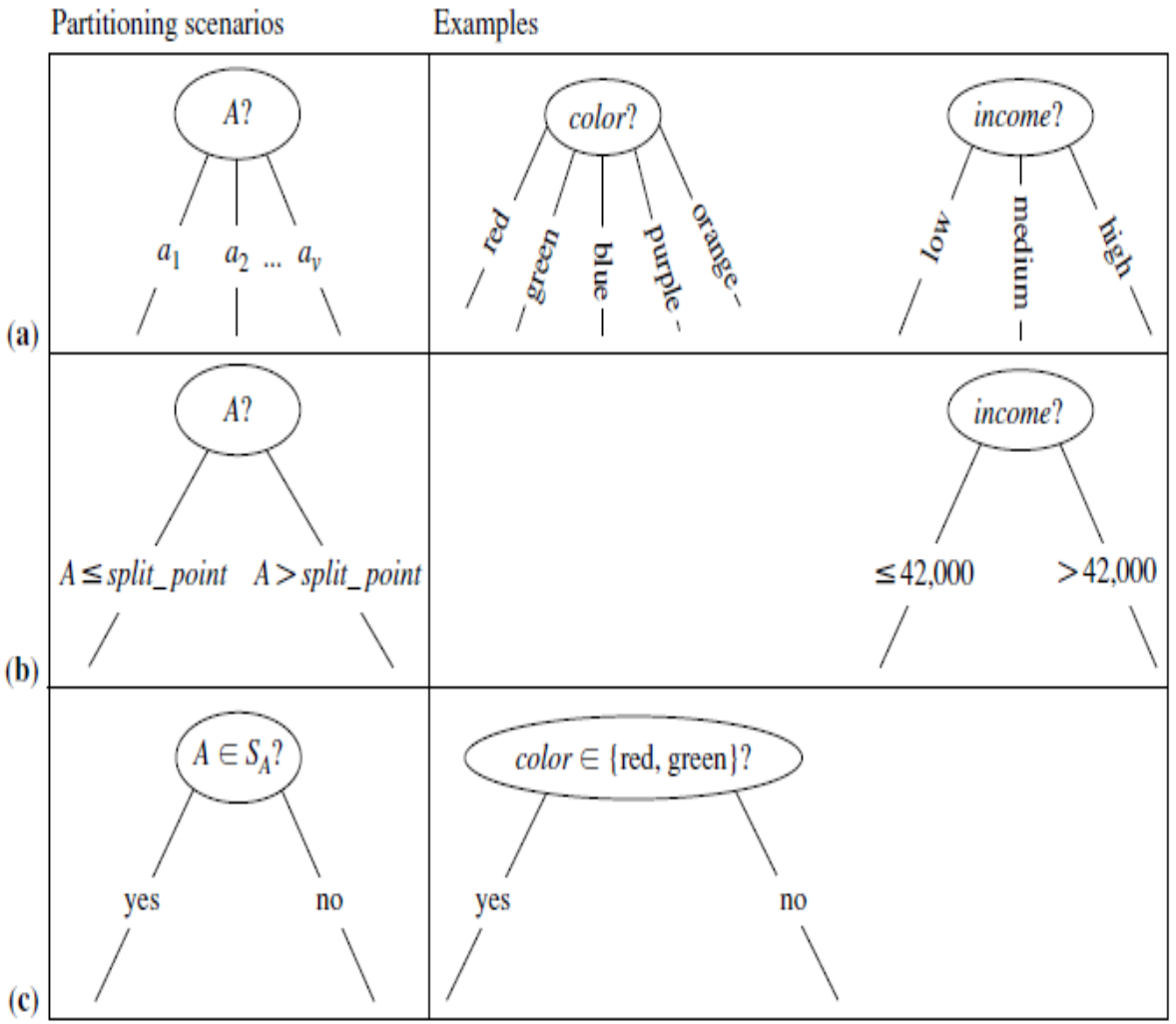


(A) A partition of the data space



(B). The decision tree

Παραδείγματα



Παράδειγμα

$$Info(D) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right) = 0.940 \text{ bits}$$

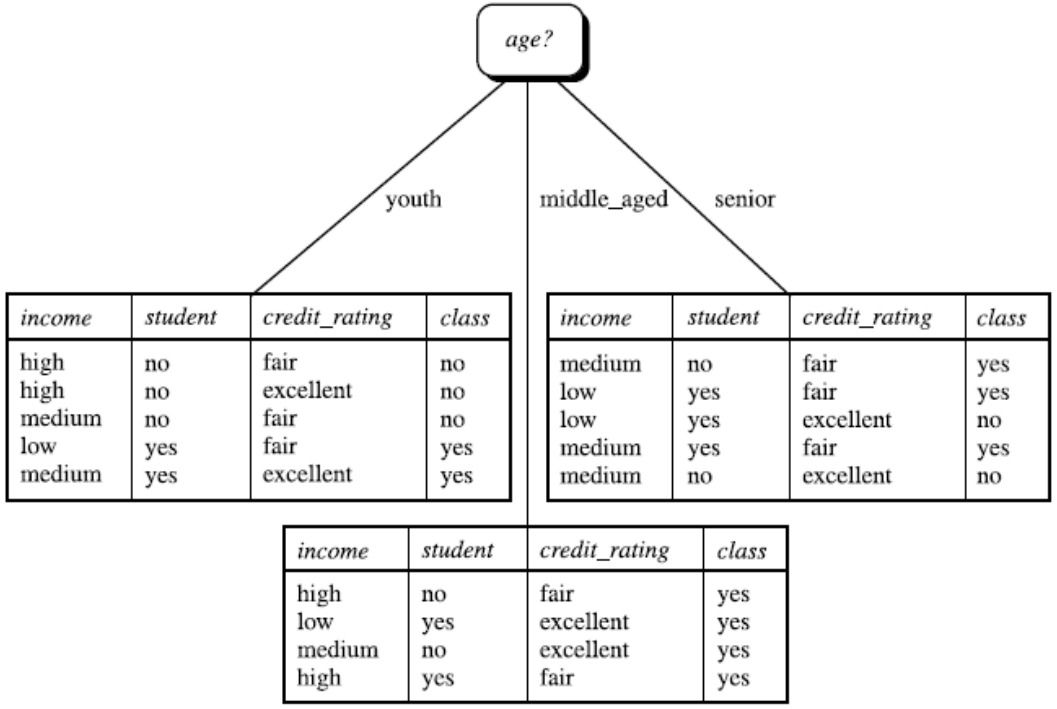
$$\begin{aligned} Info_{age}(D) = {} & \frac{5}{14} \times \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) \\ & + \frac{4}{14} \times \left(-\frac{4}{4}\log_2\frac{4}{4}\right) \\ & + \frac{5}{14} \times \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right) \\ & = 0.694 \text{ bits} \end{aligned}$$

$$Gain(age) = Info(D) - Info_{age}(D) = 0.940 - 0.694 = 0.246 \text{ bits}$$

$$Gain(income) = 0.029 \text{ bits} \quad Gain(student) = 0.151 \text{ bits}$$

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Παράδειγμα



income	*student*	*credit_rating*	*class*
high	no	fair	no
high	no	excellent	no
medium	no	fair	no
low	yes	fair	yes
medium	yes	excellent	yes

income	*student*	*credit_rating*	*class*
medium	no	fair	yes
low	yes	fair	yes
low	yes	excellent	no
medium	yes	fair	yes
medium	no	excellent	no

income	*student*	*credit_rating*	*class*
high	no	fair	yes
low	yes	excellent	yes
medium	no	excellent	yes
high	yes	fair	yes

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Gain Ratio

- Η μέτρηση κέρδους πληροφοριών (information gain) είναι ‘προκατειλημμένη’ προς χαρακτηριστικά με μεγάλο αριθμό τιμών
- Το C4.5 (ένας διάδοχος του ID3) χρησιμοποιεί αναλογία κέρδους (gain ratio) για να ξεπεράσει το πρόβλημα (κανονικοποίηση σε κέρδος πληροφοριών)

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right)$$

- GainRatio(A) = Gain(A)/SplitInfo(A)
- Παράδειγμα
- GainRatio(income) = 0.029/1.557 = 0.019
- Το χαρακτηριστικό με το μέγιστο GainRatio επιλέγεται ως χαρακτηριστικό διαχωρισμού

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Gini Index

- Εάν ένα σύνολο δεδομένων D περιέχει παραδείγματα από n κατηγορίες, ο δείκτης gini, το gini(D) ορίζεται ως

$$gini(D) = 1 - \sum_{j=1}^{n} p_j^2$$

- όπου p_j είναι η σχετική συχνότητα της κλάσης j στο D
- Εάν ένα σύνολο δεδομένων D χωριστεί στο A σε δύο υποσύνολα D1 και D2, ο δείκτης gini(D) ορίζεται ως

$$gini_A(D) = \frac{|D_1|}{|D|} gini(D_1) + \frac{|D_2|}{|D|} gini(D_2)$$

- Μείωση της αβεβαιότητας (impurity):

$$\Delta gini(A) = gini(D) - gini_A(D)$$

- Το χαρακτηριστικό παρέχει το μικρότερο $gini_{split}(D)$ (ή τη μεγαλύτερη μείωση του impurity) επιλέγεται για τον διαχωρισμό του κόμβου (πρέπει να απαριθμηθούν όλα τα πιθανά σημεία διαχωρισμού για κάθε χαρακτηριστικό)

Παράδειγμα

- 9 tuples in buys_computer = “yes” and 5 in “no”
- Suppose the attribute income partitions D into 10 in D1: {low, medium} and 4 in D2

$$gini_{income \in \{low,medium\}}(D) = \left(\frac{10}{14}\right) Gini(D_1) + \left(\frac{4}{14}\right) Gini(D_2)$$

$$= \frac{10}{14}\left(1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2\right) + \frac{4}{14}\left(1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right)$$
$$= 0.443$$
$$= Gini_{income \in \{high\}}(D).$$

- Gini{low,high} is 0.458; Gini{medium,high} is 0.450.
- Split on the {low,medium} (and {high}) since it has the lowest Gini index

$$gini(D) = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2 = 0.459$$

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Σύγκριση

- Οι τρεις μετρικές, σε γενικές γραμμές, επιστρέφουν καλά αποτελέσματα αλλά
  - Information Gain - Κέρδος πληροφοριών:
    - προκατειλημμένη προς ιδιότητες με πολλές τιμές
  - Gain Ratio - Αναλογία κέρδους:
    - τείνει να προτιμά μη ισορροπημένους διαχωρισμούς στους οποίους το ένα τμήμα είναι πολύ μικρότερο από τα άλλα
  - Δείκτης Gini:
    - προκατειλημμένη προς ιδιότητες με πολλές τιμές
    - δυσκολεύεται όταν το πλήθος τάξεων/κλάσεων είναι μεγάλο
    - τείνει να ευνοεί τις δοκιμές που καταλήγουν σε διαχωρισμούς ίσου μεγέθους και αγνότητα/καθαρότητα και στα δύο τμήματα

Άλλες Τεχνικές

- CHAID: a popular decision tree algorithm, measure based on χ2 test for independence
- C-SEP: performs better than info. gain and gini index in certain cases
- G-statistic: has a close approximation to χ2 distribution
- MDL (Minimal Description Length) principle (i.e., the simplest solution is preferred):
- The best tree as the one that requires the fewest # of bits to both (1) encode the tree, and (2) encode the exceptions to the tree
- Multivariate splits (partition based on multiple variable combinations)
- CART: finds multivariate splits based on a linear comb. of attrs.
- Which attribute selection measure is the best?
- Most give good results, none is significantly superior than others

Μετρικές

$$P = \frac{|TP|}{|TP| + |FP|}$$

$$R = \frac{|TP|}{|TP| + |FN|}$$

		Predicted class: *yes*	Predicted class: *no*	Total
Actual class	*yes*	*TP*	*FN*	*P*
	no	*FP*	*TN*	*N*
	Total	*P′*	*N′*	*P* + *N*

$$F_\beta = \frac{(1 + \beta^2) \times P \times R}{(\beta^2 \times P) + R}$$

$$\text{Accuracy} = \frac{\text{correct classifications}}{\text{total classifications}} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{FPR} = \frac{\text{incorrectly classified actual negatives}}{\text{all actual negatives}} = \frac{FP}{FP + TN}$$

Overfitting

- Υπερπροσαρμογή (overfitting): Ένα δέντρο μπορεί να προσαρμόζεται υπερβολικά στα δεδομένα εκπαίδευσης
  - Πάρα πολλά κλαδιά, μερικά μπορεί να αντικατοπτρίζουν ανωμαλίες λόγω θορύβου ή ακραίων σημείων
  - Κακή ακρίβεια για μη εμφανή δείγματα
- Υπερπροσαρμογή είναι "η παραγωγή μιας ανάλυσης που αντιστοιχεί πολύ στενά ή ακριβώς σε ένα συγκεκριμένο σύνολο δεδομένων, και επομένως μπορεί να αποτύχει να κατηγοριοποιήσει πρόσθετα δεδομένα ή να προβλέψει μελλοντικές παρατηρήσεις με αξιοπιστία".
- Δύο προσεγγίσεις για να αποφευχθεί
  - Προκλάδεμα - Prepruning: Σταματούμε νωρίς την κατασκευή δέντρων - μην χωρίσετε έναν κόμβο εάν αυτό θα είχε ως αποτέλεσμα οι μετρικές να πέσουν κάτω από ένα όριο
    - Δύσκολη η επιλογή κατάλληλου ορίου
  - Postpruning: Αφαιρούμε κλαδιά από ένα «πλήρως αναπτυγμένο» δέντρο - σε μια σειρά από δέντρα που κλαδεύονται σταδιακά
    - Χρησιμοποιήστε ένα σύνολο δεδομένων διαφορετικών από τα δεδομένα εκπαίδευσης για να αποφασίσετε ποιο είναι το «καλύτερο κλαδεμένο δέντρο»

Underfitting

- Η υποπροσαρμογή (underfitting) αναφέρεται σε ένα μοντέλο που δεν μπορεί ούτε να μοντελοποιήσει τα δεδομένα εκπαίδευσης ούτε να γενικεύσει σε νέα δεδομένα.

- Ένα τέτοιο μοντέλο μηχανικής εκμάθησης δεν είναι κατάλληλο και θα είναι προφανές καθώς θα έχει κακή απόδοση στα δεδομένα εκπαίδευσης.

- Η υποπροσαρμογή συχνά δεν συζητείται, καθώς είναι εύκολο να εντοπιστεί με βάση μια καλή μετρική απόδοσης.

Ενίσχυση

- Να επιτρέπονται χαρακτηριστικά με συνεχείς τιμές
  - Ορίστε δυναμικά νέα χαρακτηριστικά διακριτής τιμής που διαχωρίζουν τη συνεχή τιμή σε ένα διακριτό σύνολο διαστημάτων
- Χειριστείτε τις τιμές χαρακτηριστικών που λείπουν
  - Εκχωρήστε την πιο κοινή τιμή του χαρακτηριστικού
  - Εκχωρήστε πιθανότητα σε καθεμία από τις πιθανές τιμές
- Κατασκευή χαρακτηριστικών
  - Δημιουργήστε νέα χαρακτηριστικά με βάση τα υπάρχοντα που παρουσιάζονται αραιά
  - Αυτό μειώνει τον κατακερματισμό, την επανάληψη και την αναπαραγωγή

Παράδειγμα

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	***No***
Sunny	Hot	High	True	***No***
Overcast	Hot	High	False	***Yes***
Rainy	Mild	High	False	***Yes***
Rainy	Cool	Normal	False	***Yes***
Rainy	Cool	Normal	True	***No***
Overcast	Cool	Normal	True	***Yes***
Sunny	Mild	High	False	***No***
Sunny	Cool	Normal	False	***Yes***
Rainy	Mild	Normal	False	***Yes***
Sunny	Mild	Normal	True	***Yes***
Overcast	Mild	High	True	***Yes***
Overcast	Hot	Normal	False	***Yes***
Rainy	Mild	High	True	***No***

Παράδειγμα

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	*No*
Sunny	Hot	High	True	*No*
Overcast	Hot	High	False	*Yes*
Rainy	Mild	High	False	*Yes*
Rainy	Cool	Normal	False	*Yes*
Rainy	Cool	Normal	True	*No*
Overcast	Cool	Normal	True	*Yes*
Sunny	Mild	High	False	*No*
Sunny	Cool	Normal	False	*Yes*
Rainy	Mild	Normal	False	*Yes*
Sunny	Mild	Normal	True	*Yes*
Overcast	Mild	High	True	*Yes*
Overcast	Hot	Normal	False	*Yes*
Rainy	Mild	High	True	*No*

$$H(Y) = -\sum_{i=1}^{K} p_k \log_2 p_k$$

$$= -\frac{5}{14}\log_2\frac{5}{14} - \frac{9}{14}\log_2\frac{9}{14}$$

$$= 0.94$$

Παράδειγμα

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	*No*
Sunny	Hot	High	True	*No*
Overcast	Hot	High	False	*Yes*
Rainy	Mild	High	False	*Yes*
Rainy	Cool	Normal	False	*Yes*
Rainy	Cool	Normal	True	*No*
Overcast	Cool	Normal	True	*Yes*
Sunny	Mild	High	False	*No*
Sunny	Cool	Normal	False	*Yes*
Rainy	Mild	Normal	False	*Yes*
Sunny	Mild	Normal	True	*Yes*
Overcast	Mild	High	True	*Yes*
Overcast	Hot	Normal	False	*Yes*
Rainy	Mild	High	True	*No*

$$InfoGain(\text{Humidity}) = H(Y) - \frac{m_L}{m} H_L - \frac{m_R}{m} H_R$$

$$= 0.94 - \frac{7}{14} H_L - \frac{7}{14} H_R$$

Παράδειγμα

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	***No***
Sunny	Hot	High	True	***No***
Overcast	Hot	High	False	***Yes***
Rainy	Mild	High	False	***Yes***
Rainy	Cool	**Normal**	False	***Yes***
Rainy	Cool	**Normal**	True	***No***
Overcast	Cool	**Normal**	True	***Yes***
Sunny	Mild	High	False	***No***
Sunny	Cool	**Normal**	False	***Yes***
Rainy	Mild	**Normal**	False	***Yes***
Sunny	Mild	**Normal**	True	***Yes***
Overcast	Mild	High	True	***Yes***
Overcast	Hot	**Normal**	False	***Yes***
Rainy	Mild	High	True	***No***

$$InfoGain(\text{Humidity}) = H(Y) - \frac{m_L}{m} H_L - \frac{m_R}{m} H_R$$

$$0.94 - \frac{7}{14} H_L - \frac{7}{14} H_R$$

$$H_L = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}$$

Παράδειγμα

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	**High**	False	*No*
Sunny	Hot	**High**	True	*No*
Overcast	Hot	**High**	False	*Yes*
Rainy	Mild	**High**	False	*Yes*
Rainy	Cool	Normal	False	***Yes***
Rainy	Cool	Normal	True	***No***
Overcast	Cool	Normal	True	***Yes***
Sunny	Mild	**High**	False	*No*
Sunny	Cool	Normal	False	***Yes***
Rainy	Mild	Normal	False	***Yes***
Sunny	Mild	Normal	True	***Yes***
Overcast	Mild	**High**	True	*Yes*
Overcast	Hot	Normal	False	***Yes***
Rainy	Mild	**High**	True	*No*

$$InfoGain(\text{Humidity}) = H(Y) - \frac{m_L}{m} H_L - \frac{m_R}{m} H_R$$

$$0.94 - \frac{7}{14} H_L - \frac{7}{14} H_R$$

$$H_L = -\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7} = 0.592$$

$$H_R = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.985$$

Παράδειγμα

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	False	***No***
Sunny	Hot	High	True	***No***
Overcast	Hot	High	False	***Yes***
Rainy	Mild	High	False	***Yes***
Rainy	Cool	Normal	False	***Yes***
Rainy	Cool	Normal	True	***No***
Overcast	Cool	Normal	True	***Yes***
Sunny	Mild	High	False	***No***
Sunny	Cool	Normal	False	***Yes***
Rainy	Mild	Normal	False	***Yes***
Sunny	Mild	Normal	True	***Yes***
Overcast	Mild	High	True	***Yes***
Overcast	Hot	Normal	False	***Yes***
Rainy	Mild	High	True	***No***

$$InfoGain(\text{Humidity}) = H(Y) - \frac{m_L}{m} H_L - \frac{m_R}{m} H_R$$

$$0.94 - \frac{7}{14} 0.592 - \frac{7}{14} 0.985$$

$$= 0.94 - 0.296 - 0.4925$$

$$= 0.1515$$

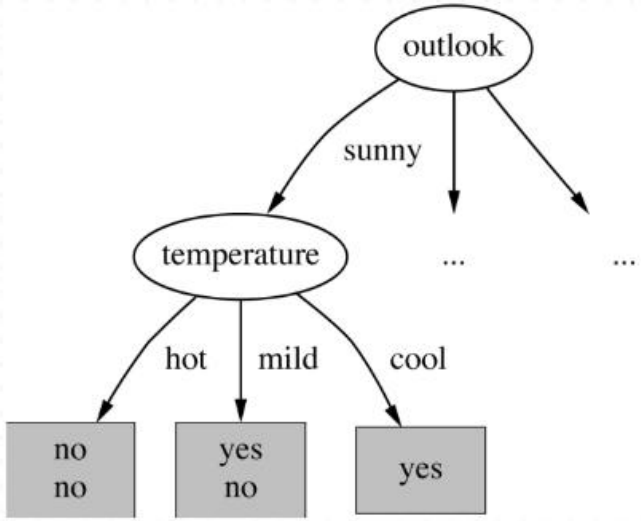
Παράδειγμα

Information gain for each feature:

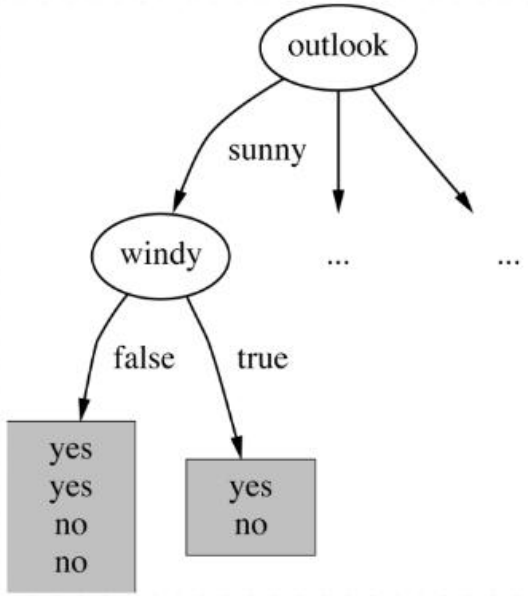
- Outlook = 0.247
- Temperature = 0.029
- Humidity = 0.152
- Windy = 0.048

Initial split is on outlook, because it is the feature with the highest information gain.

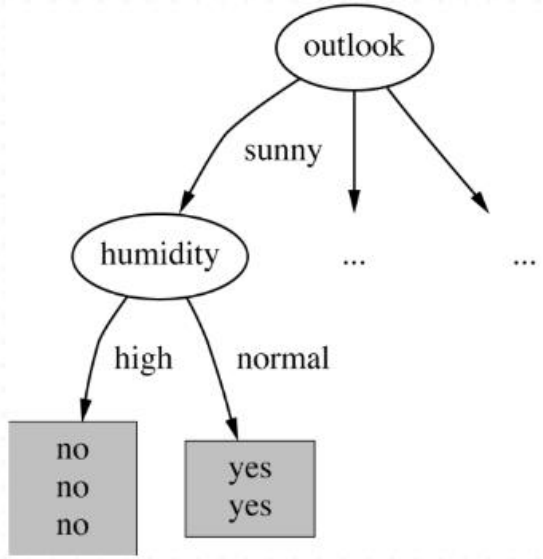
Παράδειγμα



Temperature = 0.571

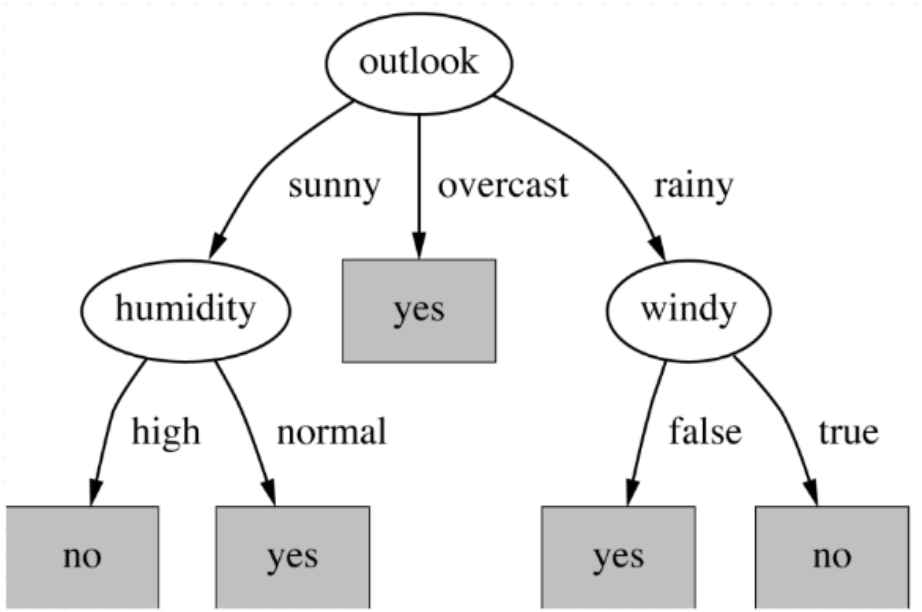


Windy = 0.020



Humidity = 0.971

Παράδειγμα


outlook
sunny
overcast
rainy
humidity
yes
windy
high
normal
false
true
no
yes
yes
no