ΤΟΜΕΑΣ ΗΛΕΚΤΡΟΝΙΚΗΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ
ΣΥΣΤΗΜΑΤΩΝ ΠΛΗΡΟΦΟΡΙΚΗΣ
ΕΡΓΑΣΤΗΡΙΟ ΑΝΑΛΥΣΗΣ ΗΛΕΚΤΡΙΚΩΝ ΚΥΚΛΩΜΑΤΩΝ
ΜΑΘΗΜΑ: ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ – 9ο ΕΞΑΜΗΝΟ

Εργασία 1

(Προθεσμία: Κυριακή 5 Νοεμβρίου 2023, 23:59)

1. Έστω ότι θέλετε να κατασκευάσετε ένα σύστημα που ταξινομεί email σε τρείς κατηγορίες {n:Normal, s:Spam, m:Malicious}. Έχετε στη διάθεσή σας έναν αλγόριθμο που διαβάζει τα περιεχόμενα κάθε email και βαθμολογεί την επικινδυνότητά του D σε πέντε στάθμες από το 1 έως το 5. Χρησιμοποιώντας έναν μεγάλο όγκο επισημασμένων email, μπορείτε να υπολογίσετε την κατανομή της εκτιμηθείσας επικινδυνότητας για καθεμία από τις κατηγορίες email, η οποία δίδεται από τον ακόλουθο πίνακα:

	Normal	Spam	Malicious
D=1	0.5	0.05	0.02
D=2	0.23	0.15	0.13
D=3	0.16	0.4	0.15
D=4	0.1	0.3	0.3
D=5	0.01	0.1	0.4

Δεδομένου ότι για τις a priori πιθανότητες των κλάσεων γνωρίζετε ότι P(s)=0.3 και P(m)=0.1, να υπολογίσετε:

α) Την απόφαση που θα παίρνει το σύστημα για κάθε πιθανή τιμή D με βάση τον κανόνα απόφασης Bayes.

β) Το ολικό σφάλμα ταξινόμησης.

γ) Το ολικό σφάλμα ταξινόμησης στην περίπτωση που δεν γνωρίζατε τίποτα για τις a priori πιθανότητες των κλάσεων (τις θεωρούσατε ισοπίθανες).

2. Για ένα πρόβλημα ταξινόμησης 2 κλάσεων $\Omega = \{\omega_1,\omega_2\}$, όπου $P(x|\omega_1) = N(2, 0.5)$ και $P(x|\omega_2) = N(1.8, 0.2)$, με εκ των προτέρων πιθανότητα $P(\omega_1) = 1/4$, όπου το κόστος έχει ορισθεί ως $\lambda = \begin{bmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 3 & 0 \end{bmatrix}$

α) Να βρεθεί η βέλτιστη λύση (κανόνας απόφασης) και να υπολογισθεί το ολικό κόστος.

β) Να προσομοιωθεί η διαδικασία υπολογιστικά, δημιουργώντας τυχαία δείγματα που ακολουθούν την κανονική κατανομή, και να εκτιμηθεί αριθμητικά το κόστος από την λύση (α).

3. Με στόχο την ταξινόμηση δεδομένων με Bayesian κριτήρια, να εκτελεστούν τα παρακάτω βήματα:

α) Να υλοποιήσετε τρεις συναρτήσεις στη γλώσσα προγραμματισμού της αρεσκείας σας που να υπολογίζουν:

i. Την τιμή της συνάρτησης διάκρισης της μορφής

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln(|\boldsymbol{\Sigma}_i|) + \ln(P(\omega_i))$$

για μια δεδομένη κανονική κατανομή $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ d διαστάσεων και εκ των προτέρων πιθανότητα $P(\omega_i)$.

ii. Την Ευκλείδεια απόσταση μεταξύ δύο αυθαίρετων σημείων $\mathbf{x_1}$ και $\mathbf{x_2}$ σε χώρο d διαστάσεων.

iii. Την απόστασης Mahalanobis μεταξύ του μέσου $\boldsymbol{\mu}$ και ενός αυθαίρετου σημείου $\mathbf{x}$ σε χώρο d διαστάσεων, δεδομένου του πίνακα συνδιασποράς $\boldsymbol{\Sigma}$.

β) Να κατεβάσετε από εδώ το αρχείο δεδομένων που περιέχει δεδομένα από 3 κλάσεις ω_i, $i = 1,2,3$ και τρία χαρακτηριστικά x_i, $i = 1,2,3$. Υποθέτοντας πως οι υποκείμενες κατανομές των κλάσεων είναι Gaussian, να κάνετε εκτίμηση των παραμέτρων των κατανομών με τη μέθοδο της μεγίστης πιθανοφάνειας για τις περιπτώσεις που χρησιμοποιείται μόνο το πρώτο χαρακτηριστικό x_1, το πρώτο και το δεύτερο χαρακτηριστικό (x_1 και x_2) και τέλος για όλα τα χαρακτηριστικά.

γ) Αφού επιλέξτε τον κατάλληλο ταξινομητή (αιτιολογώντας την επιλογή σας) προσδιορίζεται υπολογιστικά το εμπειρικό σφάλμα ταξινόμησης, δηλαδή το ποσοστό των σημείων που ταξινομείται εσφαλμένα, υποθέτωντας ότι οι a priori πιθανότητες είναι $P(\omega_1) = P(\omega_2) = ½$, κάνοντας χρήση μόνο του πρώτου χαρακτηριστικού x_1.

δ) Επαναλάβετε το προηγούμενο βήμα χρησιμοποιώντας πρώτα δύο χαρακτηριστικά (x_1 και x_2) και τέλος όλα τα χαρακτηριστικά. Σχολιάστε τα αποτελέσματά σας και ιδιαίτερα τη σχέση του εμπειρικού σφάλματος με τις διαστάσεις του προβλήματος.

ε) Επαναλάβετε την εκτίμηση του σφάλματος χρησιμοποιώντας όλα τα χαρακτηριστικά για $P(\omega_1) = 0.8$ και $P(\omega_2) = P(\omega_3) = 0.1$. Σε ποια περίπτωση μπορούμε να χρησιμοποιήσουμε καθέναν από τους ταξινομητές που υλοποιήσατε στο βημα (α);

4. Έστω ότι σχεδιάζετε ένα Bayesian σύστημα ταξινόμησης σε 2 κλάσεις $\omega = \{\omega_1, \omega_2\}$, και καταλήγετε πως η πιθανότητα σφάλματος του συστήματος εξαρτάται από την a priori πιθανότητα P_1 της κλάσης ω_1 σύμφωνα με τη σχέση:

$$P_{error}(p_1) = p_1^2(1-p_1), \ 0 \le p_1 \le 1$$

α) Εάν δεν γνωρίζετε το P_1, για ποια τιμή του θα πρέπει να σχεδιάσετε τα όρια απόφασης του συστήματός σας ώστε να ελαχιστοποιήσετε το μέγιστο πιθανό σφάλμα, και ποια θα είναι η πιθανότητα σφάλματος στην περίπτωση αυτή;

β) Εάν ρυθμίσατε το σύστημά σας υποθέτοντας πιθανότητα $P_1=0.3$, ποια θα είναι η πιθανότητα σφάλματος εάν η πραγματική a priori πιθανότητα της κλάσης ω_1 είναι $P_1=0.7$;

5. Αναδρομική εκτίμηση Bayes: Πετάμε ένα νόμισμα N=10 φορές και φέρνουμε κατά σειρά {κ,γ,κ,κ,κ,γ,κ,κ,γ,κ} (κ=κεφάλι, γ=γράμματα). Ποια είναι η πιθανότητα θ να φέρουμε κεφάλι;

Η αρχική κατανομή του θ είναι $p(\theta \mid D^0) = A \cdot \theta(1-\theta)^4$, για $0 \le \theta \le 1$ (beta distribution).

Ζητείται:

α) Να υπολογισθεί το A .

β) Να σχεδιασθούν σε κοινό διάγραμμα τα $p(\theta \mid D^1)$, $p(\theta \mid D^5)$, $p(\theta \mid D^{10})$

γ) Να βρεθεί (αριθμητικά) το $p(x = \gamma \mid D^{10})$ μετά τη 10^{η} ρίψη.

Οδηγίες: *Υποβάλετε τις απαντήσεις σας στην πλατφόρμα του e-class στην ενότητα εργασίες, πριν τη λήξη της προθεσμίας υποβολής. Η υποβολή σας πρέπει να αποτελείται από* ***ένα μόνο συμπιεσμένο αρχείο*** *(.rar, .zip κλπ.) το οποίο θα περιέχει όλα τα απαραίτητα αρχεία για την υποβολή σας. Καταγράφετε τις απαντήσεις σας σε μία αναφορά που θα αποστείλετε σε μορφή pdf ή docx. Για τις ασκήσεις που απαιτούν και συγγραφή κώδικα, οι αποδεκτές γλώσσες προγραμματισμού είναι Matlab ή Python, όπου και θα πρέπει να συμπεριλάβετε στο συμπιεσμένο αρχείο και τα αρχεία του πηγαίου κώδικα (αρχεία .m, .py κλπ) με τα απαραίτητα σχόλια για την τεκμηρίωση εντός του κώδικα. υποβάλλοντας το πολύ ένα αρχείο για κάθε άσκηση. Σε περίπτωση που χρησιμοποιήσετε Matlab, μπορείτε να συμπεριλάβετε τον κώδικα για όλες τις ασκήσεις σε ένα αρχείο, χωρίζοντας τις απαντήσεις σε διαφορετικά cells. Επίσης, για τις ασκήσεις που περιλαμβάνουν κώδικα, αντί απάντησης στην αναφορά μπορείτε να υποβάλετε Matlab live scripts ή Python notebooks κατάλληλα δομημένα ώστε να περιλαμβάνουν και τις απαντήσεις/σχολιασμούς που ζητούνται.*