13/3/2025

Παρασκευή 14/3 : Άσκουσα Γ εργαστήριο

Classification Problems

$Y \in \{g_1, g_2, \dots, g_K\}$

$X = (X_1, \dots, X_p)$ ανεξάρτητες μεταβλητές

(X, Y) από κοινή (άγνωστη) κατανομή

$P(Y = g_k \mid X = x) \quad k = 1, 2, \dots, K$

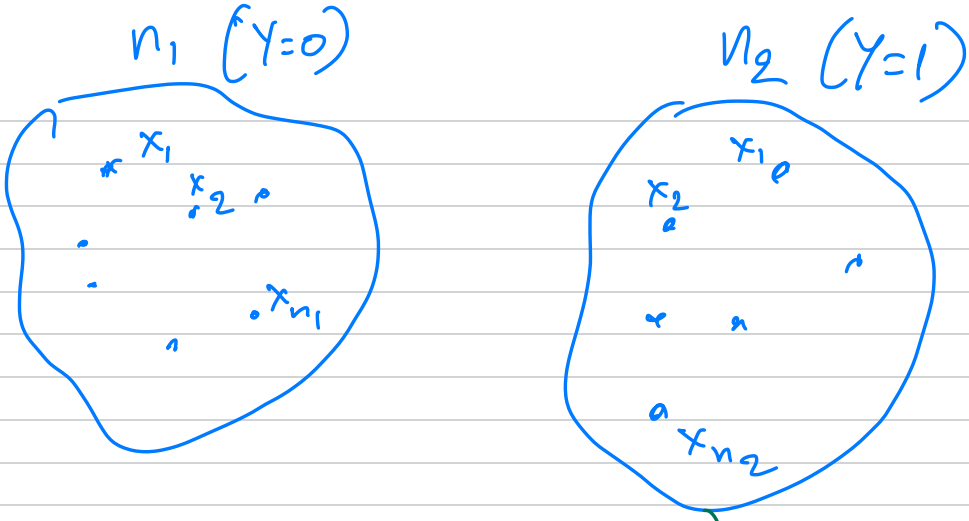
$F(X \mid Y = g_k) = P(X \le x \mid Y = g_k)$ (για $p=1$, X συνεχής $f(x \mid g_k) =$ δεν μπορεί να οπτ)

π.χ. $X = (X_1, \dots, X_p)$ (μετρήσεις αίματος/ασθενούς)

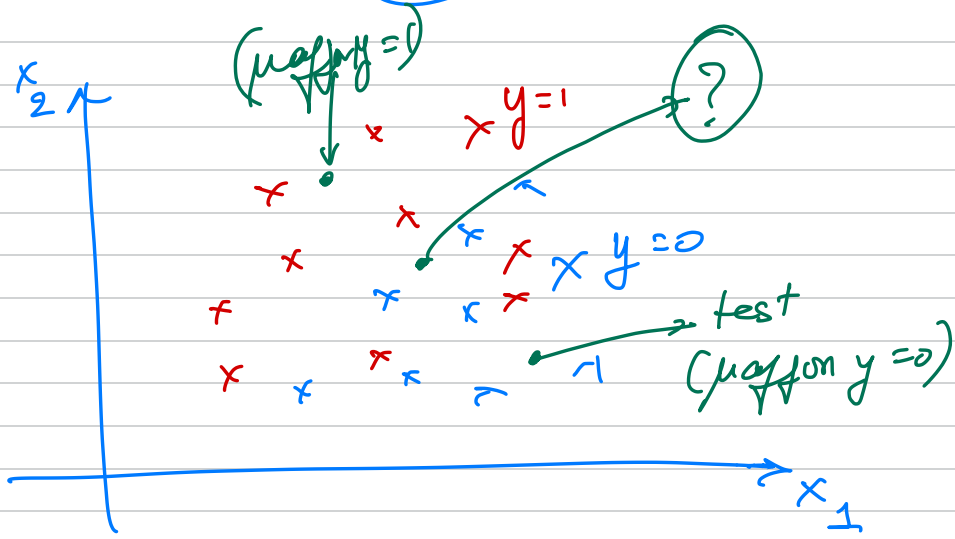
$Y = \begin{cases} 1, & \text{έχει ασθένεια} \\ 0, & \text{υγιής} \end{cases}$

$P(Y = 1 \mid X = x)$ ← δίνει τον πιθανότητα νόσησης δεδομένων των μετρήσεων (θέλουμε να το εκτιμήσουμε ⇔ "πρόβλεψη")

$F(X \mid Y = 0)$: κατανομή X για υγιείς
$F(X \mid Y = 1)$ " X " ασθενείς } ⇒ κλινικές δοκιμές (case-control)



$X = (X_1, X_{\alpha\kappa})$



Έστω $P(Y=1) = p$
$P(Y=0) = 1-p$
} περιθώρια (Y)

Bayes $P(Y=1 \mid X=x) = \dfrac{p \cdot f(x \mid y=1)}{p f(x \mid y=1) + (1-p) f(x \mid y=0)}$

Κριτήριο / Συνάρτηση Απώλειας

Γενικό μοντέλο $Y \in \{g_1, \dots, g_K\}$

Training set $(x_1, y_1), \dots, (x_n, y_n) \Rightarrow \dots \Rightarrow$ (μεθοδολογία)

$\hat{G}(x)$: προβλεπτική συνάρτηση

$\hat{G}(x) \in \{g_1, \dots, g_K\}$

} πόσο καλή είναι η $\hat{G}(x)$?

Test Set $\{(x_0, y_0), \dots\}$

$\hat{y}_i = \hat{G}_i(x_i)$ $i \in$ TestSet.

Κριτήριο (Συνάρτηση απώλειας) = $L(y, \hat{G})$

Έστω $L(y_i, \hat{y}_i) = \begin{cases} 1, & y_i \neq \hat{y}_i \\ 0, & y_i = \hat{y}_i \end{cases} = 1 - I(y_i = \hat{y}_i)$

$(i \in \text{Test})$ $\qquad \delta_{y_i, \hat{y}_i}$

$$L(y, \hat{G}) = \frac{\sum_{i \in \text{Test}} L(y_i, \hat{G}(x_i))}{|\text{Test set}|} = \text{\% λάθος ταξινομήσεων}$$

(misclassification rate)

$$\text{MPE} = E_{(X,Y)}\left(L(Y, \hat{G}(X))\right) \leftarrow$$

Για $L = 1 - I(y_i = \hat{y}_i)$

$\min_{\hat{G}\in\{\text{class}\}}$

$$E\left[L(Y,\hat{G}(X))\right] = E_X\left[E_{Y|X}\left(L(Y,\hat{G}(X))\,|\,X\right)\right]$$

$$= E_X\left[\sum_{k=1}^{K} P(Y=g_k|X)\cdot L(g_k,\hat{G}(X))\right]$$

Bayes risk.

Σημειακή βελτιστοποίηση

$$\forall x: \min_{\hat{G}(x)} \sum_{k=1}^{K} P(Y=g_k|X=x)\cdot L(g_k,\hat{G}(x)) \qquad \hat{G}\in\{g_1,\dots,g_K\}$$

Αν $L(y,\hat{y}) = 1 - I(y=\hat{y})$

Τότε $\sum_{k=1}^{K} P(Y=g_k|X=x)\,L(g_k,\hat{G}(x)) =$

$$= \sum_{g_k\neq\hat{G}(x)} P(Y=g_k|X=x)\cdot 1 = \boxed{1-P(Y=\hat{G}(x)|X=x)}$$

Βέλτιστη πρόβλεψη $|X=x$

$$\min_{\hat{G}}\left\{1-P(Y=\hat{G}|X=x)\right\} \Leftrightarrow \boxed{\max_{\hat{G}} P(Y=\hat{G}|X=x)}$$

$$\Rightarrow \hat{G}(x) = \underset{g\in\{g_1,\dots,g_K\}}{\operatorname{argmax}} P(Y=g|X=x)$$

Bayes classifier

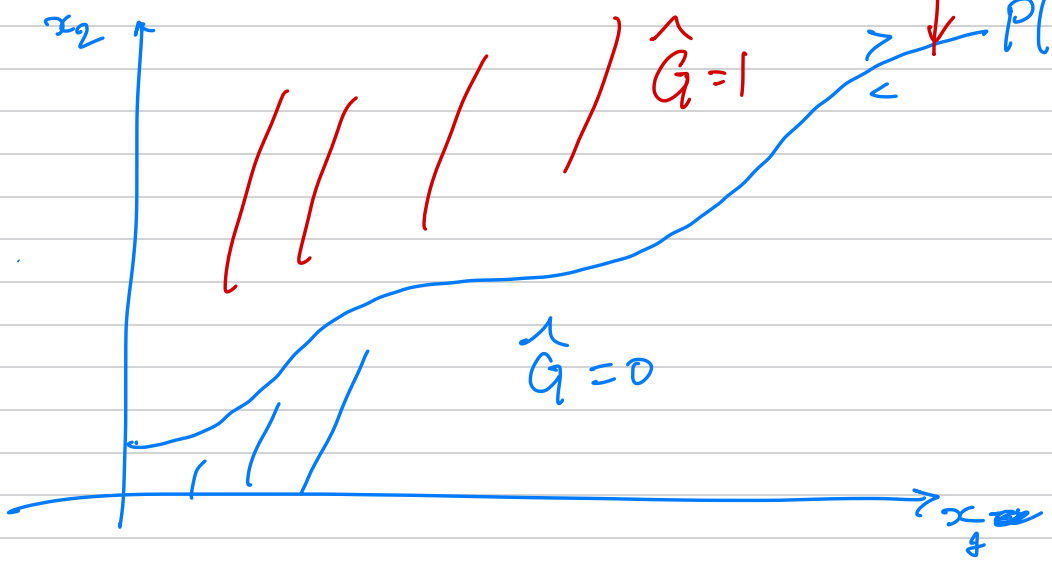
Για $X=(X_1,X_2)$, $Y\in\{0,1\}$

$$\hat{G}(x)=\begin{cases}0, & P(Y=1|X=x)<1/2\\ 1, & P(Y=1|X=x)>1/2\end{cases}$$

Τύπος Bayes

Bayes boundary ΑΠΟΣΤΟ!!

$P(Y=1|x)=1/2$



Παραμετρικά μοντέλα

1) Logistic Regression ($Y\in\{0,1\}$)

$\rightarrow P(Y=1|X)=p(x)$

$$\log\frac{p(x)}{1-p(x)}=b_0+b_1^T\cdot x$$

μοντέλο

2) Generative Models (Discriminant Analysis)

$F(x|0)\sim$ .... αρνητικά παραδείγματα

$F(x|1)\sim$ .... " θετικά παραδείγματα

Training set $\Rightarrow$ ... $\left.\begin{matrix}\hat{F}(x|0)\\ \hat{F}(x|1)\end{matrix}\right\}\Rightarrow$ Bayes $\cdot\ \hat{P}(Y=1|X=x)$

π.χ. Αν $X = X_1$

$X|Y=0 \quad X \sim N(\mu_0, \sigma^2)$

$X|Y=1 \quad X \sim N(\mu_1, \sigma^2)$ με $(\mu_1 > \mu_0)$

$$f(x|y=0) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu_0)^2}{2\sigma^2}}$$

$$f(x|y=1) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu_1)^2}{2\sigma^2}}$$

$$P(Y=1|X=x) = \frac{p\, f(x|1)}{p\, f(x|1) + (1-p) f(x|0)}$$

$$P(Y=0|X=x) = \frac{(1-p) f(x|0)}{\quad}$$

Bayes classifier

$$\hat{G} = 1 \quad \text{αν} \quad p f(x|1) > (1-p) f(x|0) \Rightarrow$$

$$\Leftrightarrow \frac{f(x|1)}{f(x|0)} > \frac{1-p}{p} \Leftrightarrow$$

$$\log \frac{f(x|1)}{f(x|0)} > \log \frac{1-p}{p} = A$$

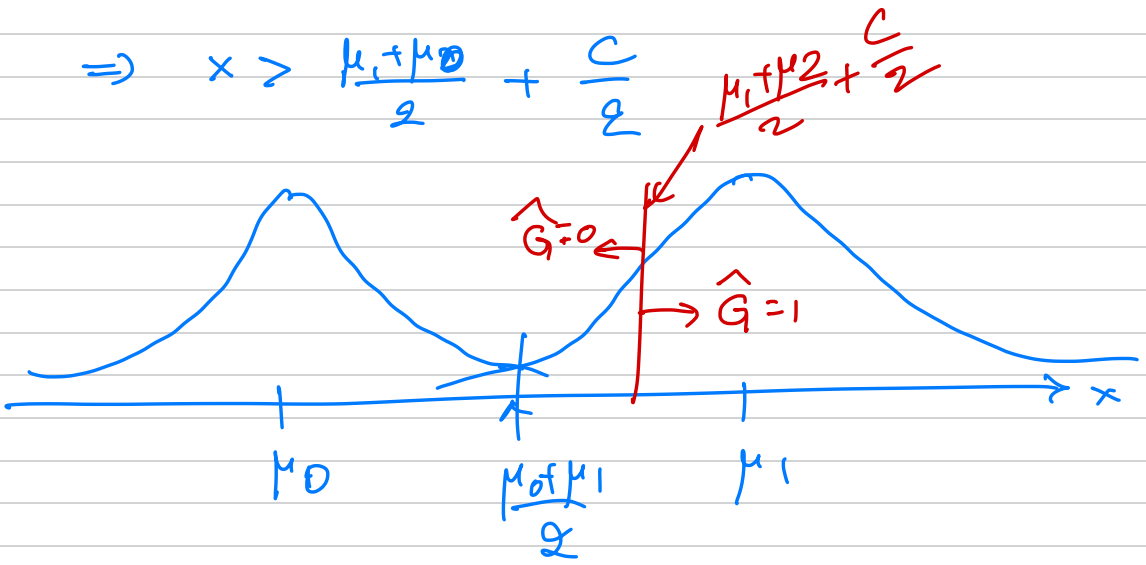
$$\log f(x|1) - \log f(x|0) > A \Leftrightarrow$$

$$\Leftrightarrow \frac{(x-\mu_0)^2}{2\sigma^2} - \frac{(x-\mu_1)^2}{2\sigma^2} > A \Leftrightarrow$$

$$\underbrace{(\mu_1-\mu_0)}_{\geq 0}(2x-(\mu_1+\mu_0)) > 2A\sigma^2 \Rightarrow$$

$$\Rightarrow \quad 2x-(\mu_1+\mu_0) > \frac{2A\sigma^2}{\mu_1-\mu_0} = C \Rightarrow$$

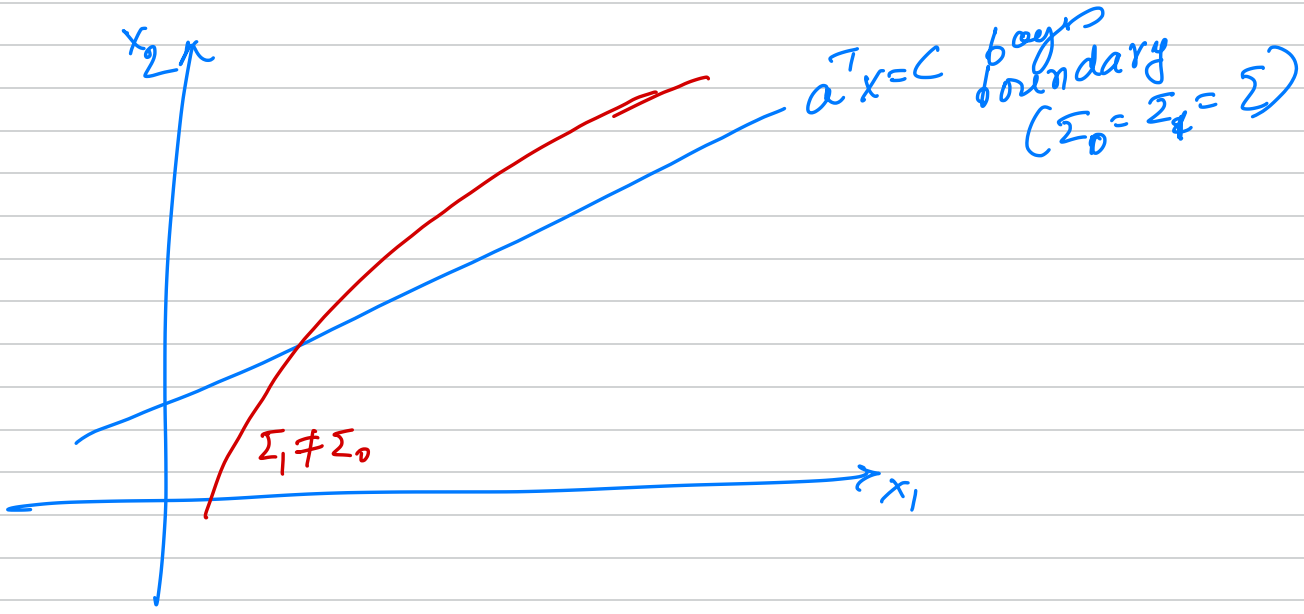
$$\Rightarrow \quad x > \frac{\mu_1+\mu_0}{2} + \frac{C}{2}$$



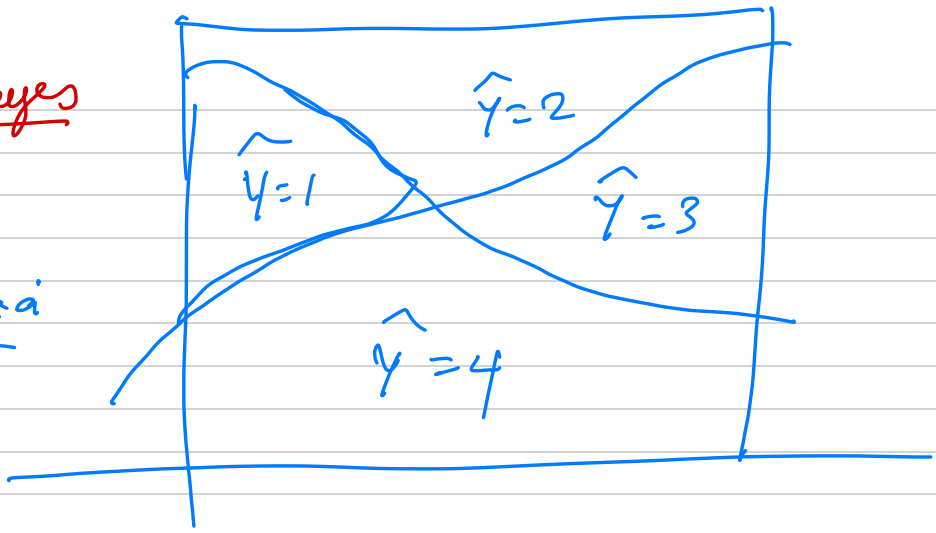
Av $X=(X_1,X_2)$

$$\left.\begin{aligned} X\,|\,Y=0 &\sim \mathcal{N}(\underline{\mu}_0, \Sigma) \\ X\,|\,Y=1 &\sim \mathcal{N}(\underline{\mu}_1, \Sigma) \end{aligned}\right\} \Rightarrow \dots$$

$$P(Y=1\,|\,X=x) > P(Y=0\,|\,X=x) \Leftrightarrow \boxed{a^T x > C}$$



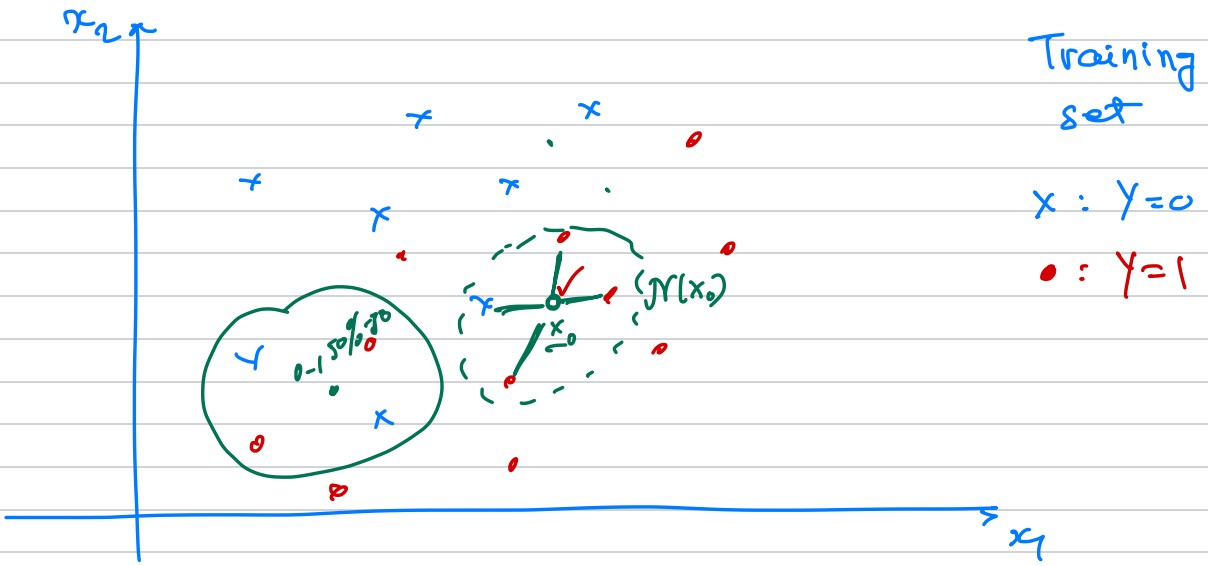
Σύνορο Bayes



γραμμικά

Μέθοδος knn (k nearest neighbor)

Εκτίμηση $P(Y=j \mid X=x_0)$ με παραμετρικά



Training set

x : $Y=0$

$\bullet$: $Y=1$

Έστω $K=4$

$N_0(x_0) = \{4$ κοντινότερα σημεία του training set στο x_0 (ως προς $x)\}$

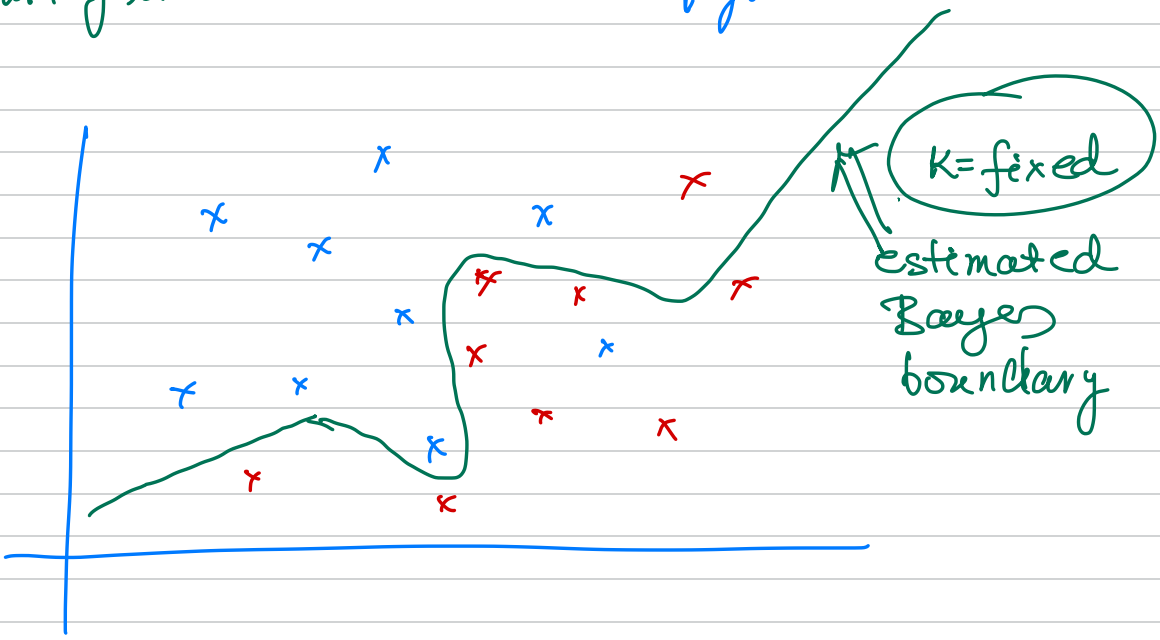
κοντινότερα : $\|x-x_0\|$

$|N_0(x_0)| = K$ τ.ω. $\|x_i-x_0\| < \|x_j-x_0\| \quad \forall i \in N_0, j \notin N_0$.

$\hat{G}(x_0) = g_i$ αν $\frac{1}{K}\sum_{j=1}^{K} I(y_j = g_i) \leftarrow \max_i$

↓ εξαρτάται από training set

% παρατ. στο $\mathcal{N}_0$ που είναι $y=g_i$



$K=1$ η' $K \to n$ $(n = |\text{Training set}|)$

ευέλιξα ⟵ (+ —)

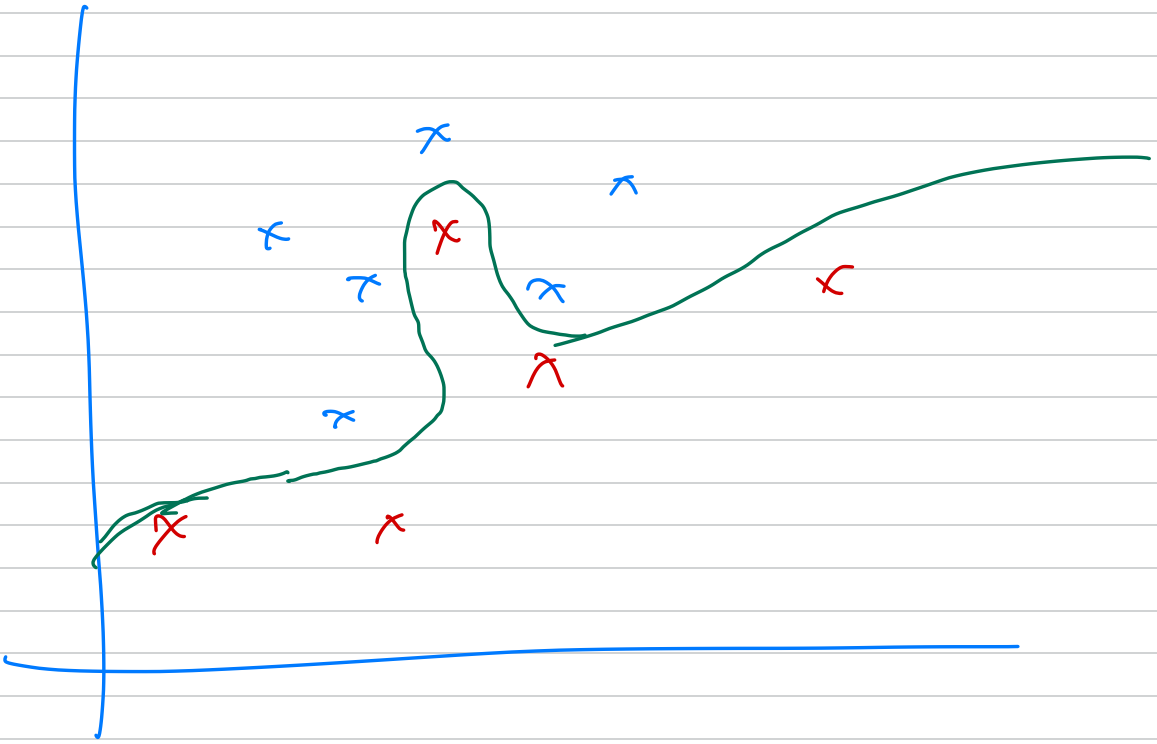
1 Όταν $K \to n$ (μεγάλο)

$\forall x_0$: % $(Y=1)$ = σταθερό = % $(Y=1)$ στο training set $= q_1$

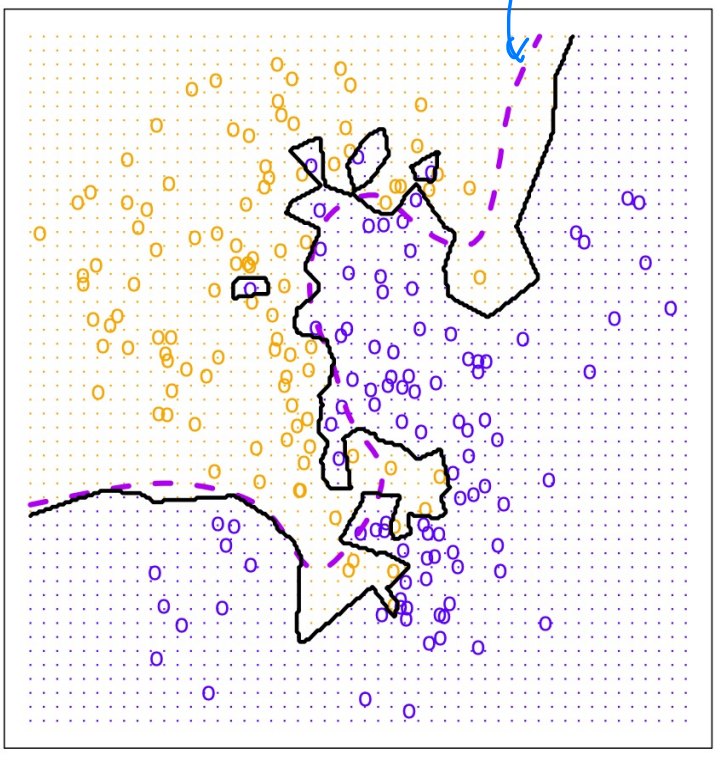
Αν $q_1 > 1/2 \Rightarrow \hat{G}(x_0) = 1 \ \forall x_0$ } ελάχιστη ευελιξία

$q_1 < 1/2 \quad \hat{G}(x_0) = 0$

2 Ozav $k \to 1$

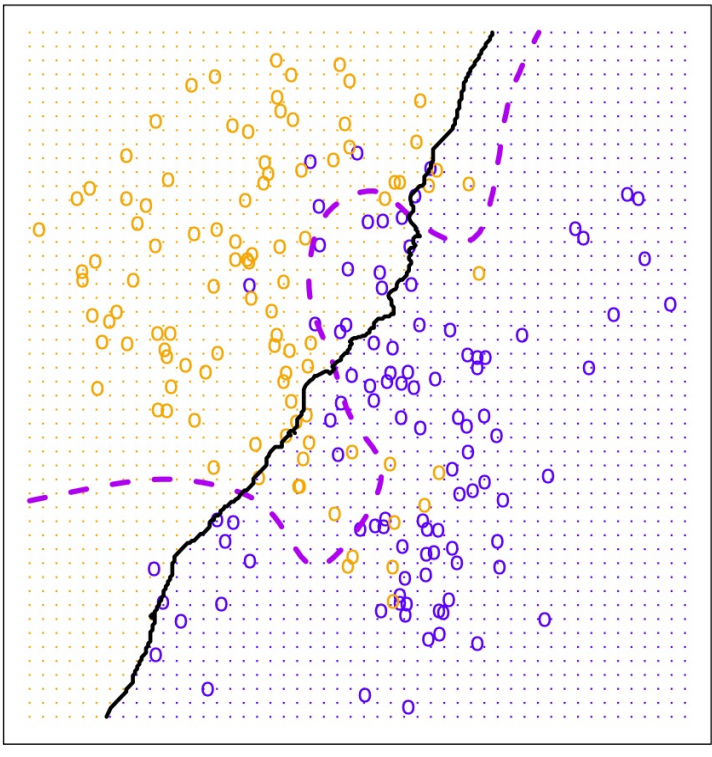


KNN: K=1



Granica nivelo Bayes

KNN: K=100



Παράδειγμα

$Y \in \{0,1\}$

$X = (X_1, X_2)$

$(X_1, X_2) \overset{iid}{\sim} U(0,1)$

$$P(Y=1 \mid X=(x_1,x_2)) = p(x_1,x_2)$$

$$Y \mid (X_1=x_1, X_2=x_2) \sim Ber(p(x_1,x_2))$$

Προσομοίωση

Υποθέτω ακόλουθη συνάρτηση:

$$p(x_1,x_2) = \begin{cases} \frac{1}{2}\left(\frac{x_2}{b(x_1)}\right)^{\alpha}, & x_2 < b(x_1) \\ \frac{1}{2}\left(\frac{1-x_2}{1-b(x_1)}\right)^{\alpha}, & x_2 > b(x_1) \end{cases}$$

όπου $b(x_1) = 0.7 + 0.1x_1 - 0.8x_1^2 + 4x_1^3 - 2.8x^4$

Bayes boundary: $\{(x_1,x_2) : p(x_1,x_2) = \frac{1}{2}\}$

Αλγόριθμος Προσομοίωσης

Έστω $n = 1000$ (μέγεθος training set)

$\alpha = 2$

$K = 10$ (κοντά)

1) Δημιουργώ N- training sets μέσω προσομοίωσης από την παραπάνω κατανομή. $(N \to \infty)$

2) Δημιουργώ 1 test set μεγέθους
$n_{test} = 100$ από την ίδια κατανομή.

3) Για κάθε training set $r = 1, 2, \ldots, N$

(a) Εφαρμόζω τον knn ($k=10$) στο r

(b) βρίσκω προβλέψεις $\hat{y}_j$ $j \in$ Test set

(c) υπολογίζω % misclassification

% λαθών στο test set όπου $\hat{y} = y$

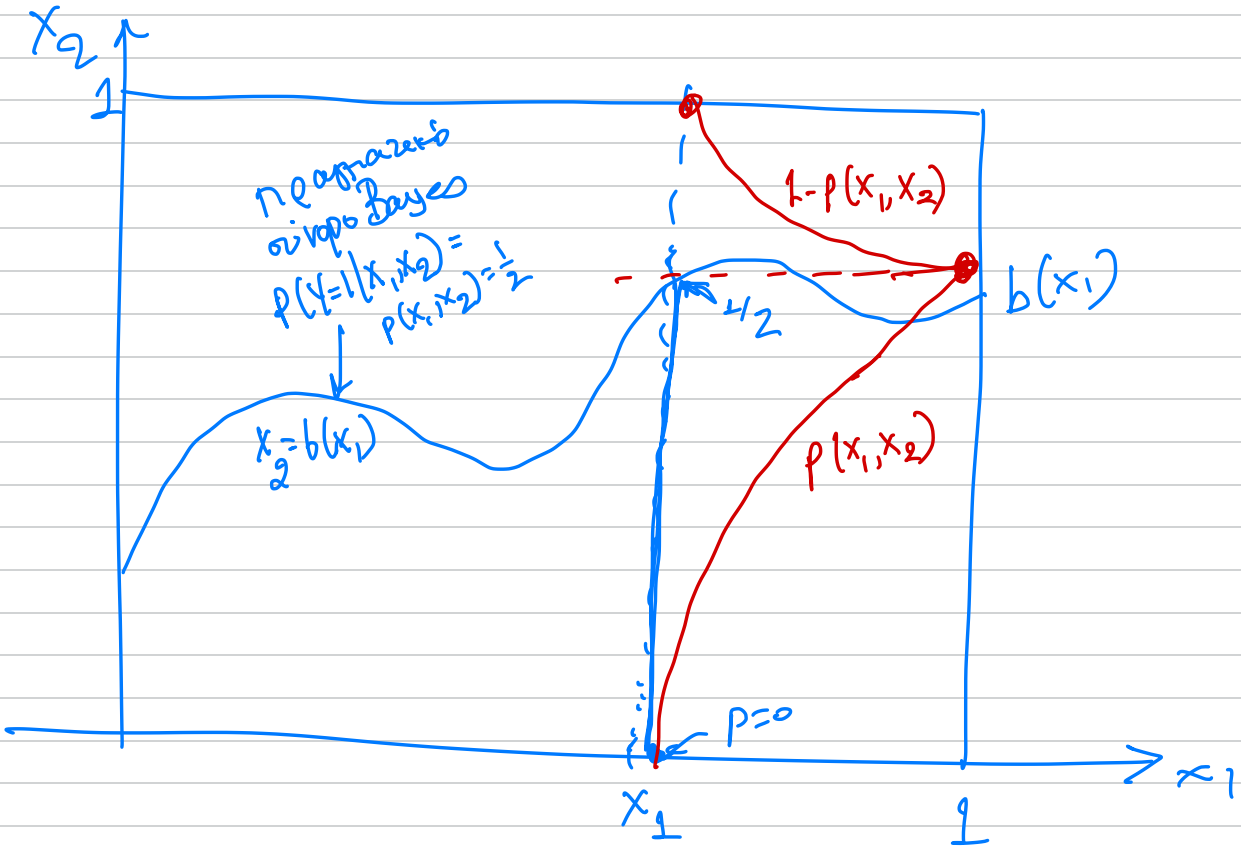
MPE_r

4) $$\widehat{MPE} = \frac{1}{N} \sum_{r=1}^{N} MPE_r$$

MPE_r, $r = 1, 2, \ldots$ i.i.d

$$\widehat{MPE} \xrightarrow[N \to \infty]{} MPE \quad (\text{ν.ν. } 1)$$

(Monte Carlo Simulation)



$$p(x_1,x_2)=\begin{cases}\left(\frac{x_2}{b(x_1)}\right)^2/2 & x_2<b(x_1)\\ 1-\left(\frac{1-x_2}{1-b(x_1)}\right)^2/2 & x_2>b(x_1)\end{cases}$$