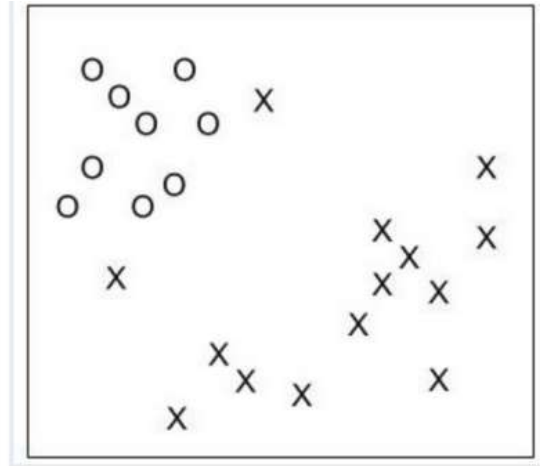


## Άσκηση 6

6



Δίνεται το binary dataset του σχήματος. Διαθέτουμε ένα SVM με τετραγωνικό πυρήνα δηλαδή δευτεροβάθμιο πολυωνυμικό πυρήνα. Η παράμετρος κόστους " $C$ " καθορίζει τη θέση και τη μορφή του διαχωριστικού συνόρου μεταξύ των δύο κλάσεων.

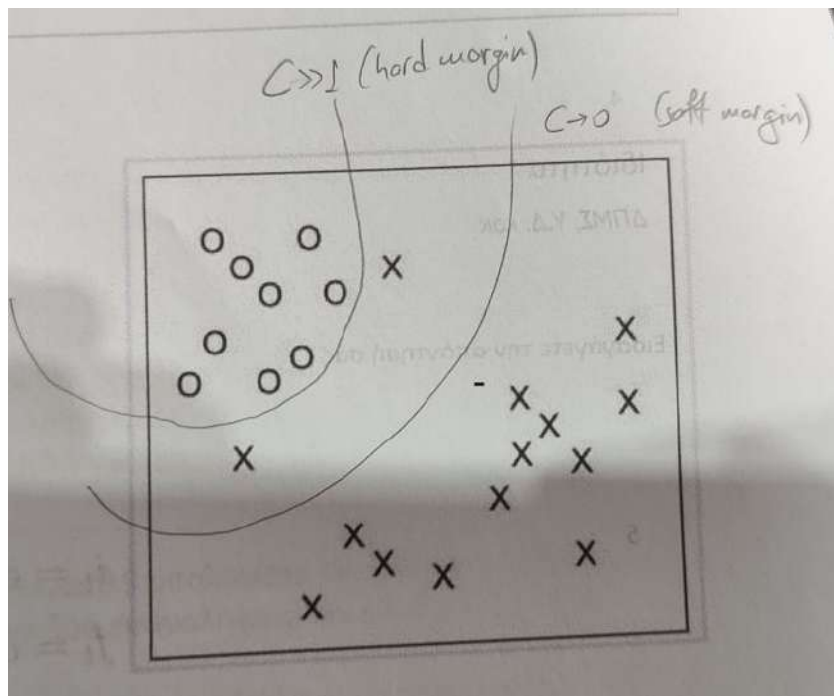
Που θα είναι το διαχωριστικό σύνορο για τιμές του  $C$  που τείνουν στο 0;  
Απαντήστε σε μια πρόταση και σχεδιάστε το πάνω στο σχήμα  
(ξανασχεδιάστε το πρόχειρα). (Μη ανώνυμη ερώτηση ☺)  
(3 βαθμοί)

## Λύση 6

Για  $C \rightarrow 0$  θα υπάρχει λάθος στην ταξινόμηση (misclassification). Η παράμετρος  $C$  στα SVMs εκφράζει το πόσο θέλουμε να αποφύγουμε το misclassification για κάθε σημείο εκπαίδευσης.

Όταν  $C \rightarrow 0$ , τότε το SVM ψάχνει για επίπεδο με μεγάλο margin, ακόμα και αν γίνονται misclassify κάποια δείγματα/σημεία. Αντιθέτως για  $C$  πολύ μεγάλο, έχουμε hard margin. Το μοντέλο με hard margin είναι πιο ευαίσθητο σε overfitting και σε outliers.=

Όταν έχουμε linearly separable data και δε θέλουμε misclassifications, τότε πηγαίνουμε σε SVM με hard margin. Ωστόσο, όταν δεν είναι εφικτό ένα linear boundary ή θέλουμε να επιτρέψουμε κάποια misclassifications αναμένοντας καλύτερη γενίκευση τότε επιλέγουμε SVM με soft margin ( $C \rightarrow 0$ ).



## Άσκηση 7

7

Έστω ότι έχουμε πρόβλημα δυαδικής ταξινόμησης σε χώρο δύο διαστάσεων με τα δείγματα  $(-1,0)$ ,  $(-1,9)$ ,  $(10,0)$ ,  $(10,10)$  και  $(a,b)$  να ανήκουν στην κλάση  $-1$  και τα δείγματα  $(4.5,4)$ ,  $(4.5,5)$ ,  $(5.5,4)$ ,  $(5.5,5)$  και  $(c,d)$  ανήκουν στην κλάση  $+1$ . Να βρείτε τον βέλτιστο PDS πυρήνα που καθιστά τις κλάσεις διαχωρίσιμες και να τεκμηριώσετε την επιλογή σας. Τα  $abcd$  είναι τα 4 τελευταία ψηφία του αριθμού μητρώου σας.  
(7 βαθμοί)

## Λύση 7

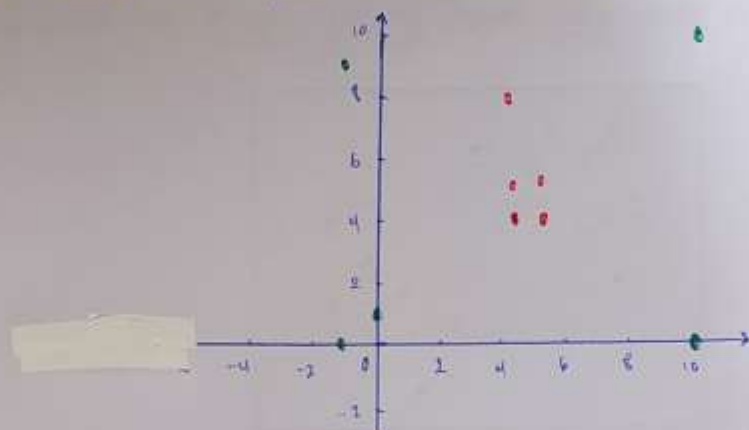
### Άσκηση 7:

Έστω προβλήματα συνδυαστικής ταξινόμησης σε χώρο δύο διαστάσεων με διαγράμματα

κλάσης -1:  $(-1,0), (-1,9), (10,0), (10,10), (a,b) = (0,1)$

κλάσης +1:  $(4,5,4), (4,5,5), (5,5,4), (5,5,5), (c,d) = (4,8)$

Βρείτε βέλτιστο PDS σύμφωνα με τον κανόνα της αλφαριθμητικής διακρίσεως.



### Λύση:

- ο βέλτιστος PDS σύμφωνα με τη μέθοδο να επιλεγεί σε αυτά τα δεδομένα είναι ο πυρήνας Radial Basis Function Kernel ή άλλος ο Γκαουσιανός πυρήνας, με τύπο:

$$K(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right), \text{ 6 ελεύθερη παραμέτρους}$$

Ανάλυστε αυτόν διότι από τις κυριότερες ιδιότητες των δεδομένων της παραπάνω της η κλάση +1 είναι "πυκνότερη" συγκριτικά με, δηλαδή τα σημεία είναι πολύ κοντά τους, ότι ο πυρήνας θα δώσει μεγάλες τιμές για αυτά, ενώ για τις κλάσεις -1 θα δώσει μικρές τιμές, λόγω των μεγάλων αποστάσεων των σημείων συγκριτικά τους.

(Η άσκηση είναι να βρούμε τον Gaussiανό πυρήνα στον  $\mathbb{R}^3$ )

## Άσκηση 8

8

Ποια από τις ακόλουθες προτάσεις δεν είναι αληθής για έναν εκπαιδευμένο αυτοοργανούμενο χάρτη (SOM);  
(3 βαθμοί)

- ☐ Ο χάρτης πραγματοποιεί διανυσματικό κβαντισμό του χώρου εισόδου.
- ☐ Δεν απαντώ
- ☐ Κατά την απεικόνιση της εισόδου στην έξοδο, η μείωση της διάστασης εξασφαλίζει τη διατήρηση της διάταξης.
- ☐ Η απεικόνιση της εισόδου στην έξοδο μπορεί να παρασταθεί με τη βοήθεια ενός διαγράμματος Voronoi.
- ☐ Καμία από τις υπόλοιπες απαντήσεις (είναι όλες αληθείς).
- ☐ Συνήθως, η απεικόνιση της εισόδου στην έξοδο χαρακτηρίζεται από μείωση της διάστασης.

## Άσκηση 9

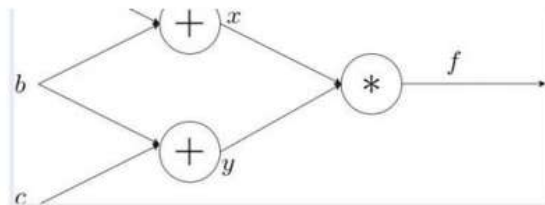
9



10/2/202

ικής Μάθησης 2020 - 2021 5

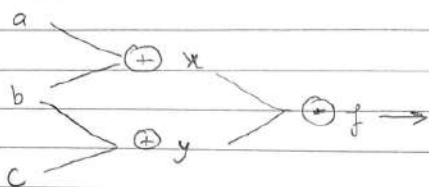
<https://forms.office.com/Pages/ResponsePage.aspx?id=swxeBy>



Δίνεται ο υπολογιστικός γράφος του σχήματος. Αν  $a=3$ ,  $b=-2$ ,  $c=2$  υπολογίστε τις μερικές παραγώγους προς όλες τις μεταβλητές του γράφου χρησιμοποιώντας τον κανόνα της αλυσίδας. (Μη ανώνυμη ερώτηση ①)  
(5 βαθμοί)

## Λύση 9

Λύση 9



$$x = a + b$$

$$y = b + c$$

$$f = x \cdot y$$

$$a = 3, \quad b = -2, \quad c = 2$$

$$\hookrightarrow x = 1, \quad y = 0$$

$$\frac{\partial f}{\partial x} = y = 0$$

$$\frac{\partial f}{\partial y} = x = 1$$

$$\frac{\partial f}{\partial a} = \frac{\partial f}{\partial x} \cdot \frac{\partial x}{\partial a} = y \cdot 1 = 0$$

$$\frac{\partial f}{\partial b} = \frac{\partial f}{\partial x} \cdot \frac{\partial x}{\partial b} + \frac{\partial f}{\partial y} \cdot \frac{\partial y}{\partial b} = y \cdot 1 + x \cdot 1 = 1$$

$$\frac{\partial f}{\partial c} = \frac{\partial f}{\partial y} \cdot \frac{\partial y}{\partial c} = x \cdot 1 = 1$$

## Άσκηση 10

10

Έστω ότι έχουμε ένα σύστημα online μάθησης το οποίο καλείται να επιλέξει από 5 πιθανές δράσεις (με κωδικό 0 ως 4). Να γράψετε την κατανομή  $p_2$  πάνω στις δράσεις καθώς και τα βάρη  $w_{2,i}$  των δράσεων που προκύπτουν μετά την εφαρμογή του πρώτου γύρου του randomized αλγορίθμου σταθμισμένης πλειοψηφίας, αν το διάνυσμα απώλειας  $l_1$  του πρώτου γύρου έχει μη-μηδενική τιμή για τις ενέργειες  $(100 + A) \bmod 5$  και  $(A - 99) \bmod 5$ , όπου  $A$  είναι τα 2 τελευταία ψηφία του αριθμού μητρώου σας. Θεωρείστε επίσης ότι η παράμετρος  $\beta$  είναι ίση με  $0.x$ , όπου  $x = 9 - k$  και  $k$  το τελευταίο ψηφίο του αριθμού μητρώου σας (θέστε  $x = 0.6$  αν το τελευταίο ψηφίο του αριθμού μητρώου σας είναι ίσο με το 9).  
(5 βαθμοί)

## Λύση 10



Online Learning.

5 πιθανές σπέρτες (καθίστοι 0, 1, 2, 3, 4).

randomized algo για τη διαφοροποίηση.

Αδυναμία σπέρτας ή του 1ου τύπου με πιθανότητα 0 ή να επιλεγεί  $(100+A) \bmod 5$ ,  $(A-99) \bmod 5$  (Α στο τελευταίο ψηφίο ΑΗ).

$B=0, x$ ,  $x=9-k \rightarrow$  τελευταίο ψηφίο ΑΗ (0.6 αν  $k=9$ )

Πρέπει να κατανοήσουμε πώς τὰ βάρη  $w_i$  και τὰ  $p_{2,i}$  των σπέρσεων που προκύπτουν μετά την εφαρμογή του 1ου τύπου του αλγορίθμου.

Για έναν ΑΗ, έχουμε  $A=42$ ,  $k=2$ :  $x=9-k=9-2=7$  και  $B=0.7$   
και  $\rightarrow$  ή σπέρτα του 1ου τύπου με πιθανότητα 0 ή να επιλεγεί  $(100+A) \bmod 5 = 142 \bmod 5 = 2$   
 $N=5$  επιλογές/σπέρτες  $\rightarrow (A-99) \bmod 5 = (-57) \bmod 5 = 3$

$w_{2,i} = 1$ ,  $p_{2,i} = \frac{1}{N} = \frac{1}{5}$ ,  $i = 0, 1, 2, 3, 4$ .

1ος τύπος:

$$l_1 = (0, 0, 1, 1, 0)$$

από τα βάρη  $w_{2,0}, w_{2,1}, w_{2,4}$  δεν αλλάζουν ενώ  $w_{2,0} = w_{2,1} = w_{2,4} = w_{1,i} = 1$ .

ενώ  $w_{2,2} = w_{2,3} = 0.7$   $\cdot w_{2,2} = 0.7 \cdot 1 = 0.7$

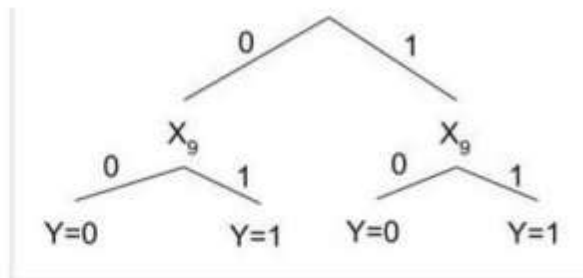
$w_{2,3} = 0.7$   $\cdot w_{2,3} = 0.7 \cdot 1 = 0.7$ .

$$W_2 = \sum_{i=1}^N w_{2,i} = 1 + 1 + 1 + 0.7 + 0.7 = 4.4.$$

$$\text{και } p_{2,i} = \frac{w_{2,i}}{W_2}, \text{ από } p_{2,0} = \frac{w_{2,0}}{W_2} = \frac{1}{4.4} = p_{2,1} \neq p_{2,4}.$$

$$\text{και } p_{2,2} - p_{2,3} = \frac{w_{2,2}}{W_2} = \frac{0.7}{4.4}.$$

Άσκηση 11



Θεωρούμε έναν αλγόριθμο μάθησης βασισμένο σε δέντρα αποφάσεων (ΔΑ) ο οποίος στην είσοδο δέχεται δυαδικά χαρακτηριστικά  $X_i$  και στην έξοδο παράγει μια επίσης δυαδική απόφαση ( $Y \in \{0,1\}$ ). Ο συγκεκριμένος αλγόριθμος παράγει μόνο κανονικά (regular) ΔΑ με βάθος 2. Ένα κανονικό ΔΑ βάθους δύο είναι ένα ΔΑ με βάθος δύο (δηλαδή τέσσερα φύλλα) στο οποίο το αριστερό και το δεξί παιδί της ρίζας υποχρεωτικά ελέγχουν το ίδιο χαρακτηριστικό. Στο σχήμα δίνεται ένα παράδειγμα τέτοιου ΔΑ.

α) Θεωρήστε ότι έχετε δεδομένα για μια έννοια-στόχο  $C$  η οποία μπορεί να περιγραφεί με κανονικά ΔΑ βάθους δύο. Θεωρήστε επιπρόσθετα ότι έχετε 10 χαρακτηριστικά εισόδου (από τα οποία τελικά θα χρειαστούν μόνο δύο για την κατασκευή του ΔΑ). Πόσα δείγματα θα χρειαστούν στην εκπαίδευση έτσι ώστε με πιθανότητα 98% ο αλγόριθμος να βρίσκει ένα δέντρο με ορθότητα το λιγότερο 97%; (3 μονάδες)

β) Ξεκινώντας από την παρατήρηση ότι πολλές υποθέσεις είναι ισοδύναμες, δηλαδή κάνουν την ίδια αντιστοίχιση από τον χώρο των χαρακτηριστικών στον χώρο εξόδου -για παράδειγμα η σειρά των δύο χαρακτηριστικών δεν έχει σημασία-, βρείτε ένα πιο σφιχτό φράγμα που να απαιτεί λιγότερα δεδομένα για τις ίδιες τιμές πιθανότητας και ορθότητας.

(4 μονάδες) (Μη ανώνυμη ερώτηση ①)

(7 βαθμοί)

# Λύση 11

## Άσκηση 11

(α) Αν η έννοια μπορεί να περιγραφεί με δευτερά βαθους δυα, τότε ο χώρος υποθέσεων  $H$  είναι PAC learnable με ελάχιστο αριθμό δειγμάτων

$$m = \frac{\log |H|/\epsilon}{\epsilon}$$

Ομως  $|H| = 2 \cdot \binom{10}{2} = 90$ , γιατί επιλέγουμε 2 από 10 χαρακτηριστικά με 2 δυνατές ερωτίς

Αρα  $m = \frac{\log \frac{90}{0.02}}{0.03} = 281$  δείγματα

(β) Αν η σειρά δεν έχει επήραση,  $|H| = 45$  και ο χώρος δίνει

$$m = 258 \text{ δείγματα}$$

# Άσκηση 12

12

Έστω ότι έχουμε πρόβλημα δυαδικής ταξινόμησης (ετικέτες  $\{-1, +1\}$ ), το οποίο επιθυμούμε να μάθουμε τη χρήση ασθενών μοντέλων μάθησης και της τεχνικής AdaBoost. Έστω επίσης ότι διαθέτουμε ένα σύνολο 5 δειγμάτων  $S = \{(x_0, y_0) \dots (x_4, y_4)\}$  του εν λόγω προβλήματος και ότι μετά την πρώτη επανάληψη του αλγορίθμου το σφάλμα ταξινόμησης είναι ίσο με 0.4, όπου  $x$  το τελευταίο ψηφίο του αριθμού μητρώου σας (θεωρείστε  $x=3$  στην περίπτωση που το τελευταίο ψηφίο του αριθμού μητρώου σας είναι το 0). Έστω επίσης ότι τα δείγματα  $(100 + A) \bmod 5$  και  $(A - 99) \bmod 5$  ταξινομούνται εσφαλμένα από το ασθενές μοντέλο μάθησης  $h_1$ , ενώ τα υπόλοιπα ταξινομούνται ορθά ( $A$  είναι τα 2 τελευταία ψηφία του αριθμού μητρώου σας). Πως θα διαμορφωθεί η κατανομή επιλογής των δειγμάτων ( $D_2$ ) στη δεύτερη επανάληψη του AdaBoost; (5 βαθμοί)



# Λύση 12

Άσκηση 12

$S = \{(x_0, y_0), \dots, (x_4, y_4)\}$  τα δείγματα, με  $y_i = \pm 1$ .

$$D_1(i) = 1/5$$

0.4

$\epsilon_t = 1$  το βεβαίωμα ταξινόμησης μετά την πρώτη επανάληψη. ( $x=4$ )

$$A = 34 \rightarrow (100 + 34) \bmod 5 = 4 \quad (99 + 34) \bmod 5 = 0$$

Άρα τα δείγματα  $(x_0, y_0)$  και  $(x_4, y_4)$  ταξινομούνται λάθος

Ο αλγόριθμος AdaBoost για τα δείγματα επανάληψη δίνει:

$$\alpha_t = \frac{1}{2} \ln \frac{1 - \epsilon_t}{\epsilon_t} = 0.2027$$

$$Z_t = \frac{1}{2\sqrt{\epsilon_t(1-\epsilon_t)}} = 0.9798$$

$$D_{t+1}(i) = D_t \frac{1}{Z_t} e^{-\alpha_t \cdot y_i \cdot h(x_i)}$$

Για  $(x_0, y_0), (x_4, y_4)$ ,  $h(x_i) \cdot y(i) = -1$  (λάθος ταξινόμηση)

Για τα υπόλοιπα  $h(x_i) \cdot y(i) = 1$  (σωστή ταξινόμηση)

Επομένως οι βάρη εκτός για την επόμενη επανάληψη

$$D(0) = 0.2 \cdot \frac{1}{0.9798} e^{-0.2027 \cdot (-1)} = 0.2499 = D(4)$$

$$D(1) = 0.2 \cdot \frac{1}{0.9798} e^{-0.2027 \cdot (1)} = 0.167 = D(2) = D(3)$$

Με βάση τα misclassified δείγματα της εκφώνησης (ασάφεια άσκησης), το error  $\epsilon$  στην 1η επανάληψη θα πρέπει να είναι  $2/5 = 0.4$  ανεξαρτήτως Αριθμού μητρώου εφόσον 2 από τα 5

δείγματα εκπαίδευσης γίνονται misclassified, επομένως πλήρης λύση είναι η παρακάτω:

2020-2021 AdaBoost example

$$S = \{(x_0, y_0), (x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)\} \quad y = \pm 1 \text{ (labels)}$$

$$m=5, \text{ οπότε } D_1(i) = 1/5, \quad i=0, \dots, 4$$

$$\epsilon_t = 0.4, \quad A = 54, \quad (100+A) \bmod 5 = 4, \quad (54-99) \bmod 5 = 0$$

$$\downarrow$$

$$(x_4, y_4)$$

$$\downarrow$$

$$(x_0, y_0)$$

$$\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t} = \frac{1}{2} \ln \frac{3}{2}, \quad \beta_t = 2[\epsilon_t(1-\epsilon_t)]^{1/2} = 2\sqrt{0.6 \cdot 0.4} = 0.4\sqrt{6}$$

$$\text{Γεχώς } y_i h_t(x_i) = \begin{cases} +1, & \text{εάν } i=1,2,3 \\ -1, & \text{εάν } i=0,4 \end{cases}$$

$$\text{Άρα, } \bullet D_2(0) = \frac{1}{0.4\sqrt{6}} \cdot \frac{1}{5} \cdot e^{\ln \sqrt{\frac{3}{2}}} = \frac{1}{2\sqrt{6}} \cdot \sqrt{\frac{3}{2}} = \frac{1}{4}$$

$$\bullet D_2(1) = \frac{1}{0.4\sqrt{6}} \cdot \frac{1}{5} \cdot e^{-\ln \sqrt{\frac{3}{2}}} = \frac{1}{2\sqrt{6}} \cdot \sqrt{\frac{2}{3}} = \frac{1}{6}$$

$$\text{αυτοίμοια, } D_2(2) = D_2(3) = \frac{1}{6} \text{ και } D_2(4) = \frac{1}{4}$$

$$\text{Sanity check: } \bullet D_2(\text{misclassified}) > D_2(\text{correctly classified}) \checkmark$$

$$\bullet \sum_{i=0}^4 D_2(i) = \frac{1}{4} \cdot 2 + \frac{1}{6} \cdot 3 = 1 \checkmark$$

## Ασκηση 13

13

Μια συνάρτηση  $h$  που ορίζεται όπως φαίνεται παρακάτω είναι συμμετρική αν η τιμή της εξαρτάται μόνο από το πλήθος των 1 στην είσοδο. Έστω  $S$  το σύνολο όλων των συμμετρικών συναρτήσεων. Ποια είναι η διάσταση VC του  $S$ ;

(5 βαθμοί)

$$h: \{0, 1\}^n \rightarrow \{0, 1\}$$

# Λυση 13

## 3.26 Symmetric functions

(a) For  $i = 0, \dots, n$ , let  $x_i \in \{0, 1\}^n$  be defined by  $x_i = (\underbrace{1, \dots, 1}_{i \text{ 1's}}, 0, \dots, 0)$ . Then,  $\{x_0, \dots, x_n\}$

can be shattered by  $\mathcal{C}$ . Indeed, let  $y_0, \dots, y_n \in \{0, 1\}$  be an arbitrary labeling of these points. Then, the function  $h$  defined by:

$$h(x) = y_i \quad (\text{E.70})$$

for all  $x$  with  $i$  1's is symmetric and  $h(x_i) = y_i$ . Thus,  $\text{VCdim}(\mathcal{C}) \geq n + 1$ . Conversely, a set of  $n + 2$  points cannot be shattered by  $\mathcal{C}$ , since at least two points would then have the same number of 1's and will not be distinguishable by  $\mathcal{C}$ . Thus,

$$\text{VCdim}(\mathcal{C}) = n + 1. \quad (\text{E.71})$$

# Ασκηση 14

14

Η συνάρτηση ενεργοποίησης  $\text{ReLU}(z) = \max(0, z)$  μπορεί να σβήσει/κορεστεί όταν η είσοδος είναι αρνητική. Σας προτείνεται η χρήση μιας άλλης συνάρτησης ενεργοποίησης  $f(z) = 1.5z$ . Θα λυθεί έτσι το πρόβλημα? Εξηγείστε.

(3 βαθμοί)

# Λύση 14

Το πρόβλημα του κορεσμού θα λυθεί με την  $f(z)=1.5z$ , δηλαδή με τη νέα συνάρτηση ενεργοποίησης θα ενεργοποιούνται και οι αρνητικοί εισόδοι. Όμως, καθώς αυτή η συνάρτηση είναι γραμμική δε μας χρησιμεύει καθώς στα Νευρωνικά Δίκτυα επιθυμούμε μη γραμμικές ιδιότητες στις συναρτήσεις ενεργοποίησης, καθώς η γραμμικότητα εξασφαλίζεται από τα βάρη. Επιπλέον, τα gradients της συνάρτησης αυτής είναι παντού σταθερά (1.5) και με έναν αλγόριθμο όπως το backpropagation το δίκτυο δεν μπορεί να μάθει τίποτα αφού το gradient είναι μόνιμα ίδιο ανεξαρτήτως εισόδου. Αν πάρουμε γραμμικές συναρτήσεις ως συναρτήσεις ενεργοποίησης τότε το αποτέλεσμα θα είναι γραμμική συνάρτηση => άρα δεν μπορούμε να προσεγγίσουμε άλλες, μη γραμμικές συναρτήσεις.

# Ασκηση 15

15

$$h_r(x) = 1 \text{ if } x \leq r \text{ και } h_r(x) = 0 \text{ αλλιώς (1)}$$

Έχουμε μονοδιάστατα δεδομένα. Για έναν πραγματικό  $r$ , ορίζουμε τη συνάρτηση (1). Έστω  $H = \{h_r\}$ . Θεωρούμε ένα σύνολο  $S$  από  $m$  ξεχωριστά δείγματα πάνω στην ευθεία. Ποια είναι η εμπειρική πολυπλοκότητα Rademacher  $R_m(H)$  της  $H$  στο  $S$ ; (Μη ανώνυμη ερώτηση ①)  
(5 βαθμοί)



# Λύση 15

Θέματα 9020-9021 :

Ερώτηση 15: Για  $r \in \mathbb{R}$  θεωρούμε τη συνάρτηση

$$h_r(x) = \begin{cases} 1, & x \leq r \\ 0, & x > r. \end{cases}$$

Έστω  $H = \{h_r : r \in \mathbb{R}\}$  και  $S = \{x_1, \dots, x_m\}$  τυχαία δείγματα. Ζητάμε την Rademacher υπολογιστικότητα  $\hat{R}_S(H)$  ως  $H$  στο  $S$ .

Λύση: Εξ' ορισμού έχουμε ότι  $\hat{R}_S(H) = E_{\sigma} \left[ \sup_{r \in \mathbb{R}} \frac{1}{m} \sum_{i=1}^m \sigma_i h_r(x_i) \right]$ .

όπου οι ανεξαρτησίες  $\sigma_i$  με  $P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2 \quad \forall i = 1, \dots, m$ .

• Δύο χρήσιμες ιδιότητες: (\* αποδείχτηκατος).

- (α)  $\sup(A+B) = \sup(A) + \sup(B)$
- (β)  $\sup(-A) = -\inf(A)$

Πρώτα, επιπλέον για ένα  $x_i \in S$  μπορούμε να βρούμε  $r_1 < x_i < r_2$  που έχουμε

ότι  $\sup_{r \in \mathbb{R}} h_r(x_i) \geq h_{r_1}(x_i) = 1$  και  $\inf_{r \in \mathbb{R}} h_r(x_i) \leq h_{r_2}(x_i) = 0$  και επειδή  $h_r(x_i) = \begin{cases} 1 \\ 0 \end{cases}$

θα έχουμε ότι  $\sup_{r \in \mathbb{R}} h_r(x_i) = 1$  και  $\inf_{r \in \mathbb{R}} h_r(x_i) = 0$ .

Τώρα χρησιμοποιώντας την (α) θα έχουμε ότι

$$\hat{R}_S(H) = E_{\sigma} \left[ \sup_{r \in \mathbb{R}} \frac{1}{m} \sum_{i=1}^m \sigma_i h_r(x_i) \right] = \frac{1}{m} E_{\sigma} \left[ \sup_{r \in \mathbb{R}} \sigma_1 h_r(x_1) \right] + \dots + \frac{1}{m} E_{\sigma} \left[ \sup_{r \in \mathbb{R}} \sigma_m h_r(x_m) \right]$$

όπως  $\forall i = 1, \dots, m$  έχουμε

$$\begin{aligned} E_{\sigma} \left[ \sup_{r \in \mathbb{R}} \sigma_i h_r(x_i) \right] &= P(\sigma_i = -1) \cdot \sup_{r \in \mathbb{R}} -h_r(x_i) + P(\sigma_i = 1) \cdot \sup_{r \in \mathbb{R}} h_r(x_i) \\ &\stackrel{(β)}{=} -\frac{1}{2} \inf_{r \in \mathbb{R}} h_r(x_i) + \frac{1}{2} \sup_{r \in \mathbb{R}} h_r(x_i) \\ &= \frac{1}{2} \sup_{r \in \mathbb{R}} h_r(x_i) = \frac{1}{2}. \end{aligned}$$

Αρα τελικά  $\hat{R}_S(H) = \frac{1}{m} E_{\sigma} \left[ \sup_{r \in \mathbb{R}} \sigma_1 h_r(x_1) \right] + \dots + \frac{1}{m} E_{\sigma} \left[ \sup_{r \in \mathbb{R}} \sigma_m h_r(x_m) \right]$

$$= \underbrace{\frac{1}{m} \cdot \frac{1}{2} + \dots + \frac{1}{m} \cdot \frac{1}{2}}_m = \frac{m}{2} \cdot \frac{1}{m} = \frac{1}{2}.$$

0

Απόδειξη της (\*):

(α)  $\sup(A+B) = \sup(A) + \sup(B)$ .

• Έστω  $\alpha \in A$  και  $\beta \in B$  τότε  $\alpha \leq \sup(A), \beta \leq \sup(B) \Rightarrow \alpha + \beta \leq \sup(A) + \sup(B)$ .

Επομένως, επειδή  $\alpha + \beta \in A+B$  θα έχουμε ότι το  $\sup(A) + \sup(B)$  θα είναι ένα άνω φράγμα για το  $A+B$ . Αρα,  $\sup(A+B) \leq \sup(A) + \sup(B)$ .

• Αντίστοιχα, αν  $\varepsilon > 0$  από χαρακτηριστικό supremum θα υπάρξουν  $\alpha \in A, \beta \in B$  με

$$\sup(A) - \frac{\varepsilon}{2} < \alpha \leq \sup(A) \quad (1)$$

και

$$\sup(B) - \frac{\varepsilon}{2} < \beta \leq \sup(B) \quad (2)$$

① + ②

$$\Rightarrow \sup(A) + \sup(B) - \varepsilon < \alpha + \beta.$$

Όμως,  $\alpha + \beta \in A+B$  άρα  $\alpha + \beta \leq \sup(A+B)$ . Συνεπώς,

$$\sup(A) + \sup(B) - \varepsilon < \alpha + \beta \leq \sup(A+B).$$

Άρα δείξαμε ότι για κάθε  $\varepsilon > 0$  ισχύει  $\sup(A) + \sup(B) - \varepsilon \leq \sup(A+B)$ .

Αντίστοιχα έπεται ότι  $\sup(A) + \sup(B) \leq \sup(A+B)$ .

$$\text{Άρα, } \boxed{\sup(A+B) = \sup(A) + \sup(B)}.$$

(β) - Για κάθε  $\alpha \in A$  έχουμε  $\alpha \geq \inf(A) \Rightarrow -\alpha \leq -\inf(A)$ . Άρα το  $-\inf(A)$  άνω

$$\text{φράγμα του } -A \text{ ανήκει έπειτα ότι } \boxed{\sup(-A) \leq -\inf(A)} \quad (1).$$

- Για κάθε  $\alpha \in A$  έχουμε  $-\alpha \leq \sup(-A) \Rightarrow \alpha \geq -\sup(-A)$ . Άρα το  $-\sup(-A)$

$$\text{κάτω φράγμα του } A \text{ ανήκει έπειτα ότι } \inf(A) \geq -\sup(-A) \Rightarrow \boxed{\inf(A) \leq \sup(-A)} \quad (2)$$

$$\text{Αν } (1), (2) \text{ έχουμε } \boxed{\sup(-A) = -\inf(A)}.$$

## Ασκηση 16

16

Εξηγήστε εν συντομία σε τί διαφέρουν το bipartite ranking από το k-partite ranking και δώστε ένα παράδειγμα χρήσης (ένα υποθετικό task) για το καθένα.

(3 βαθμοί)

## Λύση 16

Στο bi-partite ranking ο αλγόριθμος δέχεται δεδομένα εκπαίδευσης με binary πχ για αρνητικά και θετικά. Αυτός κατασκευάζει μια ranking function η οποία θα διατάσσει σωστά τα νέα δεδομένα με βάση τα labels.

Στο k-partite ranking έχουμε k κατηγορίες, άρα k διατεταγμένους αριθμούς ως ετικέτες.

Πχ στο Spotify, ο αλγόριθμος πρότασης για ένα τραγούδι:

Bi-partite ranking: Not relevant (or recommended), Relevant

K-partite ranking: Βαθμός recommendation: 1,2,3,4,5

## Άσκηση 17

17

Ας υποθέσουμε ότι εκπαιδεύετε ένα νευρωνικό δίκτυο για ταξινόμηση, αλλά παρατηρείτε ότι το training error είναι πολύ χαμηλότερο από το validation error. Ποιο από τα ακόλουθα θα χρησιμοποιήσετε για την αντιμετώπιση του ζητήματος (επιλέξτε όλα όσα ισχύουν); (3 βαθμοί)

- ☐ Μειώστε την dropout probability
- ☒ Χρησιμοποιήστε ένα δίκτυο με λιγότερα επίπεδα
- ☒ Αυξήστε το βάρος της L2 κανονικοποίησης (regularization)
- ☐ Αυξήστε το μέγεθος κάθε κρυμμένου επιπέδου

## Άσκηση 18

18

Στη γραμμική παλινδρόμηση, έχουμε  $XX^T W = XY$ . Τί πρέπει να ισχύει για να είναι αντιστρέψιμος ο πίνακας  $XX^T$ ; (3 βαθμοί)

## Λύση 18

When  $X^T X$  is invertible, eq (3) directly implies  $w^* = (X^T X)^{-1} X^T y$  is the **unique** solution of linear regression. This often happens when we face an over-determined system—number of samples is much larger than number of variables ( $N \gg D$ ). An intuitive way to see this: When  $N \gg D$ , we have many training samples to fit but don't have enough degree of freedom (number of variables), so it's unlikely to fit data very well and the minimizer can be uniquely determined.

## Άσκηση 19

19

Έχουμε μια συλλογή εικόνων 32x32 pixels σε χρωματική αναπαράσταση RGB. Για την ταξινόμησή τους σε 5 κατηγορίες σας δίνεται συνελκτικό δίκτυο με τα εξής επίπεδα: CONV (8 φίλτρα, F=6, S=2, P=1), POOL (F=3, S=2) FC (32), Softmax(5). Υπολογίστε τον αριθμό των παραμέτρων του δικτύου. (5 βαθμοί)

# Λύση 19

## Λύση

### 1. Convolution Layers

$$L_{in} = 32$$

$$L_{out} = \frac{L_{in} - F + 2P}{S} + 1 = \frac{32 + 2 \cdot 1 - 6}{2} + 1 = 15$$

$$\text{After } 32 \times 32 \times 8 \rightarrow 15 \times 15 \times 8.$$

$$\text{Parameters: } 6 \times 6 \times 3 = 108 + 1 (\text{bias}) = 109 \text{ per filter}$$

$$\Rightarrow \text{Params} = 8 \times 109 = 872.$$

### 2 Pooling

$$L_{out} = \frac{L_{in} - F}{S} + 1 = 7$$

$$\text{After } 15 \times 15 \times 8 \rightarrow 7 \times 7 \times 8.$$

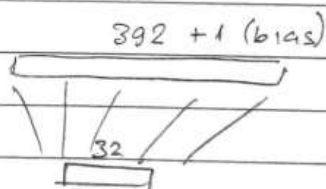
$$\text{Params: } 0!$$

### 3 Fully Connected.

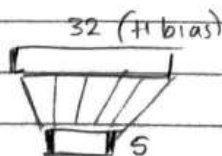
$$\text{To } 7 \times 7 \times 8 \rightarrow \text{Flatten} = 392$$

$$\Rightarrow \text{Fully conn to 32}$$

$$\text{Params } (392 + 1) \cdot 32 = 12576 \text{ param}$$



### 4. Softmax



$$(32 + 1) \cdot 5 = 165 \text{ param}$$

$$\text{Summed: } 872 + 12576 + 165 = \underline{\underline{13613 \text{ params}}}$$



# Άσκηση 20

20

Έστω ότι για ένα πρόβλημα δυαδικής ταξινόμησης έχουμε 100 δείγματα δεδομένων με τις αντίστοιχες ετικέτες τους. Το προσεγγίζουμε με ένα πολυωνμικό μοντέλο 20ου βαθμού και παρατηρούμε ότι εμφανίζει υπερπροσαρμογή. Για το λόγο αυτό αποφασίζουμε να εφαρμόσουμε L2 κανονικοποίηση με  $\lambda=0.xy$ , όπου  $x, y$  το προτελευταίο και το τελευταίο ψηφίο του αριθμού μητρώου σας αντίστοιχα (θέστε  $\lambda=0.27$  αν  $xy=00$ ). Πως θα μεταβάλλονται οι οροι του πολυωνύμου σε κάθε επανάληψη, αν χρησιμοποιήσουμε ως μέθοδο εκμάθησης τη στοχαστική κατάβαση κλήσης με ρυθμό μάθησης  $\eta=0.y$  (θέστε  $\eta=0.7$  αν  $y=0$ ); (3 βαθμοί)

## Λύση 20

Έστω  $\hat{y} = \vec{w}^T \vec{x} + b$ . Για κάθε συνιστώσα  $\vec{w}$  το update γίνεται ως:

$$w_i^{k+1} = w_i^k - \eta \left[ \frac{\partial L(\hat{y}^k, y)}{\partial w_i^k} + \lambda \frac{\partial R(w_i^k)}{\partial w_i^k} \right]$$

$k$ : iteration step  
 $\lambda = \text{const.}$  ή  $\lambda = \frac{c}{N}$  (sample size)

L1 Regularization:  $R(\vec{w}) = \|\vec{w}\|_1 = \sum_i |w_i|$ ,  
οπότε  $\frac{\partial R(w_i)}{\partial w_i} = \text{sign}(w_i)$

L2 Regularization:  $R(\vec{w}) = \frac{1}{2} \|\vec{w}\|_2^2$ , οπότε  $\frac{\partial R(w_i)}{\partial w_i} = w_i$

Συνολικά περιπτώσεις του δείκτη:

$$w_i^{k+1} = w_i^k - \eta \left[ \frac{\partial L}{\partial w_i^k} + \lambda w_i \right]$$

(Βοηθάει π.χ. σε ισοκύβητες μεταξύ άλλων loss functions που by default αφορούν #δείγματα, π.χ. Cross Entropy)

# Άσκηση 21

21

Ένα αυτόνομο ρομπότ κινείται στο χώρο που απεικονίζεται στο σχήμα (α). Οι καταστάσεις αντιστοιχούν στα κελιά και αναφερόμαστε σε αυτά με τη σύμβαση (γραμμή, κολόνα). Το ρομπότ ξεκινάει πάντα από την κατάσταση "S". Υπάρχουν δύο τερματικές καταστάσεις-στόχοι στις θέσεις (1,3) και (2,3) στις οποίες αναγράφεται η ανταμοιβή του πράκτορα. Στις μη-τερματικές θέσεις η ανταμοιβή είναι μηδέν. Ο πράκτορας μπορεί να κινηθεί σε τέσσερις κατευθύνσεις (πάνω, κάτω, αριστερά, δεξιά). Ωστόσο, η κίνησή του είναι ελαφρά στοχαστική: μετακινείται με πιθανότητα 0.8 στην κατεύθυνση που θέλει, αλλά και με πιθανότητα 0.1 για την καθεμία, προς τις δύο κατευθύνσεις που είναι κάθετες στην ηθελημένη (σχήμα β). Αν το ρομπότ συγκρουστεί με τοίχο (πάει να βγει εκτός του grid), μένει στην ίδια θέση.

- 1) αν ο παράγοντας έκπτωσης είναι 0.9 υπολογίστε τις τιμές της αξίας για τα σημεία (2,1), (1,2) και (2,2) για δύο επαναλήψεις του αλγόριθμου value iteration (4 μονάδες)
- 2) ο πράκτορας ξεκινάει με την πολιτική που διαλέγει να πηγαίνει πάντοτε δεξιά. Εκτελεί τρία πειράματα και οι μετακινήσεις του καταγράφονται ως εξής: α) (1,1)-(1,2)-(1,3), β) (1,1)-(1,2)-(2,2)-(2,3), και γ) (1,1)-(2,1)-(2,2)-(2,3). Ποια είναι η κατά monte carlo εκτίμηση της αξίας των καταστάσεων (1,1) και (2,2) με βάση αυτά τα δειγματικά μονοπάτια; (3 μονάδες)

(Μη ανώνυμη ερώτηση ①)  
(7 βαθμοί)

	1	2	3
2			+10
1	S		-10


Σχήμα α

0.8

0.1   ↑   0.1

Σχήμα β

## Λύση 21

To πρόβλημα με το αυτοκίνητο  robot

	1	2	3
2			10
1			-10

intended dir: 0.8

⊥ dir: 0.1

backwards: 0.0

$\gamma = 0.9$

$\alpha) \underline{k=1}$ :

$$\begin{pmatrix} 0 & 0 & 10 \\ 0 & 0 & -10 \end{pmatrix} \text{ (initialization)}$$

$k=2$ : Τα  $(1,1)$ ,  $(2,1)$  δεν γίνονται update γιατί στις περιπτώσεις αυτές έχω μηδενικά.

$(1,2)$ : Βέλτιστη δράση = απιστρέφα, γιατί αν πάει νότια τότε έχω 20% πιθανότητα να καταλήξει στο -10

$(2,2)$ : Βέλτιστη δράση = δεξιά, με  $V = 0.9(0.8 \times 10 + 0.1 \times 0 + 0.1 \times 0) = 7.2$

$$\begin{pmatrix} 0 & 7.2 & 10 \\ 0 & 0 & -10 \end{pmatrix}$$

$k=3$ : (Τελικά η εκπαίδευση τελειώνει μέχρι  $k=2$ )

$(1,1)$ : No updates

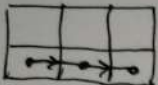
$(1,2)$ : Βέλτιστη δράση = πάνω:

$$V = 0.9(7.2 \times 0.8 + (-10) \times 0.1 + 0 \times 0.1) = 4.284$$

$(2,1)$ : Βέλτιστη δράση = δεξιά:  $V = 0.9(7.2 \times 0.8 + 0 \times 0.1 + 0 \times 0.1) = 5.184$

$(2,2)$ : Βέλτιστη δράση = δεξιά:  $V = 0.9(10 \times 0.8 + 0 \times 0.1 + \overset{\text{ταίριας}}{7.2} \times 0.1) = 7.848$

β) Αρχικά, ισχύει  $2 \begin{array}{c|cc} & 1 & 2 & 3 \\ \hline 1 & 0 & 0 & +10 \\ & 0 & 0 & -10 \end{array}$ . Αντ.  $V_1(i,j) = 0, i,j = 1,2$

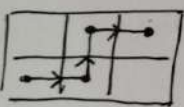
α):   $U_1 = -10$ . Το (2,1) και (2,2) δεν γίνονται update.

Για τα (1,1), (1,2) ισχύει:

$$V_2 = V_1 + \alpha(U_1 - V_1) \Rightarrow V_2 = 0 + \alpha \cdot (-10 - 0) = -10\alpha,$$

όπου  $\alpha$ : κάποιο learning rate. Αρα μετά το (α) έχουμε

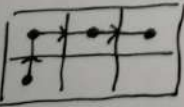
0	0	+10
-10α	-10α	-10

β):   $U_2 = +10$ . Το (2,1) δεν γίνεται update.

$$(2,2): V_3 = V_2 + \alpha(U_2 - V_2) \Rightarrow V_3 = 10\alpha$$

$$(1,1), (1,2): V_3 = -10\alpha + \alpha(10 + 10\alpha) \Rightarrow V_3 = 10\alpha^2$$

0	10α	+10
10α <sup>2</sup>	10α <sup>2</sup>	-10

γ):   $U_3 = +10$ . Το (1,2) δεν γίνεται update.

$$(2,1): V_4 = 10\alpha \quad (2,2): V_4 = 10\alpha + \alpha(10 - 10\alpha) \Rightarrow V_4 = 20\alpha - 10\alpha^2$$

$$(1,1): V_4 = 10\alpha^2 + \alpha(10 - 10\alpha^2) \Rightarrow V_4 = 10\alpha + 10\alpha^2 - 10\alpha^3$$