Παν.Θεσσαλίας,ΤΜΗΥΤΔ ΗΥ345: ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΤΩΝ Ακ.Έτος 2012-13
20-06-2013 **Τελική Εξέταση** Ομάδα **A**

Ονομα/νυμο: **Υπογραφή:**

ΑΜ: **Εξάμηνο:** **Αριθμός διφύλλων:**

ΠΑΡΑΤΗΡΗΣΕΙΣ: *Ανοιχτό βιβλίο μαθήματος & σημειώσεις μαθήματος. Κλειστά κινητά.*

Θέμα 1: (30%) Τα (a), (b), και (c) είναι ανεξάρτητα ερωτήματα. Απαντήστε αναλυτικά.

(a) Έστω πολυδιάστατα διανύσματα χαρακτηριστικών $\mathbf{x} = [\,x_1\,,x_2\,,...,x_L\,]^{\mathrm{T}}$ με στατιστικά ανεξάρτητα στοιχεία (statistically independent components) που παίρνουν δυαδικές τιμές (δηλ. $x_l = 0$ ή 1, για $l = 1\,,2\,,...,L\,$). Τα διανύσματα αυτά χρησιμοποιούνται για την ταξινόμηση σε K κλάσεις, $\omega_1\,,\omega_2\,,...,\omega_K$, με εκ των προτέρων πιθανότητες (prior probabilities) $P(\omega_k)\,,\ k = 1\,,2\,,...,K$. Έστω επίσης ότι οι υπό συνθήκη (class conditional) πιθανότητες δίνονται από τις

$$P_{l\,k} = P\,[\,x_l = 1\,|\,\omega_k\,]\ \ , \quad \text{για} \quad l = 1\,,2\,,...,L \quad \text{και} \quad k = 1\,,2\,,...,K\ \ .$$

Βρείτε τη μορφή της συνάρτησης διάκρισης (κανόνα ταξινόμησης) $g_k(\mathbf{x})$ σύμφωνα με την θεωρία αποφάσεων κατά Bayes.

Βρείτε επίσης πώς αυτή απλοποιείται σε μία απλή συνάρτηση του αθροίσματος $\sum_{l=1}^{L} x_l\,$, γράφοντας την εξίσωση που περιγράφει το υπερεπίπεδο απόφασης, για την ειδική περίπτωση δύο ισοπίθανων κλάσεων, δηλ. $K = 2$ και $P(\omega_1) = P(\omega_2) = 1/2\,$, και με ίσες υπό συνθήκη πιθανότητες για όλα τα στοιχεία του διανύσματος χαρακτηριστικών, δηλ. $P_{l\,1} = p > 1/2\,$, για $l = 1\,,2\,,...,L\,$.

(b) Βρείτε τον εκτιμητή μέγιστης πιθανοφάνειας (maximum likelihood estimator) της παραμέτρου $\theta > 0$ της εκθετικής συνάρτησης πυκνότητας πιθανότητας,

$$p\,(\,x\,|\,\theta\,) \ = \ \begin{cases} \theta\, e^{-\,\theta\, x}\ , & x \ \geq \ 0 \\ 0\ , & \text{αλλού} \end{cases}\ \ ,$$

με βάση N στατιστικά ανεξάρτητα δείγματά της $x_n\,,\ n = 1\,,2\,,...,N\,$.

(c) Έστω s μία μετρική ομοιότητας (similarity metric) ορισμένη στο σύνολο $X\,$, για την οποία ισχύει $s\,(\mathbf{x}\,,\mathbf{y}) > 0\,,\ \forall\,\mathbf{x}\,,\mathbf{y}\ \in\ X\,$. Έστω επίσης μία συνεχής μονοτονικά φθίνουσα συνάρτηση $f:\ \mathcal{R}^{+}\ \rightarrow\ \mathcal{R}^{+}\,$, για την οποία ισχύει ότι

$$f(a) + f(b)\ \geq\ f\left(\frac{1}{\dfrac{1}{a}+\dfrac{1}{b}}\right)\ \ , \quad \forall\, a\,,\ b\ \in\ \mathcal{R}^{+}\ \ ,$$

Δείξτε ότι η $d = f(s)$ είναι μετρική ανομοιότητας (dissimilarity metric) στο σύνολο $X\,$.

Θέμα 2: (45%) Έστω δύο κλάσεις ω_1 και ω_2 με τα ακόλουθα 2-D διανύσματα χαρακτηριστικών εκπαίδευσης (γραμμένα σε μορφή $\mathbf{x} = [x_1, x_2]^{\mathrm{T}}$):

$$\omega_1 : \ [1,0]^{\mathrm{T}}, \ [-1,0]^{\mathrm{T}}, \ [3,1]^{\mathrm{T}}, \ [-3,1]^{\mathrm{T}}, \ [3,-1]^{\mathrm{T}}, \ [-3,-1]^{\mathrm{T}} ,$$

$$\omega_2 : \ [1,2]^{\mathrm{T}}, \ [1,-2]^{\mathrm{T}}, \ [-1,2]^{\mathrm{T}}, \ [-1,-2]^{\mathrm{T}} .$$

Έστω επίσης και ένα διάνυσμα δοκιμής $\mathbf{x}_{\text{test}} = [3,2]^{\mathrm{T}}$.

(a) Σχεδιάστε τα παραπάνω διανύσματα στο 2-D χώρο με άξονες τα x_1, x_2. Είναι οι κλάσεις γραμμικά διαχωρίσιμες (με βάση τα διανύσματα εκπαίδευσης);

(b) Ταξινομήστε το διάνυσμα δοκιμής με βάση τον κανόνα των τριών πλησιέστερων γειτόνων (3-NN), χρησιμοποιώντας την Ευκλείδεια απόσταση (L_2) μεταξύ διανυσμάτων.

(c) Κατασκευάστε (και σχεδιάστε) ένα perceptron δύο επιπέδων που να ταξινομεί τα διανύσματα εκπαίδευσης σωστά στις δύο κλάσεις. Σε ποια κλάση ταξινομείται το διάνυσμα δοκιμής $\mathbf{x}_{\text{test}}$;

(d) Λύστε το πρόβλημα χρησιμοποιώντας πολυωνυμικό ταξινομητή και ταξινομήστε το διάνυσμα δοκιμής με αυτόν. Επίσης, σχεδιάστε τις περιοχές απόφασης για κάθε κλάση στον αρχικό χώρο.

(e) Τέλος, μειώστε την διάσταση των δεδομένων χρησιμοποιώντας ανάλυση κύριων συνιστωσών (Principal Component Analysis – PCA). Είναι οι κλάσεις διαχωρίσιμες στο μονοδιάστατο χώρο που προκύπτει;

Θέμα 3: (25%) Έστω τα πρώτα 4 διανύσματα της κλάσης ω_1 του Θέματος 2, δηλ.

$$\mathbf{x}_1 = [1,0]^{\mathrm{T}}, \ \mathbf{x}_2 = [-1,0]^{\mathrm{T}}, \ \mathbf{x}_3 = [3,1]^{\mathrm{T}}, \ \mathbf{x}_4 = [-3,1]^{\mathrm{T}} .$$

(a) Ομαδοποιήστε τα διανύσματα ακολουθιακά, χρησιμοποιώντας το βασικό ακολουθιακό αλγοριθμικό σχήμα (BSAS) με βάση την απόσταση L_1 (Manhattan distance) και κατώφλι $\theta = 3.5$. Θεωρήστε δύο περιπτώσεις όσον αφορά την ακολουθία που εμφανίζονται τα διανύσματα στον αλγόριθμο: Στην πρώτη περίπτωση με τη σειρά της εκφώνησης (αριστερά προς δεξιά), και στην δεύτερη περίπτωση με την αντίθετη σειρά. Και στις δύο περιπτώσεις, θεωρήστε ότι οι αντιπρόσωποι των ομάδων που προκύπτουν είναι η μέση τιμή των μελών της ομάδας.

(b) Ομαδοποιήστε τα διανύσματα ιεραρχικά, χρησιμοποιώντας τον συσσωρευτικό αλγόριθμο ομαδοποίησης (agglomerative clustering) με βάση την απόσταση L_1 (Manhattan distance). Χρησιμοποιείστε δύο διαφοροποιήσεις του αλγορίθμου, όσον αφορά την ενημέρωση των αποστάσεων του πίνακα ανομοιότητας, τον πρώτο με βάση τον αλγόριθμο απλού δεσμού (single link), και τον δεύτερο με βάση τον αλγόριθμο πλήρους δεσμού (complete link). Και στις δύο περιπτώσεις σχεδιάστε τα δεντρογράμματα ανομοιότητας (dendrograms) που προκύπτουν.

Ονομα/νυμο: **Υπογραφή:**

ΑΜ: **Εξάμηνο:** **Αριθμός διφύλλων:**

ΠΑΡΑΤΗΡΗΣΕΙΣ: *Ανοιχτό βιβλίο μαθήματος & σημειώσεις μαθήματος. Κλειστά κινητά.*

Θέμα 1: (30%) Τα (a), (b), και (c) είναι ανεξάρτητα ερωτήματα. Απαντήστε αναλυτικά.

(a) Έστω πολυδιάστατα διανύσματα χαρακτηριστικών $\mathbf{x} = [\,x_1\,,x_2\,,...,x_L\,]^{\mathrm{T}}$ με στατιστικά ανεξάρτητα στοιχεία (statistically independent components) που παίρνουν δυαδικές τιμές (δηλ. $x_l = 0$ ή 1, για $l = 1\,,2\,,...,L\,$). Τα διανύσματα αυτά χρησιμοποιούνται για την ταξινόμηση σε K κλάσεις, $\omega_1\,,\omega_2\,,...,\omega_K$, με εκ των προτέρων πιθανότητες (prior probabilities) $P(\omega_k)\,,\ k = 1\,,2\,,...,K$. Έστω επίσης ότι οι υπό συνθήκη (class conditional) πιθανότητες δίνονται από τις

$$P_{l\,k} = P\,[\,x_l = 1\,|\,\omega_k\,]\ \ , \quad \text{για} \quad l = 1\,,2\,,...,L \quad \text{και} \quad k = 1\,,2\,,...,K\ \ .$$

Βρείτε τη μορφή της συνάρτησης διάκρισης (κανόνα ταξινόμησης) $g_k(\mathbf{x})$ σύμφωνα με την θεωρία αποφάσεων κατά Bayes.

Βρείτε επίσης πώς αυτή απλοποιείται σε μία απλή συνάρτηση του αθροίσματος $\sum_{l=1}^{L} x_l\,$, γράφοντας την εξίσωση που περιγράφει το υπερεπίπεδο απόφασης, για την ειδική περίπτωση δύο ισοπίθανων κλάσεων, δηλ. $K = 2$ και $P(\omega_1) = P(\omega_2) = 1/2\,$, και με ίσες υπό συνθήκη πιθανότητες για όλα τα στοιχεία του διανύσματος χαρακτηριστικών, δηλ. $P_{l\,1} = p > 1/2\,$, για $l = 1\,,2\,,...,L\,$.

(b) Βρείτε τον εκτιμητή μέγιστης πιθανοφάνειας (maximum likelihood estimator) της παραμέτρου $\theta > 0$ της συνάρτησης πυκνότητας πιθανότητας Erlang,

$$p\,(\,x\,|\,\theta\,) \ = \ \begin{cases} \theta^{\,2}\,x\,e^{\,-\,\theta\,x}\ , & x \ \geq \ 0 \\ 0\ , & \text{αλλού} \end{cases}\ \ ,$$

με βάση N στατιστικά ανεξάρτητα δείγματά της $x_n\,,\ n = 1\,,2\,,...,N\,$.

(c) Έστω s μία μετρική ομοιότητας (similarity metric) ορισμένη στο σύνολο $X\,$, για την οποία ισχύει $s\,(\mathbf{x}\,,\mathbf{y}) > 0\,,\ \forall\,\mathbf{x}\,,\mathbf{y} \ \in \ X\,$. Έστω επίσης μία συνεχής μονοτονικά φθίνουσα συνάρτηση $f\,:\ \mathcal{R}^{+} \ \rightarrow \ \mathcal{R}^{+}\,$, για την οποία ισχύει ότι

$$f(a) + f(b) \ \geq \ f\left(\frac{1}{\dfrac{1}{a} + \dfrac{1}{b}}\right) \ \ , \quad \forall\,a\,,\ b \ \in \ \mathcal{R}^{+}\ \ ,$$

Δείξτε ότι η $d = f(s)$ είναι μετρική ανομοιότητας (dissimilarity metric) στο σύνολο $X\,$.

[Ομάδα **B**]

Θέμα 2: (45%) Έστω δύο κλάσεις ω_1 και ω_2 με τα ακόλουθα 2-D διανύσματα χαρακτηριστικών εκπαίδευσης (γραμμένα σε μορφή $\mathbf{x} = [x_1, x_2]^{\mathrm{T}}$):

$$\omega_1 : \ [1,2]^{\mathrm{T}}, \ [1,-2]^{\mathrm{T}}, \ [-1,2]^{\mathrm{T}}, \ [-1,-2]^{\mathrm{T}},$$

$$\omega_2 : \ [3,-1]^{\mathrm{T}}, \ [-3,-1]^{\mathrm{T}}, \ [1,0]^{\mathrm{T}}, \ [-1,0]^{\mathrm{T}}, \ [3,1]^{\mathrm{T}}, \ [-3,1]^{\mathrm{T}} \ .$$

Έστω επίσης και ένα διάνυσμα δοκιμής $\mathbf{x}_{\text{test}} = [-3,-2]^{\mathrm{T}}$.

(a) Σχεδιάστε τα παραπάνω διανύσματα στο 2-D χώρο με άξονες τα x_1, x_2. Είναι οι κλάσεις γραμμικά διαχωρίσιμες (με βάση τα διανύσματα εκπαίδευσης);

(b) Ταξινομήστε το διάνυσμα δοκιμής με βάση τον κανόνα των τριών πλησιέστερων γειτόνων (3-NN), χρησιμοποιώντας την Ευκλείδεια απόσταση (L_2) μεταξύ διανυσμάτων.

(c) Κατασκευάστε (και σχεδιάστε) ένα perceptron δύο επιπέδων που να ταξινομεί τα διανύσματα εκπαίδευσης σωστά στις δύο κλάσεις. Σε ποια κλάση ταξινομείται το διάνυσμα δοκιμής $\mathbf{x}_{\text{test}}$;

(d) Λύστε το πρόβλημα χρησιμοποιώντας πολυωνυμικό ταξινομητή και ταξινομήστε το διάνυσμα δοκιμής με αυτόν. Επίσης, σχεδιάστε τις περιοχές απόφασης για κάθε κλάση στον αρχικό χώρο.

(e) Τέλος, μειώστε την διάσταση των δεδομένων χρησιμοποιώντας ανάλυση κύριων συνιστωσών (Principal Component Analysis – PCA). Είναι οι κλάσεις διαχωρίσιμες στο μονοδιάστατο χώρο που προκύπτει;

Θέμα 3: (25%) Έστω τα πρώτα 4 διανύσματα της κλάσης ω_2 του Θέματος 2, δηλ.

$$\mathbf{x}_1 = [3,-1]^{\mathrm{T}}, \ \mathbf{x}_2 = [-3,-1]^{\mathrm{T}}, \ \mathbf{x}_3 = [1,0]^{\mathrm{T}}, \ \mathbf{x}_4 = [-1,0]^{\mathrm{T}} .$$

(a) Ομαδοποιήστε τα διανύσματα ακολουθιακά, χρησιμοποιώντας το βασικό ακολουθιακό αλγοριθμικό σχήμα (BSAS) με βάση την απόσταση L_1 (Manhattan distance) και κατώφλι $\theta = 3.5$. Θεωρήστε δύο περιπτώσεις όσον αφορά την ακολουθία που εμφανίζονται τα διανύσματα στον αλγόριθμο: Στην πρώτη περίπτωση με τη σειρά της εκφώνησης (αριστερά προς δεξιά), και στην δεύτερη περίπτωση με την αντίθετη σειρά. Και στις δύο περιπτώσεις, θεωρήστε ότι οι αντιπρόσωποι των ομάδων που προκύπτουν είναι η μέση τιμή των μελών της ομάδας.

(b) Ομαδοποιήστε τα διανύσματα ιεραρχικά, χρησιμοποιώντας τον συσσωρευτικό αλγόριθμο ομαδοποίησης (agglomerative clustering) με βάση την απόσταση L_1 (Manhattan distance). Χρησιμοποιείστε δύο διαφοροποιήσεις του αλγορίθμου, όσον αφορά την ενημέρωση των αποστάσεων του πίνακα ανομοιότητας, τον πρώτο με βάση τον αλγόριθμο απλού δεσμού (single link), και τον δεύτερο με βάση τον αλγόριθμο πλήρους δεσμού (complete link). Και στις δύο περιπτώσεις σχεδιάστε τα δεντρογράμματα ανομοιότητας (dendrograms) που προκύπτουν.