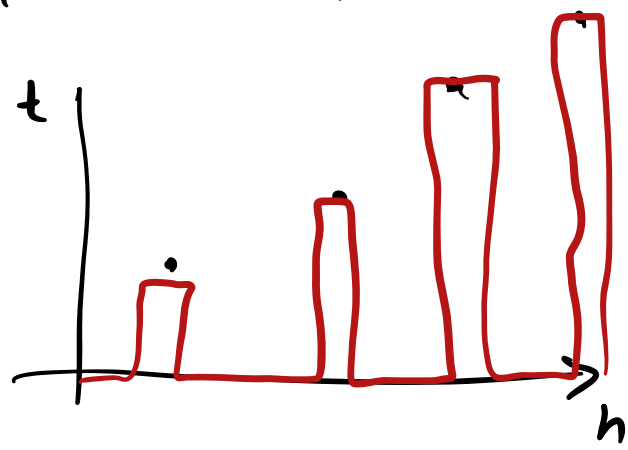
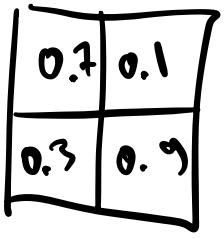
Χρόνος πτώσης

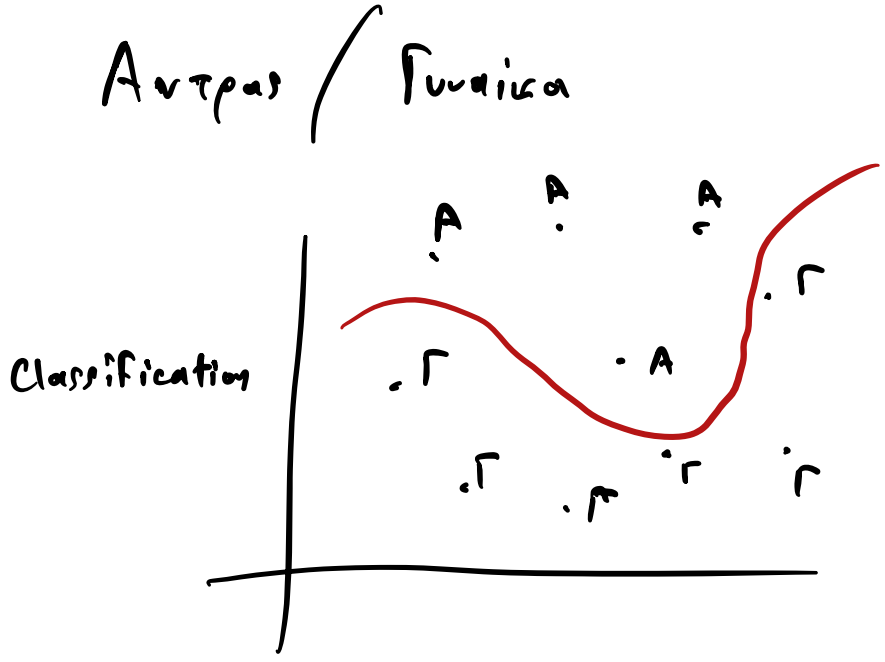



Regression

Φωτογραφίες




Άντρας / Γυναίκα




Μοντελοποίηση

$(x, y) \sim D$

Features

$x \in X$ (domain)

$y \in Y$

Στόχος είναι να μάθουμε συνάρτηση

$h : X \to Y$ ώστε να ελαχιστοποιείται το $\underset{(x,y)\sim D}{E}[\ell(h(x), y)]$

για κάποια συνάρτηση κόστους $\ell$

- Παραδείγματα: Για regression

$$L(\hat{y}, y) = (\hat{y} - y)^2$$

Για classification

$$L(\hat{y}, y) = 1\{\hat{y} \neq y\} = \begin{cases} 1 & \text{αν } \hat{y} \neq y \\ 0 & \text{αλλιώς} \end{cases}$$

Training set

$$S = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$$

όπου $(x_i, y_i) \sim D$

Κλάση υποθέσεων $H$

$$\rightarrow \min_{h \in H} \underbrace{\mathop{E}_{(x,y)\sim D}[\ell(h(x), y)]}_{L_D(h)}$$

Empirical Risk Minimization

$$\min_{h \in H} \underbrace{\mathop{E}_{(x,y)\sim S}[\ell(h(x), y)]}_{L_S(h)} \rightarrow h_S \in ERM$$

$$h(x) = \begin{cases} y & \text{αν υπάρχει } (x,y) \in S \\ 0 & \text{αλλιώς} \end{cases}$$

overfitting

Απλή περίπτωση

$$L_D(h^*) = 0$$
$$L_D(h) \in [0,1]$$

$$\ell(h(x), y) = 1[h(x) \neq y]$$

Θέλω να βρω $h \in H$

$$\Pr_{(x,y)\sim D}[h(x) \neq y] \leq \varepsilon$$

πραγματικό loss

Έστω $H_B$ όλες οι συναρτήσεις με

$$\Pr_{(x,y)\sim D}[h(x) \neq y] > \varepsilon$$

Για κάθε $h \in H_B$

Η πιθανότητα να μην την έχω απορρίψει μετά από $m$ δείγματα είναι το πολύ $(1-\varepsilon)^m \leq e^{-m\varepsilon}$

Αν $m = \frac{\log(1/\delta)}{\varepsilon}$ απορρίπτω την κακή $h$ με πιθανότητα $\geq 1-\delta$

Αν το $H$ είναι πεπερασμένο

$$\Pr[\exists h \in H_B \text{ που επιβιώνει}]$$
$$\leq \sum_{h \in H_B} \Pr[h \text{ επιβιώνει}]$$
$$\leq |H_B|\,\delta \leq \underbrace{|H|\,\delta}_{\delta'}$$

$$m = \frac{\log(|H|/\delta')}{\varepsilon}$$

τότε η $h_S \in H \setminus H_B$ με πιθανότητα $1-\delta'$

Γενίκευση μέσω Uniform Convergence

Στόχος είναι να μάθουμε καλή συνάρτηση:

$$L_D(h_S) - L_D(h^*) \leq \varepsilon$$

Uniform Convergence

$$\max_{h \in H} |L_S(h) - L_D(h)| \leq \varepsilon$$

Ξέρουμε

$$\mathop{E}_{S}[L_S(h)] = L_D(h)$$

$\forall h$

Γιατί είναι καλό το Uniform Convergence;

$$L_D(h_S) - L_D(h^*) = L_D(h_S) - L_S(h_S)$$

$$+ L_S(h_S) - L_S(h^*)$$

$$+ L_S(h^*) - L_D(h^*)$$

$$\leq |L_D(h_S) - L_S(h_S)|$$

$$+ \underbrace{L_S(h_S) - L_S(h^*)}_{\leq 0}$$

$$+ |L_S(h^*) - L_D(h^*)|$$

$$\leq 2 \max_{h \in H} |L_D(h) - L_S(h)| \leq 2\varepsilon$$

Περιορισμένη κλάση $H$

Με $m$ δείγματα

Για κάθε $h \in H$

$$\Pr[\,|L_S(h) - L_D(h)| > \varepsilon\,] \leq e^{-m\varepsilon^2}$$

(Από Hoeffding Inequality)

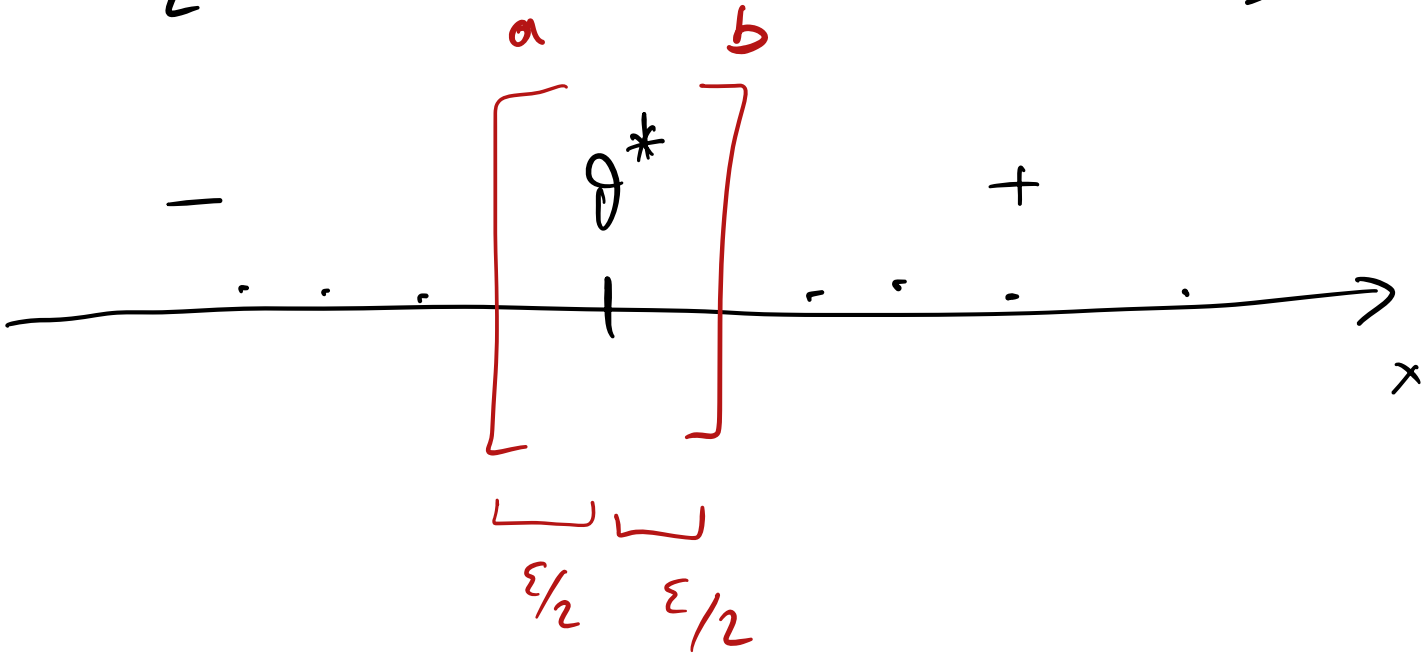
$$\Pr[\exists h \in H : |L_S(h) - L_D(h)| > \varepsilon]$$

$$\leq |H| e^{-m\varepsilon^2} \quad \text{αν} \quad m = \frac{\log(|H|/\delta)}{\varepsilon^2}$$

$$\leq \delta$$

# Threshold Functions

$$H = \{ h_\theta : h_\theta(x) = \text{sign}(x - \theta) \}$$




a t.w. $\Pr[[a, \theta^*]] \leq \frac{\varepsilon}{2}$

b t.w. $\Pr[[\theta^*, b]] \leq \frac{\varepsilon}{2}$

$$\Pr[x_1, \ldots, x_m \notin [a, \theta^*]] \leq (1 - \tfrac{\varepsilon}{2})^m \leq e^{-m\varepsilon/2} \leq \delta/2$$

$$m = \frac{2}{\varepsilon} \log(2/\delta)$$

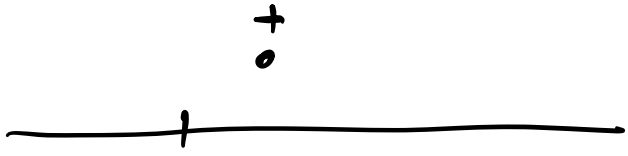
# VC-dimension

**Shattering:** Ένα σύνολο σημείων $X$ είναι shattered από μια κλάση $H$ αν για κάθε labelling $y$ του $X$ υπάρχει $h \in H$ που να δίνει αυτό το labelling δηλ $\forall i \quad h(x_i) = y_i$

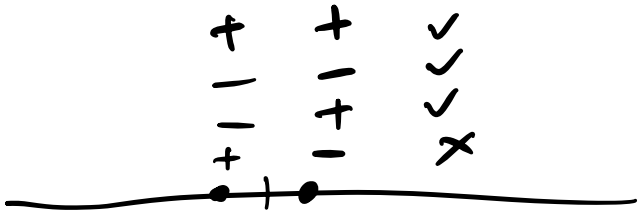
**VC-dimension:** Το μεγαλύτερο μήκος συνόλου σημείων που είναι shattered από την $H$.

**Παραδείγματα:**

Thresholds



2 σημεία



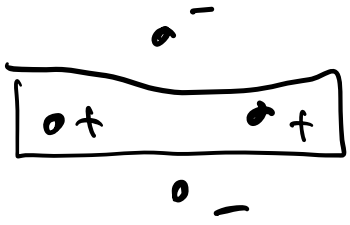

VC-dim = 1

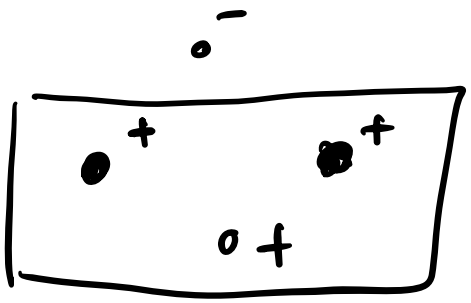
Thresholds με φορά
ή intervals


2 σημεία ✓
3 σημεία ✗

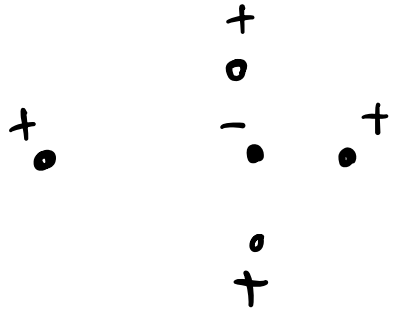

VC-dim = 2

Ορθογώνια (Axis-Aligned)




4 σημεία ✓




5 σημεία ✗
VC-dim = 4

Βασικό Θεώρημα του Statistical Learning

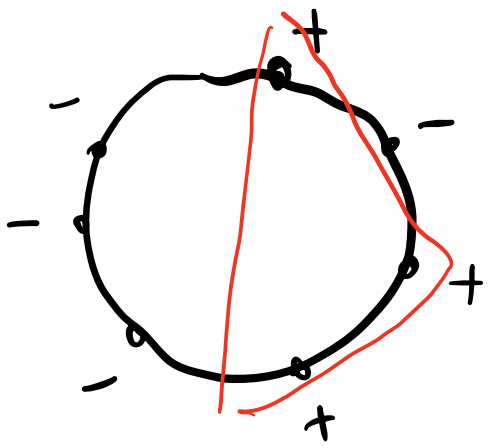
Αν έχουμε μια κλάση H με VC-dim = d

Τότε για binary classification χρειαζόμαστε

$$m = \Theta\left(\frac{d + \log(1/\delta)}{\varepsilon^2}\right)$$

δείγματα για uniform convergence.

H = "Όλα τα κυρτά υποσύνολα του $\mathbb{R}^2$"




VC-dim(H) = ∞

Αν $H$ έχει μέγεθος $k$

τότε

$$VC\text{-}dim(H) \leq \log k$$

Για Realizable Learning

$$m \leq O\left(\frac{d\log(1/\varepsilon) + \log(1/\delta)}{\varepsilon}\right)$$