Θεωρία Αποφάσεων

Παρακάτω παρουσιάζονται συνοπτικά τα παραδείγματα τα οποία παρουσιάστηκαν κατά τη διάρκεια των διαλέξεων του μαθήματος και δεν συμπεριλαμβάνονται στο βιβλίο του μαθήματος.

ΚΕΦΑΛΑΙΟ 2^{Ο}

Υπο Συνθήκη Πιθανότητες

P(A | B) = Υποσύνολο του δειγματικού χώρου στο οποίο το B είναι αληθές και στο οποίο επίσης και το A είναι αληθές

Παράδειγμα

H = "Have a headache"
F = "Coming down with Flu"

P(H) = 1/10
P(F) = 1/40
P(H | F) = 1/2

"Headaches are rare and flu is rarer, but if you're coming down with 'flu there's a 50-50 chance you'll have a headache."

P(H | F) = Fraction of flu-inflicted worlds in which you have a headache

$$= \frac{\text{\#worlds with flu and headache}}{\text{\#worlds with flu}}$$

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

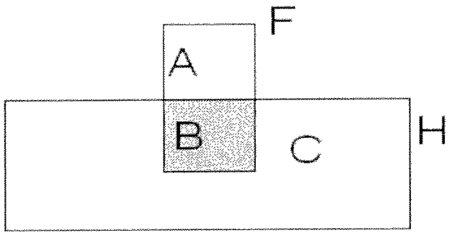
$$= \frac{P(H \wedge F)}{P(F)}$$

Πιθανοτικό Inference

One day you wake up with a headache. You think: "Drat! 50% of flus are associated with headaches so I must have a 50-50 chance of coming down with flu"

Is this reasoning good?

Another way to understand the intuition



Let's say we have P(F), P(H), and P(H|F), like in the example in class.

Areawise, P(F) = A + B, P(H) = B + C,

Also, $P(H|F) = \frac{B}{A + B}$

Thus, to get the opposite conditional probability, ie, P(F|H), we need to figure out $\frac{B}{B + C}$

Since we know B / (A+B), we can get B / (B+C) by multiplying by (A+B) and dividing by (B+C). But since we already calculated, A+B = P(F), and B+C = P(H), so we are actually multiplying by P(F) and dividing by P(H). Which is Bayes Rule:

$$P(F|H) = P(H|F) * \frac{P(F)}{P(H)}$$

Παράδειγμα

Classify cars on UNR campus whether they are more or less than \$50K:

- C1: price > \$50K
- C2: price < \$50K
- Feature x: height of car

From Bayes' rule, we know how to compute the posterior probabilities:

$$P(C_i / x) = \frac{p(x / C_i) P(C_i)}{p(x)}$$

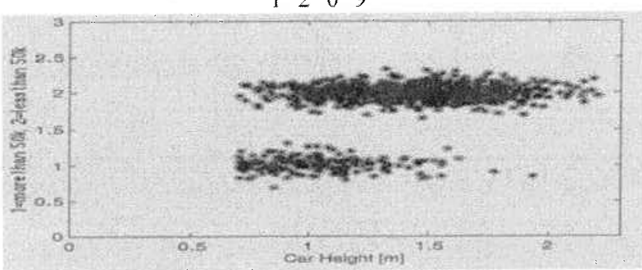
Need to compute $p(x/C_1)$, $p(x/C_2)$, $P(C_1)$, $P(C_2)$

Determine prior probabilities

- Collect data: ask drivers how much their car was and measure height.
- e.g., 1209 samples: $\#C_1=221$ $\#C_2=988$

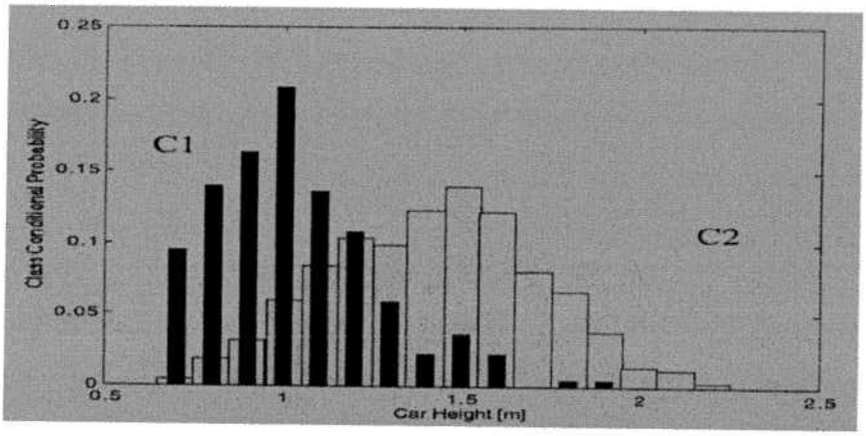
$$P(C_1) = \frac{221}{1209} = 0.183$$

$$P(C_2) = \frac{988}{1209} = 0.817$$



Determine class conditional probabilities (*likelihood*).

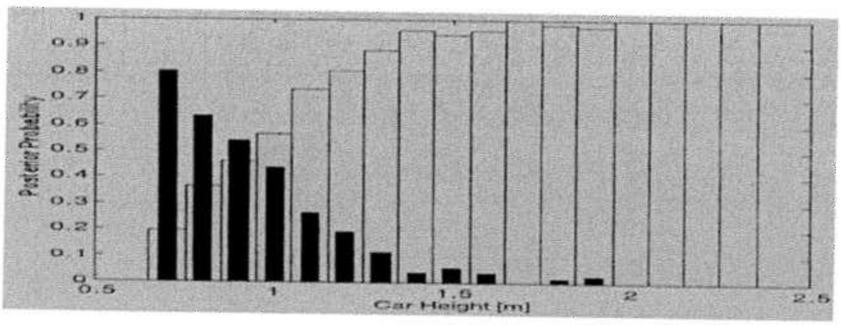
- Discretize car height into bins and use normalized histogram.



Calculate the posterior probability for each bin:

$$P(C_1 / x = 1.0) = \frac{p(x = 1.0 / C_1)\, P(C_1)}{p(x = 1.0 / C_1)\, P(C_1) + p(x = 1.0 / C_2)\, P(C_2)} = $$

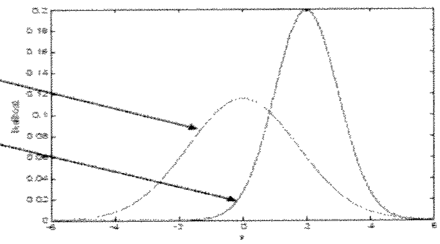
$$= \frac{0.2081 * 0.183}{0.2081 * 0.183 + 0.0597 * 0.817} = 0.438$$



- Consider a classification problem with two classes defined by the following likelihood functions

$$P(x \mid \omega_1) = \frac{1}{\sqrt{2\pi}\sqrt{3}} e^{-\frac{1}{2}\frac{x^2}{3}}$$

$$P(x \mid \omega_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2}$$



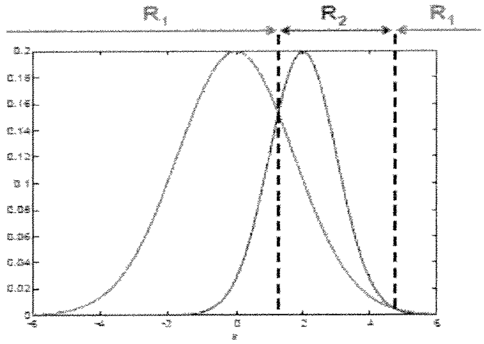
- What is the decision rule that minimizes P[error]?
  - Assume $P[\omega_1]=P[\omega_2]=0.5$, $C_{11}=C_{22}=0$, $C_{12}=1$ and $C_{21}=3^{1/2}$

$$\Lambda(x) = \frac{\frac{1}{\sqrt{2\pi}\sqrt{3}} e^{-\frac{1}{2}\frac{x^2}{3}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2}} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{1}{\sqrt{3}} \Rightarrow$$

$$\frac{e^{-\frac{1}{2}\frac{x^2}{3}}}{e^{-\frac{1}{2}(x-2)^2}} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1 \Rightarrow$$

$$-\frac{1}{2}\frac{x^2}{3} + \frac{1}{2}(x-2)^2 \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0 \Rightarrow$$

$$2x^2 - 12x + 12 \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0 \Rightarrow x = 4.73, 1.27$$



Επιλογή της κλάσης που έχει την μεγαλύτερη εκ των υστέρων πιθανότητα (Bayes Decision Rule)

Παράδειγμα

Ταξινόμησε αυτοκίνητα στο UNR campus σύμφωνα με το αν αξίζουν περισσότερα ή λιγότερα από $50K:

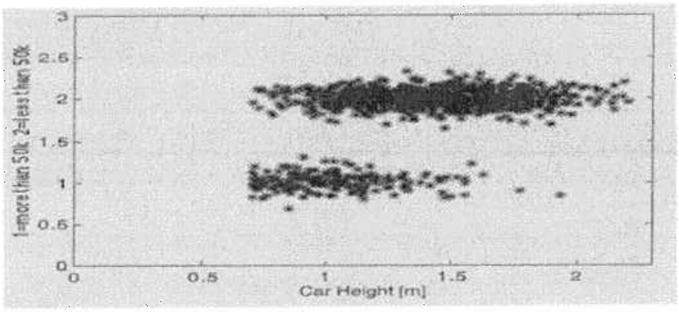
- C1: τιμή > $50K
- C2: τιμή < $50K
- Χαρακτηριστικό x: ύψος του αυτοκινήτου

Από τον νόμο του Bayes, ξέρουμε να υπολογίζουμε τις εκ των υστέρων πιθανότητες.
$$P(C_i / x) = \frac{p(x / C_i)P(C_i)}{p(x)}$$

Πρέπει να υπολογίσουμε τις *p(x/C1), p(x/C2), P(C1), P(C2)*

Καθόρισε τις εκ των προτέρων πιθανότητες

- Συλλογή δεδομένων: Ρώτησε οδηγούς πόσο αγόρασαν το αυτοκίνητο τους και μέτρησε το ύψος.
- π.χ., 1209 δείγματα: #C1=221 #C2=988

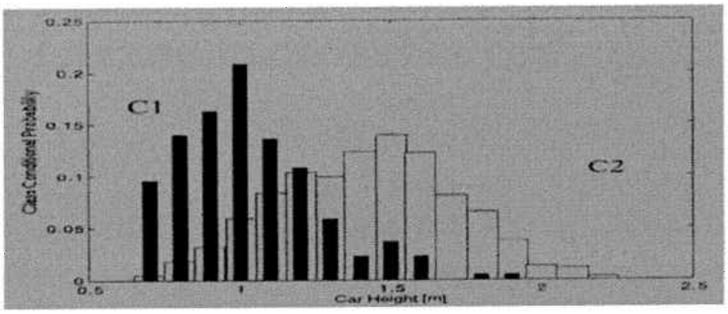


$$P(C_1) = \frac{221}{1209} = 0.183$$
$$P(C_2) = \frac{988}{1209} = 0.817$$

Καθόρισε τις υπο συνθήκη συναρτήσεις π.π. (πιθανοφάνεια)

- Διακριτοποίησε το ύψος των αυτοκινήτων σε bins και χρησιμοποίησε κανονικοποιημένο ιστόγραμμα

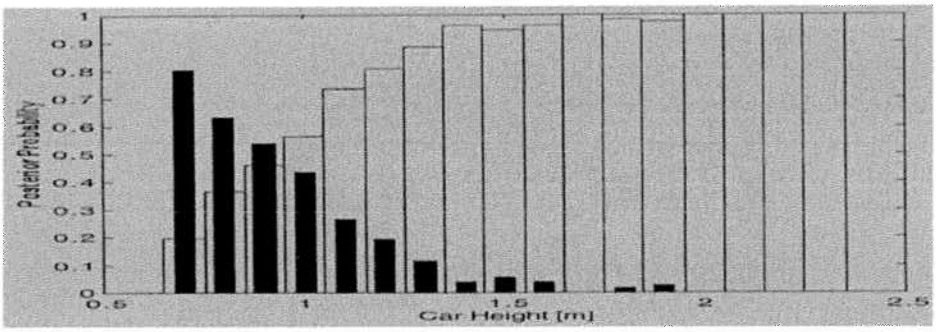
p(x | Ci)



Υπολόγισε την εκ των υστέρων πιθανότητα για κάθε bin:

$$P(C_1 / x = 1.0) = \frac{p(x = 1.0 / C_1)\, P(C_1)}{p(x = 1.0 / C_1)\, P(C_1) + p(x = 1.0 / C_2)\, P(C_2)} =$$

$$= \frac{0.2081 * 0.183}{0.2081 * 0.183 + 0.0597 * 0.817} = 0.438$$



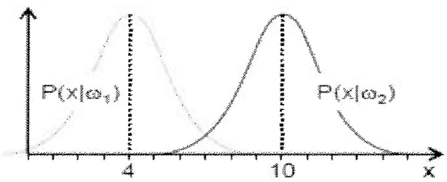
Κανόνες Απόφασης

Παράδειγμα

Για ένα πρόβλημα ταξινόμησης δίνονται οι παρακάτω δεσμευμένες συναρτήσεις π.π

$$P(x \mid \omega_1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-4)^2}$$

$$P(x \mid \omega_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-10)^2}$$



Θεωρώντας συνάρτηση κόστους μηδέν-ένα και ίσες εκ των προτέρων πιθανότητες βρείτε έναν κανόνα απόφασης

Λύση:

Αντικαταστώντας της τιμές των πιθανοφανειών και των εκ των προτέρων πιθανοτήτων στο τύπο (3):

$$\Lambda(x) = \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-4)^2}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-10)^2}} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{1}{1}$$

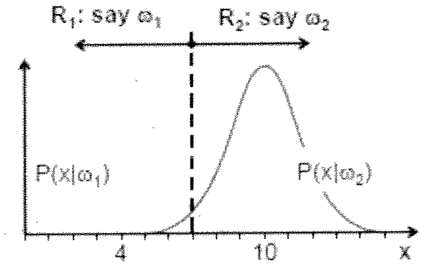
Απλοποιώντας, αλλάζοντας πρόσημο και λογαριθμίζοντας :

$$(x-4)^2 - (x-10)^2 \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 0$$

Οπότε:

$$\boxed{x \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 7}$$

Το αποτέλεσμα το αναμέναμε και διαισθητικά μια που οι πιθανοφάνειες διαφέρουν μόνο ως προς τον μέσο όρο τους

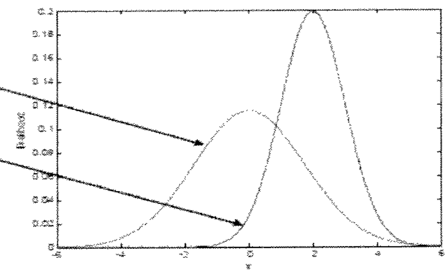


Παράδειγμα

Για ένα πρόβλημα ταξινόμησης δίνονται οι παρακάτω δεσμευμένες συναρτήσεις π.π

$$P(x \mid \omega_1) = \frac{1}{\sqrt{2\pi}\sqrt{3}} e^{-\frac{1}{2}\frac{x^2}{3}}$$

$$P(x \mid \omega_2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2}$$



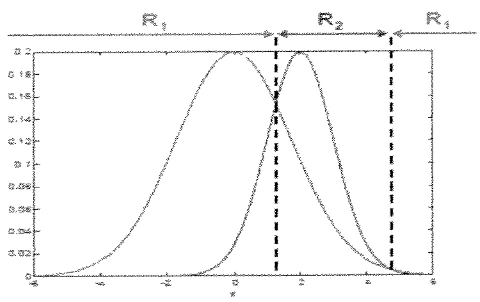
Θεωρώντας P(ω1)= P(ω2), λ11=λ22=0, λ12=1 και λ21=31/2 ποιός είναι ο κανόνας, ο οποίος μειώνει την πιθανότητα λάθους?

$$\Lambda(x) = \frac{\frac{1}{\sqrt{2\pi}\sqrt{3}} e^{-\frac{1}{2}\frac{x^2}{3}}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-2)^2}} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{1}{\sqrt{3}} \Rightarrow$$

$$\frac{e^{-\frac{1}{2}\frac{x^2}{3}}}{e^{-\frac{1}{2}(x-2)^2}} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1 \Rightarrow$$

$$-\frac{1}{2}\frac{x^2}{3} + \frac{1}{2}(x-2)^2 \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0 \Rightarrow$$

$$2x^2 - 12x + 12 \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0 \Rightarrow x = 4.73, 1.27$$



Ταξινόμηση Ελάχιστου Ρυθμού Λάθους και συναρτήσεις κανονικής κατανομής

Αριθμητικό Παράδειγμα

Έστω ότι :

$$\mu_1=[0 \quad 0 \quad 0]^T;\ \mu_2=[1 \quad 1 \quad 1]^T;\ \Sigma_1=\Sigma_2=\begin{bmatrix}1/4 & 0 & 0\\ 0 & 1/4 & 0\\ 0 & 0 & 1/4\end{bmatrix};\ p(\omega_2)=2p(\omega_1)$$

$$g_i(x)=-\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i)+\log P(\omega_i)=-\frac{1}{2}\begin{bmatrix}x-\mu_x\\ y-\mu_y\\ z-\mu_z\end{bmatrix}^T\begin{bmatrix}4 & 0 & 0\\ 0 & 4 & 0\\ 0 & 0 & 4\end{bmatrix}\begin{bmatrix}x-\mu_x\\ y-\mu_y\\ z--\mu_z\end{bmatrix}+\log P(\omega_i)$$

$$g_1(x)=-\frac{1}{2}\begin{bmatrix}x-0\\ y-0\\ z-0\end{bmatrix}^T\begin{bmatrix}4 & 0 & 0\\ 0 & 4 & 0\\ 0 & 0 & 4\end{bmatrix}\begin{bmatrix}x-0\\ y-0\\ z-0\end{bmatrix}+\log\frac{1}{3};\qquad g_2(x)=-\frac{1}{2}\begin{bmatrix}x-1\\ y-1\\ z-1\end{bmatrix}^T\begin{bmatrix}4 & 0 & 0\\ 0 & 4 & 0\\ 0 & 0 & 4\end{bmatrix}\begin{bmatrix}x-1\\ y-1\\ z-1\end{bmatrix}+\log\frac{2}{3}$$

$$g_1(x)\underset{\omega_2}{\overset{\omega_1}{\gtrless}}g_2(x)\Rightarrow -2(x^2+y^2+z^2)+\log\frac{1}{3}\underset{\omega_2}{\overset{\omega_1}{\gtrless}}-2((x-1)^2+(y-1)^2+(z-1)^2)+\log\frac{2}{3}$$

$$\boxed{x+y+z\underset{\omega_1}{\overset{\omega_2}{\gtrless}}\frac{6-\log 2}{4}=1.32}$$

x_u=[0.1 0.7 0.8]^T

$$0.1+0.7+0.8=1.6\underset{\omega_1}{\overset{\omega_2}{\gtrless}}1.32\Rightarrow x_u\in\omega_2$$

$$\mu_1=[0 \quad 0 \quad 0]^T;\ \mu_2=[1 \quad 1 \quad 1]^T;\ \Sigma_1=\Sigma_2=\begin{bmatrix}1/4 & 0 & 0\\ 0 & 1/4 & 0\\ 0 & 0 & 1/4\end{bmatrix};\ p(\omega_2)=2p(\omega_1)$$

$$g_i(x)=-\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i)+\log P(\omega_i)=-\frac{1}{2}\begin{bmatrix}x-\mu_x\\ y-\mu_y\\ z-\mu_z\end{bmatrix}^T\begin{bmatrix}4 & 0 & 0\\ 0 & 4 & 0\\ 0 & 0 & 4\end{bmatrix}\begin{bmatrix}x-\mu_x\\ y-\mu_y\\ z--\mu_z\end{bmatrix}+\log P(\omega_i)$$

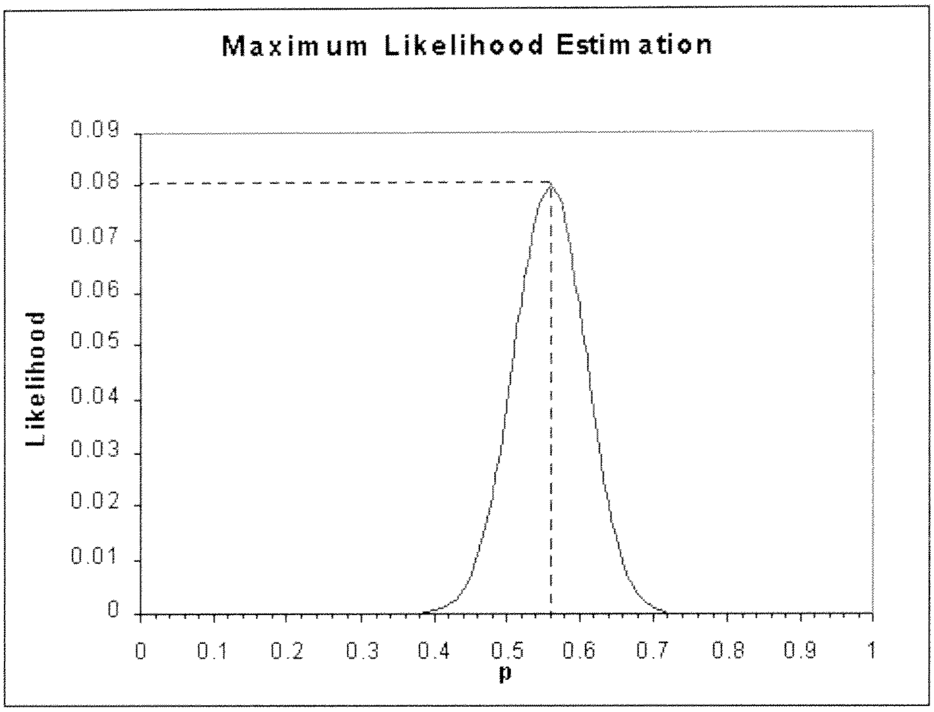
$$g_1(x)=-\frac{1}{2}\begin{bmatrix}x-0\\ y-0\\ z-0\end{bmatrix}^T\begin{bmatrix}4 & 0 & 0\\ 0 & 4 & 0\\ 0 & 0 & 4\end{bmatrix}\begin{bmatrix}x-0\\ y-0\\ z-0\end{bmatrix}+\log\frac{1}{3};\qquad g_2(x)=-\frac{1}{2}\begin{bmatrix}x-1\\ y-1\\ z-1\end{bmatrix}^T\begin{bmatrix}4 & 0 & 0\\ 0 & 4 & 0\\ 0 & 0 & 4\end{bmatrix}\begin{bmatrix}x-1\\ y-1\\ z-1\end{bmatrix}+\log\frac{2}{3}$$

$$g_1(x)\underset{\omega_2}{\overset{\omega_1}{\gtrless}}g_2(x)\Rightarrow -2(x^2+y^2+z^2)+\log\frac{1}{3}\underset{\omega_2}{\overset{\omega_1}{\gtrless}}-2((x-1)^2+(y-1)^2+(z-1)^2)+\log\frac{2}{3}$$

$$\boxed{x+y+z\underset{\omega_1}{\overset{\omega_2}{\gtrless}}\frac{6-\log 2}{4}=1.32}$$

x_u=[0.1 0.7 0.8]^T

$$0.1+0.7+0.8=1.6\underset{\omega_1}{\overset{\omega_2}{\gtrless}}1.32\Rightarrow x_u\in\omega_2$$



Μπεϋζιανή Εκμάθηση –Κανονική κατανομή

Αριθμητικό παράδειγμα

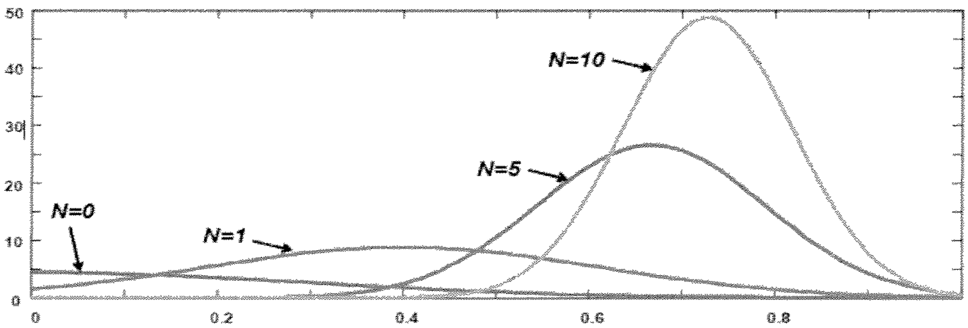
Έστω ότι ο πραγματικός μέσος όρος της p(x) είναι μ=0.8, με τυπική απόκλιση σ=0.3

- Στην πραγματικότητα δεν θα γνωρίζαμε τον πραγματικό μέσο όρο...

Παράγουμε έναν αριθμό από δείγματα της κατανομή αυτής

Εκφράζουμε την γνώση μας για τον μέσο όρο, θεωρώντας ότι η εκ των προτέρων πιθανότητα ακολουθεί κανονική κατανομή με μ0=0.0 και σ0=0.3

Όσο αυξάνει ο αριθμός των δειγμάτων εκπαίδευσης, η εκτίμηση μΝ, πλησιάζει την πραγματική τιμή μ=0.8 και η τυπική απόκλιση (δηλ. η αβεβαιότητα της εκτίμησης) ελαττώνεται.



Παράδειγμα

Classify cars on UNR campus whether they are more or less than $50K:

- C1: price > $50K
- C2: price < $50K
- Feature x: height of car

From Bayes' rule, we know how to compute the posterior probabilities:

$$P(C_i / x) = \frac{p(x / C_i) P(C_i)}{p(x)}$$

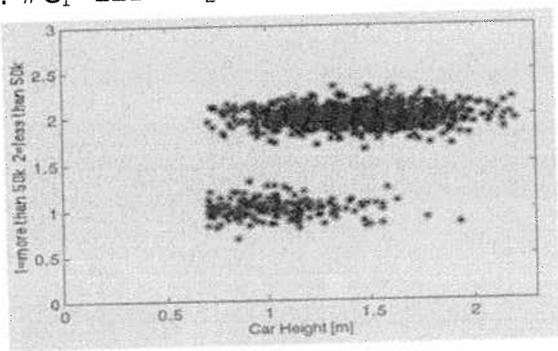
Need to compute *p(x/C₁), p(x/C₂), P(C₁), P(C₂)*

Determine prior probabilities

- Collect data: ask drivers how much their car was and measure height.
- e.g., 1209 samples: $\#C_1=221$ $\#C_2=988$

$$P(C_1) = \frac{221}{1209} = 0.183$$

$$P(C_2) = \frac{988}{1209} = 0.817$$



ΚΕΦΑΛΑΙΟ 3ο

Bias Variance Trade-off

Παράδειγμα

Υποτιθέμενο μοντέλο: $P(x \mid \theta) = \theta\, x(1-\theta)x \quad x = 0,1$
Σενάριο: πελάτες σε σούπερ μάρκετ είτε αγοράζουν είτε δεν αγοράζουν γάλα.
θ είναι η πιθανότητα ότι ένας τυχαίος πελάτης αγοράζει γάλα

$$P(\theta \mid x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^r (1-\theta)^{n-r}$$

όπου r είναι ο αριθμός των πελατών, του δείγματος των n πελατών, που αγόρασαν γάλα

$l(\theta) = r \ln \theta + (n - r) \ln(1 - \theta)$

$$\frac{\partial}{\partial \theta} l(\theta) = \frac{r}{n} - \frac{n-r}{1-\theta} = 0 \qquad \hat{\theta}_{ML} = \frac{r}{n}$$

Παράδειγμα

n=100 Ρίψεις κέρματος
h=56 Ρίψεις κορώνας
t=44 Ρίψεις γράμματα

Είναι το κέρμα «πειραμγμένο»?

Έστω ότι p=0.5.
Τότε :

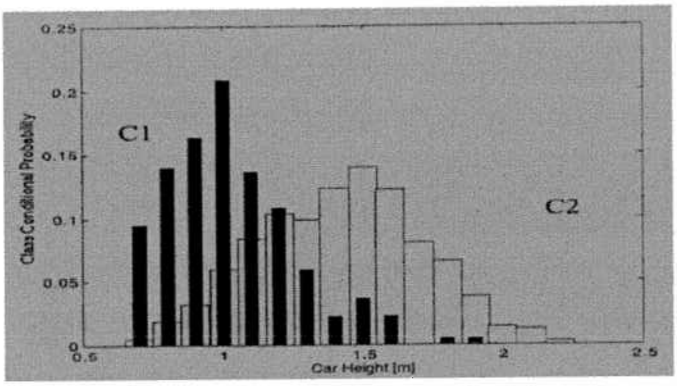
$$L(p = 0.5 \mid data) = \frac{100!}{56!44!} 0.5^{56} 0.5^{44} = 0.0389$$

$$L(p = 0.52 \mid data) = \frac{100!}{56!44!} 0.52^{56} 0.48^{44} = 0.0581$$

P	L
0.48	0.0222
0.50	0.0389
0.52	0.0581
0.54	0.0739
0.56	0.0801
0.58	0.0738
0.60	0.0576
0.62	0.0378

Determine class conditional probabilities (*likelihood*)

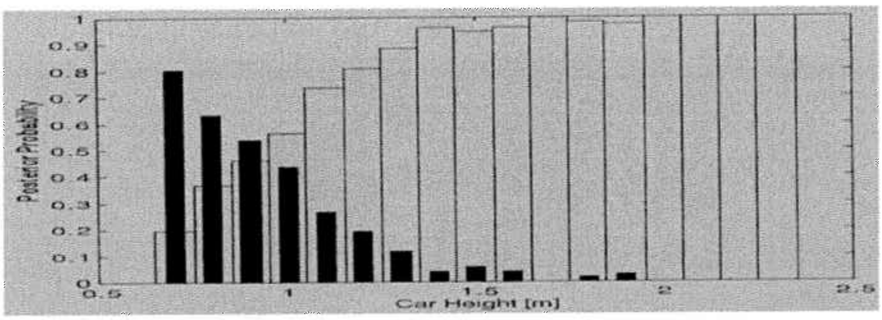
- Discretize car height into bins and use normalized histogram



Calculate the posterior probability for each bin:

$$P(C_1 / x = 1.0) = \frac{p(x = 1.0 / C_1)\, P(C_1)}{p(x = 1.0 / C_1)\, P(C_1) + p(x = 1.0 / C_2)\, P(C_2)} =$$

$$= \frac{0.2081 * 0.183}{0.2081 * 0.183 + 0.0597 * 0.817} = 0.438$$

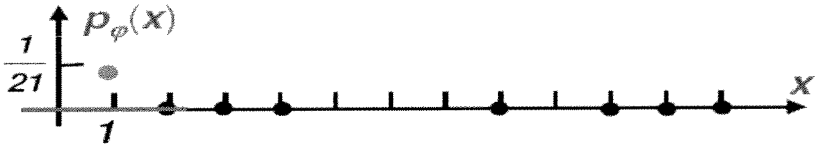


Παράδειγμα 1-D

$$p_{\varphi}(x) = \frac{1}{n}\sum_{i=1}^{i=n}\frac{1}{h^{d}}\,\varphi\left(\frac{x - x_i}{h}\right)$$

Υποθέτουμε ότι έχουμε 7 δείγματα

$$\boldsymbol{D} = \{2,3,4,8,10,11,12\}$$



Θέλουμε να βρούμε την πυκνότητα στο σημείο x=1 για h=3

$$p_{\varphi}(1) = \frac{1}{7}\sum_{i=1}^{i=7}\frac{1}{3}\,\varphi\left(\frac{1 - x_i}{3}\right) = \frac{1}{21}\left[\varphi\left(\frac{1-2}{3}\right) + \varphi\left(\frac{1-3}{3}\right) + \varphi\left(\frac{1-4}{3}\right) + \ldots + \varphi\left(\frac{1-12}{3}\right)\right]$$

$$\left|-\frac{1}{3}\right| \le 1/2 \qquad \left|-\frac{2}{3}\right| > 1/2 \qquad \left|-1\right| > 1/2 \qquad \left|-\frac{11}{3}\right| > 1/2$$

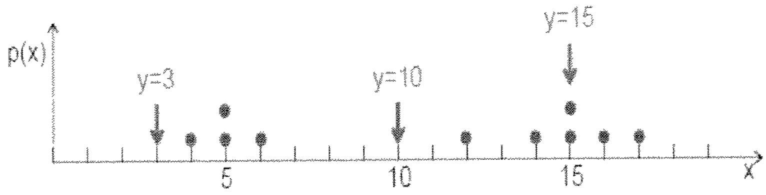
$$p_{\varphi}(1) = \frac{1}{7}\sum_{i=1}^{i=7}\frac{1}{3}\,\varphi\left(\frac{1 - x_i}{3}\right) = \frac{1}{21}\left[1 + 0 + 0 + \ldots + 0\right] = \frac{1}{21}$$

Παράδειγμα

Παράθυρο πλάτους 4

p(x) y=3,10,15 h=4

- $X = \{x^{(1}, x^{(2}, \ldots x^{(N}\} = \{4, 5, 5, 6, 12, 14, 15, 15, 16, 17\}$



- p(y=3):

$$p_{KDE}(y=3) = \frac{1}{Nh^D}\sum_{n=1}^{N} K\left(\frac{y-x^{(n}}{h}\right) = \frac{1}{10\times 4^1}\left[K\left(\underbrace{\frac{3-4}{4}}_{-1/4}\right)+K\left(\underbrace{\frac{3-5}{4}}_{-1/2}\right)+K\left(\underbrace{\frac{3-5}{4}}_{-1/2}\right)+K\left(\underbrace{\frac{3-6}{4}}_{-1}\right)+\ldots+K\left(\underbrace{\frac{3-17}{4}}_{-13/4}\right)\right] =$$

$$= \frac{1}{10\times 4^1}[1+0+0+0+0+0+0+0+0+0] = \frac{1}{10\times 4} = 0.025$$

-

$$p_{KDE}(y=10) = \frac{1}{10\times 4^1}[0+0+0+0+0+0+0+0+0+0] = \frac{0}{10\times 4} = 0$$

$$p_{KDE}(y=15) = \frac{1}{10\times 4^1}[0+0+0+0+0+1+1+1+1+0] = \frac{4}{10\times 4} = 0.1$$

Αριθμητικό Παράδειγμα

5 δείγματα: D={2, 2.5, 3, 1, 6}

Υπολογίστε την σ.π.π. για x=3 χρησιμοποιώντας ως συνάρτηση παραθύρου συνάρτηση Gauss με σ=1

Λύση:

x1 = 2

$$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x_1-x)^2}{2}\right)$$

$$= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(2-3)^2}{2}\right) = 0.2420$$

x2 = 2.5

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_2 - x)^2}{2}\right)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(2.5 - 3)^2}{2}\right) = 0.3521$$

Με τον ίδιο τρόπο υπολογίζουμε για x3 = 3, x4 = 1 και x5 = 6 αντίστοιχα:

$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_3 - x)^2}{2}\right) = 0.3989$$

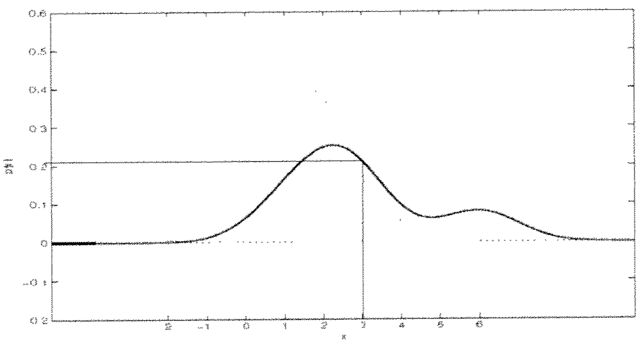
$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_4 - x)^2}{2}\right) = 0.0540$$

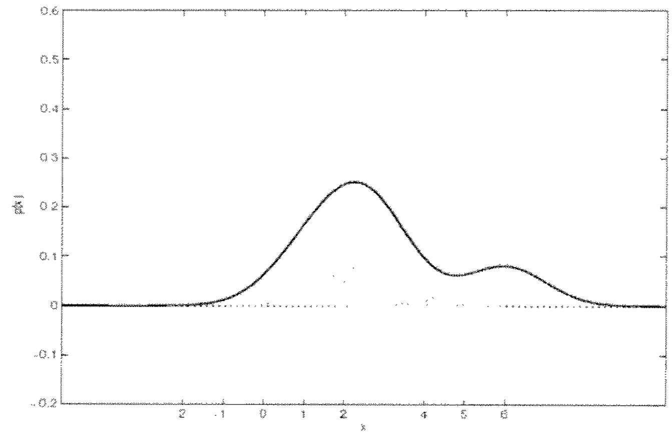
$$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_5 - x)^2}{2}\right) = 0.0044$$

Οπότε:

$$p(x = 3) = (0.2420 + 0.3521 + 0.3989$$

$$+0.0540 + 0.0044)/5 \quad = 0.2103$$





0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
p(x)
-2
-1
0
1
2
4
5
6
x

Κεφάλαιο 5ο

Γραμμικές Διακρίνουσες Συναρτήσεις

Παράδειγμα

	features				grade
name	*good attendance?*	*tall?*	*sleeps in class?*	*chews gum?*	
Jane	*yes (1)*	*yes (1)*	*no (-1)*	*no (-1)*	*A*
Steve	*yes (1)*	*yes (1)*	*yes (1)*	*yes (1)*	*F*
Mary	*no (-1)*	*no (-1)*	*no (-1)*	*yes (1)*	*F*
Peter	*yes (1)*	*no (-1)*	*no (-1)*	*yes (1)*	*A*

- ***class 1***: students who get grade *A*
- ***class 2***: students who get grade *F*

	features					grade
name	*extra*	*good attendance?*	*tall?*	*sleeps in class?*	*chews gum?*	
Jane	*1*	*yes (1)*	*yes (1)*	*no (-1)*	*no (-1)*	*A*
Steve	*1*	*yes (1)*	*yes (1)*	*yes (1)*	*yes (1)*	*F*
Mary	*1*	*no (-1)*	*no (-1)*	*no (-1)*	*yes (1)*	*F*
Peter	*1*	*yes (1)*	*no (-1)*	*no (-1)*	*yes (1)*	*A*

Αντικατάσταση όλων των δειγμάτων της κλάσης c2 από τα αρνητικά τους

$$\mathbf{y}_i \rightarrow -\mathbf{y}_i \qquad \forall \mathbf{y}_i \in c_2$$

Αναζήτηση διανύσματος βαρών *a*, ώστε

$$\mathbf{a}^t \mathbf{y}_i > 0 \qquad \forall \mathbf{y}_i$$

	features					grade
name	*extra*	*good attendance?*	*tall?*	*sleeps in class?*	*chews gum?*	
Jane	*1*	*yes (1)*	*yes (1)*	*no (-1)*	*no (-1)*	*A*
Steve	*-1*	*yes (-1)*	*yes (-1)*	*yes (-1)*	*yes (-1)*	*F*
Mary	*-1*	*no (1)*	*no (1)*	*no (1)*	*yes (-1)*	*F*
Peter	*1*	*yes (1)*	*no (-1)*	*no (-1)*	*yes (1)*	*A*

- Sample is misclassified if $\quad a^t y_i = \sum_{k=0}^{4} a_k y_i^{(k)} < 0$
- gradient descent single sample rule: $a^{(k+1)} = a^{(k)} + \eta^{(k)} y_M$
- Set ***fixed*** learning rate to $\eta^{(k)} = 1$: $\boxed{a^{(k+1)} = a^{(k)} + y_M}$

Θέτουμε ίσα αρχικά βάρη

- set equal initial weights $a^{(1)}$=[0.25, 0.25, 0.25, 0.25]
- visit all samples sequentially, modifying the weights for after finding a misclassified example

name	$a^t y$	*misclassified?*
Jane	*0.25*1+0.25*1+0.25*1+0.25*(-1)+0.25*(-1) >0*	*no*
Steve	*0.25*(-1)+0.25*(-1)+0.25*(-1)+0.25*(-1)+0.25*(-1)<0*	*yes*

- new weights

$$a^{(2)} = a^{(1)} + y_M = [0.25 \;\; 0.25 \;\; 0.25 \;\; 0.25 \;\; 0.25] +$$
$$+ [-1 \;\; -1 \;\; -1 \;\; -1 \;\; -1] =$$
$$= [-0.75 \;\; -0.75 \;\; -0.75 \;\; -0.75 \;\; -0.75]$$

$$a^{(2)} = [-0.75 \;\; -0.75 \;\; -0.75 \;\; -0.75 \;\; -0.75]$$

name	$a^t y$	*misclassified?*
Mary	*-0.75*(-1)-0.75*1 -0.75 *1 -0.75 *1 -0.75*(-1) <0*	*yes*

- new weights

$$a^{(3)} = a^{(2)} + y_M = [-0.75 \;\; -0.75 \;\; -0.75 \;\; -0.75 \;\; -0.75] +$$
$$+ [-1 \;\; 1 \;\; 1 \;\; 1 \;\; -1] =$$
$$= [-1.75 \;\; 0.25 \;\; 0.25 \;\; 0.25 \;\; -1.75]$$

$$a^{(3)} = [-1.75 \;\; 0.25 \;\; 0.25 \;\; 0.25 \;\; -1.75]$$

name	$a^t y$	*misclassified?*
Peter	*-1.75 *1 +0.25* 1+0.25* (-1) +0.25 *(-1)-1.75*1 <0*	*yes*

- new weights

$$a^{(4)} = a^{(3)} + y_M = [-1.75 \;\; 0.25 \;\; 0.25 \;\; 0.25 \;\; -1.75] +$$
$$+ [1 \;\; 1 \;\; -1 \;\; -1 \;\; 1] =$$
$$= [-0.75 \;\; 1.25 \;\; -0.75 \;\; -0.75 \;\; -0.75]$$

$$a^{(4)} = [-0.75 \;\; 1.25 \;\; -0.75 \;\; -0.75 \;\; -0.75]$$

name	$a^t y$	misclassified?
Jane	-0.75 *1 +1.25*1 -0.75*1 -0.75 *(-1) -0.75 *(-1)+0	no
Steve	-0.75*(-1)+1.25*(-1) -0.75*(-1) -0.75*(-1)-0.75*(-1)>0	no
Mary	-0.75 *(-1)+1.25*1-0.75*1 -0.75 *1 –0.75*(-1) >0	no
Peter	-0.75 *1+ 1.25*1-0.75* (-1)-0.75* (-1) -0.75 *1 >0	no

- Thus the discriminant function is

$$g(y) = -0.75 * y^{(0)} + 1.25 * y^{(1)} - 0.75 * y^{(2)} - 0.75 * y^{(3)} - 0.75 * y^{(4)}$$

- Converting back to the original features ***x***:

$$g(x) = 1.25 * x^{(1)} - 0.75 * x^{(2)} - 0.75 * x^{(3)} - 0.75 * x^{(4)} - 0.75$$

- Converting back to the original features ***x***:

$$1.25 * x^{(1)} - 0.75 * x^{(2)} - 0.75 * x^{(3)} - 0.75 * x^{(4)} > 0.75 \Rightarrow \textit{grade A}$$
$$1.25 * x^{(1)} - 0.75 * x^{(2)} - 0.75 * x^{(3)} - 0.75 * x^{(4)} < 0.75 \Rightarrow \textit{grade F}$$

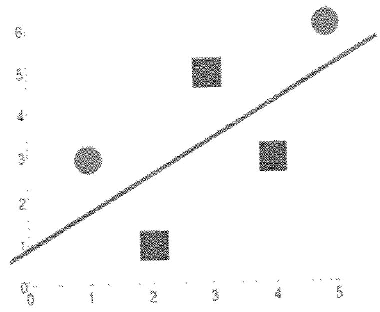
good attendance   tall   sleeps in class   chews gum

- This is just one possible solution vector
- If we started with weights ***a***$^{(1)}$=[0,0.5, 0.5, 0, 0], solution would be [-1,1.5, -0.5, -1, -1]

$$1.5 * x^{(1)} - 0.5 * x^{(2)} - x^{(3)} - x^{(4)} > 1 \Rightarrow \textit{grade A}$$
$$1.5 * x^{(1)} - 0.5 * x^{(2)} - x^{(3)} - x^{(4)} < 1 \Rightarrow \textit{grade F}$$

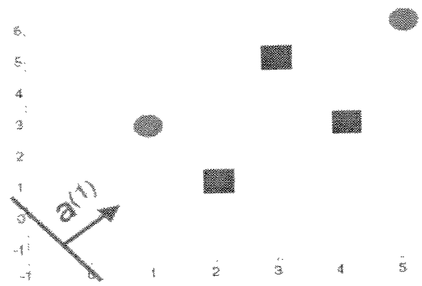
  - In this solution, being tall is the least important feature

Παράδειγμα - Μη γραμμικά διαχωρίσιμα δεδομένα



- Suppose we have 2 features and samples are:
  - Class 1: [2,1], [4,3], [3,5]
  - Class 2: [1,3] and [5,6]
- These samples are not separable by a line
- Still would like to get approximate separation by a line, good choice is shown in green
  - some samples may be "noisy", and it's ok if they are on the wrong side of the line
- Get y_1, y_2, y_3, y_4 by adding extra feature and "normalizing"

$$y_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \quad y_2 = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix} \quad y_3 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} \quad y_4 = \begin{bmatrix} -1 \\ -1 \\ -3 \end{bmatrix} \quad y_5 = \begin{bmatrix} -1 \\ -5 \\ -6 \end{bmatrix}$$



- Let's apply Perceptron single sample algorithm
- initial equal weights $a^{(1)} = [1 \;\; 1 \;\; 1]$
  - this is line $x^{(1)} + x^{(2)} + 1 = 0$
- fixed learning rate $\eta = 1$

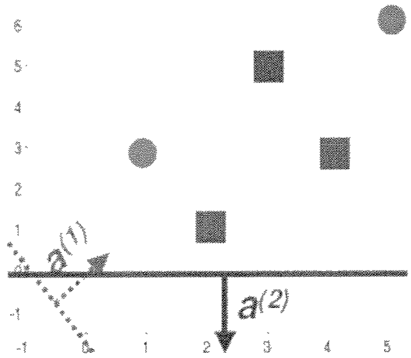
$$a^{(k+1)} = a^{(k)} + y_M$$

$$y_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \quad y_2 = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix} \quad y_3 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} \quad y_4 = \begin{bmatrix} -1 \\ -1 \\ -3 \end{bmatrix} \quad y_5 = \begin{bmatrix} -1 \\ -5 \\ -6 \end{bmatrix}$$

- $y^t_1 a^{(1)} = [1 \;\; 1 \;\; 1]^*[1 \;\; 2 \;\; 1]^t > 0$ ✓
- $y^t_2 a^{(1)} = [1 \;\; 1 \;\; 1]^*[1 \;\; 4 \;\; 3]^t > 0$ ✓
- $y^t_3 a^{(1)} = [1 \;\; 1 \;\; 1]^*[1 \;\; 3 \;\; 5]^t > 0$ ✓

$$a^{(1)} = [1\ 1\ 1] \qquad a^{(k+1)} = a^{(k)} + y_M$$

$$y_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \quad y_2 = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix} \quad y_3 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} \quad y_4 = \begin{bmatrix} -1 \\ -1 \\ -3 \end{bmatrix} \quad y_5 = \begin{bmatrix} -1 \\ -5 \\ -6 \end{bmatrix}$$



- $y^t_4 a^{(1)}$=[1 1 1]*[-1 -1 -3]^t= -5< 0

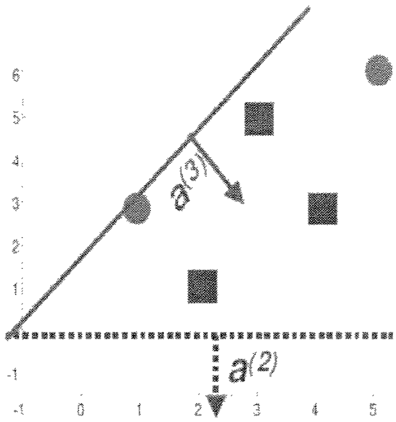
$$a^{(2)} = a^{(1)} + y_M = [1\ 1\ 1] + [-1\ -1\ -3] = [0\ 0\ -2]$$

- $y^t_5 a^{(2)}$=[0 0 -2]*[-1 -5 -6]^t= 12 > 0 ✓
- $y^t_1 a^{(2)}$=[0 0 -2]*[1 2 1]^t < 0

$$a^{(3)} = a^{(2)} + y_M = [0\ 0\ -2] + [1\ 2\ 1] = [1\ 2\ -1]$$

$$a^{(3)} = [1\ 2\ -1] \qquad a^{(k+1)} = a^{(k)} + y_M$$

$$y_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \quad y_2 = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix} \quad y_3 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} \quad y_4 = \begin{bmatrix} -1 \\ -1 \\ -3 \end{bmatrix} \quad y_5 = \begin{bmatrix} -1 \\ -5 \\ -6 \end{bmatrix}$$

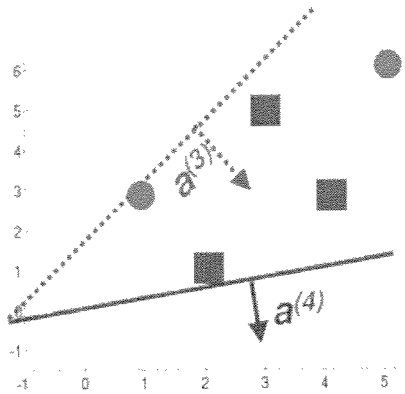


- $y^t_2 a^{(3)}$=[1 4 3]*[1 2 -1]^t=6 > 0 ✓
- $y^t_3 a^{(3)}$=[1 3 5]*[1 2 -1]^t > 0 ✓
- $y^t_4 a^{(3)}$=[-1 -1 -3]*[1 2 -1]^t = 0

$$a^{(4)} = a^{(3)} + y_M = [1\ 2\ -1] + [-1\ -1\ -3] = [0\ 1\ -4]$$

$$a^{(4)} = [0\ 1\ -4] \qquad a^{(k+1)} = a^{(k)} + y_M$$

$$y_1 = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \quad y_2 = \begin{bmatrix} 1 \\ 4 \\ 3 \end{bmatrix} \quad y_3 = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix} \quad y_4 = \begin{bmatrix} -1 \\ -1 \\ -3 \end{bmatrix} \quad y_5 = \begin{bmatrix} -1 \\ -5 \\ -6 \end{bmatrix}$$



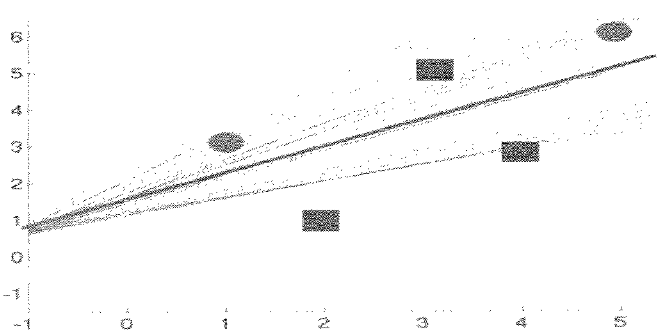
- $y^t_2\, a^{(3)} = [1\ 4\ 3]^*[1\ 2\ -1]^t = 6 > 0$ ✓
- $y^t_3\, a^{(3)} = [1\ 3\ 5]^*[1\ 2\ -1]^t > 0$ ✓
- $y^t_4\, a^{(3)} = [-1\ -1\ -3]^*[1\ 2\ -1]^t = 0$

$$a^{(4)} = a^{(3)} + y_M = [1\ 2\ -1] + [-1\ -1\ -3] = [0\ 1\ -4]$$

Μπορούμε να συνεχίσουμε για πάντα !!
Δεν υπάρχει διάνυσμα λύσης **a** το οποίο να ικανοποιεί για κάθε **i** τη σχέση:

$$a^t y_i = \sum_{k=0}^{5} a_k y_i^{(k)} > 0$$

Πρέπει να σταματήσουμε αλλά σε «καλό» σημείο.
Οι λύσεις για επαναλήψεις 900 ως 915.
Κάποιες είναι καλές, κάποιες όχι
Πώς θα σταματήσουμε σε καλή λύση?



Class 1: (6 9), (5 7)
Class 1: (5 9), (0 10)

$$Y = \begin{bmatrix} 1 & 6 & 9 \\ 1 & 5 & 7 \\ -1 & -5 & -9 \\ -1 & 0 & -10 \end{bmatrix}$$

Αρχικές τιμές:

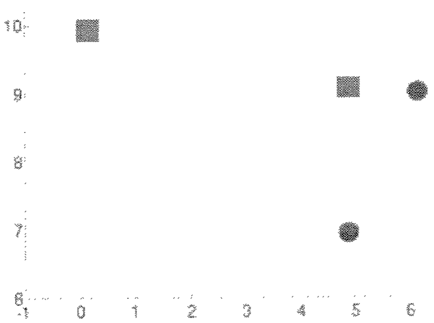
$$a^{(1)} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} b^{(1)} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Θέτουμε :

$$\eta = 0.9$$

Οπότε:

$$Ya^{(1)} = \begin{bmatrix} 16 \\ 13 \\ -15 \\ -11 \end{bmatrix}$$



Πρώτη επανάληψη:

$$e^{(1)} = Ya^{(1)} - b^{(1)} = \begin{bmatrix} 16 \\ 13 \\ -15 \\ -11 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 15 \\ 12 \\ -16 \\ -12 \end{bmatrix}$$

Επίλυση για $b^{(2)}$ χρησιμοποιώντας $a^{(1)}$ και $b^{(1)}$

$$b^{(2)} = b^{(1)} + 0.9\left[e^{(1)} + |e^{(1)}|\right] = \begin{bmatrix}1\\1\\1\\1\end{bmatrix} + 0.9\left[\begin{bmatrix}15\\12\\-16\\-12\end{bmatrix} + \begin{bmatrix}15\\12\\16\\12\end{bmatrix}\right] = \begin{bmatrix}28\\22.6\\1\\1\end{bmatrix}$$

Επίλυση για $a^{(2)}$ χρησιμοποιώντας $b^{(2)}$ και

$$a^{(2)} = \left(Y^tY\right)^{-1}Y^t\, b^{(2)} = \begin{bmatrix}-2.6 & 4.7 & 1.6 & -0.5\\ 0.16 & -0.1 & -0.1 & 0.2\\ 0.26 & -0.5 & -0.2 & -0.1\end{bmatrix} * \begin{bmatrix}28\\22.6\\1\\1\end{bmatrix} = \begin{bmatrix}34.6\\2.7\\-3.8\end{bmatrix}$$

Συνεχίζουμε τις επαναλήψεις ώσπου

$$Ya > 0$$

Στη πράξη, συνεχίζουμε ώσπου το μικρότερο στοιχείο του Ya να είναι μικρότερο του 0.01

Μετά απο 104 επαναλήψεις συγκλίνει σε λύση

$$a = \begin{bmatrix}-34.9\\27.3\\-11.3\end{bmatrix} \qquad b = \begin{bmatrix}28\\23\\1\\147\end{bmatrix}$$

Το διάνυσμα *a* ορίζει ένα διαχωριστικό υπερεπίπεδο

$$Ya = \begin{bmatrix}27.2\\22.5\\0.14\\1.48\end{bmatrix}$$