ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ

Τμήμα Ηλεκτρολόγων Μηχ. Και Μηχ. Υπολογιστών

Στατιστική Μοντελοποίηση και Αναγνώριση Προτύπων (ΤΗΛ311)

Φυλλάδιο Ασκήσεων 3

Οδηγίες:

1. Σας παρακαλώ να σεβαστείτε τον παρακάτω κώδικα τιμής τον οποίον θα θεωρηθεί ότι προσυπογράφετε μαζί με τη συμμετοχή σας στο μάθημα και τις εργασίες του:
   a) *Οι απαντήσεις στις εργασίες, τα quiz και τις εξετάσεις, ο κώδικας και γενικά οτιδήποτε αφορά τις εργασίες θα είναι προϊόν δικής μου δουλειάς.*
   b) *Δεν θα διαθέσω κώδικα, απαντήσεις και εργασίες μου σε κανέναν άλλο.*
   c) *Δεν θα εμπλακώ σε άλλες ενέργειες με τις οποίες ανέντιμα θα βελτιώνω τα αποτελέσματα μου ή ανέντιμα θα αλλάζω τα αποτελέσματα άλλων.*
2. Η εργασία είναι ατομική
3. Ημερομηνία παράδοσης: **Παρασκευή, 1/6/2018 στις 20:00**
4. **Παραδοτέα:** α) Κώδικας και β) Αναφορά με τις απαντήσεις, παρατηρήσεις, πειράματα, αποτελέσματα και οδηγίες χρήσης του κώδικα.

Θέμα 1: Ταξινόμηση λουλουδιών (k-NN)

Σε αυτή την άσκηση πρόκειται να χρησιμοποιήσετε ένα πολύ γνωστό σύνολο δεδομένων για τα λουλούδια που χρησιμοποιήθηκε αρχικά το 1936 από τον Fisher στο βιβλίο του «*Η χρήση πολλαπλών μετρήσεων σε προβλήματα ταξινόμησης*», και από τότε το σύνολο δεδομένων έχει καθιερωθεί για τη δοκιμή διαφόρων στατιστικών τεχνικών και αλγορίθμων. Το σύνολο δεδομένων, είναι ευρέως γνωστό στη βιβλιογραφία ως «*το σύνολο δεδομένων των λουλουδιών Iris*» ή «*το σύνολο δεδομένων Fisher's Iris*», αποτελείται από 50 δείγματα από καθένα από τα τρία είδη Iris (Iris setosa, Iris virginica και Iris versicolor). Από κάθε δείγμα μετρήθηκαν τέσσερα χαρακτηριστικά: το μήκος και το πλάτος των σεπάλων και των πετάλων σε εκατοστά. Με βάση το συνδυασμό αυτών των τεσσάρων χαρακτηριστικών, ο Fisher ανέπτυξε ένα γραμμικό μοντέλο διάκρισης για να διακρίνει το είδος το ένα από το άλλο. Οι μετρούμενες τιμές μπορούν να βρεθούν στο iris.dat. Κάθε σειρά αντιστοιχεί σε ένα διαφορετικό δείγμα. Οι στήλες περιέχουν τα μετρηθέντα χαρακτηριστικά ως εξής:

- Μήκος σέπαλου σε cm (feature x1)
- Πλάτος σέπαλου σε cm (feature x2)
- Μήκος πέταλου σε cm (feature x3)
- Πλάτος πέταλου σε cm (feature x4)
- Iris class ($\omega_i = 0,1,2$ όπου ω_0: Iris Setosa, ω_1: Iris Versicolor, ω_2: Iris Verginica)

α) Γράψτε κώδικα για να εφαρμόσετε τον k-NN ταξινομητή (k-Nearest Neighbor). Ο ταξινομητής θα χρησιμοποιεί την Ευκλείδεια απόσταση.

β) Παρουσιάστε τον πίνακα confusion matrix για διάφορες τιμές του k.

γ) Σχολιάστε τα αποτελέσματά σας.

Θέμα 2: Λογιστική Παλινδρόμηση

Υποθέτουμε ότι έχουμε ένα σύνολο m παραδειγμάτων $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$, όπου $x^{(i)} \in \mathbb{R}^n$ και $y^{(i)} \in \{0,1\}$. Θέλουμε να προβλέψουμε τις τιμές των $y^{(i)}$ από τις αντίστοιχες τιμές $x^{(i)}$, $i \in \{1,2, \dots, m\}$, χρησιμοποιώντας την συνάρτηση της λογιστικής παλινδρόμησης, η οποία ορίζεται ως εξής

$$h_\theta(x) = g(\theta^T x)$$

όπου

$$g(z) = \frac{1}{1+e^{-z}}$$

Αν $\hat{y}^{(i)} = h_\theta(x^{(i)})$ είναι η εκτίμηση της λογιστικής συνάρτησης για το $y^{(i)}$ η συνάρτηση σφάλματος ορίζεται ως εξής:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(-y^{(i)}\ln\left(\hat{y}^{(i)}\right) - (1-y^{(i)})\ln(1-\hat{y}^{(i)})\right)$$

και αντικαθιστώντας το $\hat{y}_i$ έχουμε

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m}\left(-y^{(i)}\ln\left(h_\theta\left(x^{(i)}\right)\right) - (1-y^{(i)})\ln(1-h_\theta(x^{(i)})\right)$$

Η κλίση του σφάλματος $J(\theta)$ είναι ένα διάνυσμα ίσης διάστασης με το θ.

a) Αν θ_j και $x_j^{(i)}$ είναι η j συνιστώσα των διανυσμάτων $\theta = [\theta_1, \theta_2, \dots, \theta_n]^T$ και

$x^{(i)} = \left[x_1^{(i)}, x_2^{(i)}, \dots x_n^{(i)}\right]^T$ αντίστοιχα, να δείξετε ότι το j-στοιχείο της κλίσης του σφάλματος είναι:

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)x_j^{(i)}$$

b) Θα χρησιμοποιήσουμε την γραμμική παλινδρόμηση για να προβλέψουμε αν ένας φοιτητής θα γίνει δεκτός σε ένα πανεπιστήμιο με βάση τους βαθμούς του σε δύο εξετάσεις. Υπάρχουν δεδομένα αρχείου από παλαιότερες αιτήσεις που θα χρησιμοποιηθούν ως δεδομένα εκμάθησης της λογιστικής παλινδρόμησης. Θα συμπληρώσετε κώδικα ώστε να τρέξει η άσκηση ex3_2_logisticRegression.m

i) Αρχικά δείτε τα δεδομένα με την συνάρτηση plotData.m

ii) Υλοποιήστε την σιγμοειδή συνάρτηση $g(z)$ στο αρχείο sigmoid.m. Αν η είσοδος z είναι πίνακας, η sigmoid θα πρέπει να εφαρμόζει την $g(z)$ σε κάθε στοιχείο του πίνακα.

iii) Υλοποιήστε την συνάρτηση κόστους $J(\theta)$ στο αρχείο costFunction.m (Σημείωση για τις αρχικές τιμές του θ το κόστος πρέπει να είναι περίπου 0.693). Επίσης στο ίδιο αρχείο υλοποιήστε την κλίση του σφάλματος (για τις αρχικές τιμές του θ η κλίση πρέπει να είναι περίπου [-0.1, -12.009217, -11.262842]).

iv) Η βελτιστοποίση των παραμέτρων γίνεται με κώδικα που υπάρχει έτοιμος στην άσκηση ex3_2_logisticRegression.m. Εσείς απλά τρέξτε τον κώδικα και βρείτε το σύνορο

απόφασης με την συνάρτηση plotDecisionBoundary.m. Επίσης τρέξτε την συνάρτηση predict.m για να προβλέψετε αν ο φοιτητής θα γίνει δεκτός με βάση διάφορες τιμές βαθμών στις δύο εξετάσεις.

Θέμα 3: HMMs

Έστω ένα κρυφό Μαρκοβιανό μοντέλο (HMM) λ τριών καταστάσεων και δύο συμβόλων εξόδου [x, y]. Οι αρχικές πιθανότητες των καταστάσεων είναι $\pi = [1 \quad 0 \quad 0]$ αντίστοιχα, ενώ οι πιθανότητες μετάβασης από τη μία κατάσταση στην άλλη έχουν τις τιμές που δίνονται στον παρακάτω πίνακα.

		Τελική Κατάσταση		
		1	2	3
Αρχική Κατάσταση	1	0.5	0.4	0.1
	2	0	0.7	0.3
	3	0	0	1.0

Έστω ότι οι πιθανότητες εξόδου είναι:

		Καταστάσεις		
		1	2	3
Παρατηρήσεις	P(x)	0.8	0.6	0.1
	P(y)	0.2	0.4	0.9

1. Σχεδιάστε το διάγραμμα μεταβάσεων των καταστάσεων του παραπάνω HMM γράφοντας πάνω από κάθε τόξο την αντίστοιχη πιθανότητα μετάβασης.

2. Σχεδιάστε το διάγραμμα trellis με όλα τα επιτρεπόμενα μονοπάτια για το μοντέλο θεωρώντας ότι έχετε μια ακολουθία παρατηρήσεων μήκους 4. Θεωρήστε ότι η τελική κατάσταση είναι η q_3.
3. Για την ακολουθία παρατηρήσεων *O=[xxxy]* βρείτε την πιθανότητα $P(O \mid \lambda)$ από όλα τα πιθανά μονοπάτια που καταλήγουν στην κατάσταση q_3 τη χρονική στιγμή *T=4*.
4. Για την ίδια ακολουθία παρατηρήσεων, υπολογίστε την πιθανότητα P^* του πιο πιθανού μονοπατιού. Βρείτε ποιο είναι το πιο πιθανό μονοπάτι.

Θέμα 4: K-means clustering

Σε αυτή την άσκηση θα υλοποιήσετε τον K-means clustering αλγόριθμο και θα τον χρησιμοποιήσετε για να συμπιέσετε μια εικόνα. Θα αρχίσετε με ένα διδιάστατο, 2D, σύνολο δεδομένων για να καταλάβετε πως λειτουργεί ο K-means. Ο K-means είναι ένας αλγόριθμος που ομαδοποιεί όμοια δεδομένα. Έστω ότι έχουμε ένα σύνολο $X = \{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$ με m n-διάστατα δείγματα ($x^{(i)} \in \mathbb{R}^n$), τα οποία θέλουμε να τα ομαδοποιήσουμε σε K-κλάσεις. Ο K-means είναι μια επαναληπτική διαδικασία που αρχικά θέτει τυχαίες τιμές στα κέντρα των κλάσεων, και στην συνέχεια βελτιώνει την αρχική του επιλογή τοποθετώντας τα παραδείγματα στην κλάση με την μικρότερη απόσταση μεταξύ του παραδείγματος και του κέντρου της κλάσης και ξαναϋπολογίζοντας τα κέντρα των κλάσεων.

Ο αλγόριθμος K-means είναι ο ακόλουθος:

```
% Initialize centroids
centroids = kMeansInitCentroids(X, K);
for iter = 1:iterations
    % Cluster assignment step: Assign each data point to the
    % closest centroid. idx(i) corresponds to cˆ(i), the index
    % of the centroid assigned to example i
    idx = findClosestCentroids(X, centroids);
    % Move centroid step: Compute means based on centroid
    % assignments
    centroids = computeMeans(X, idx, K);
end
```

Τρέξτε το Matlab/Octave script ex3_kmeans.m. Θα πρέπει να συμπληρώσετε τα επόμενα μέρη του προγράμματος.

a) Να συμπληρώσετε την συνάρτηση findClosestCentroids.m
   $c^{(i)} \coloneqq j$ that minimizes $\left\|x^{(i)} - \mu_j\right\|^2$, όπου $c^{(i)}$ είναι ο δείκτης του κοντινότερου κέντρου στο $x^{(i)}$, και μ_j είναι το διάνυσμα τιμών (συντεταγμένων) του κέντρου j. Το $c^{(i)}$ αντιστοιχεί στο `idx(i)` του παραπάνω κώδικα.

b) Να συμπληρώσετε την συνάρτηση computeCentroids.m

$$\mu_k = \frac{1}{|C_k|} \sum_{i \in C_k} x^{(i)}$$

   όπου C_k είναι το σύνολο των παραδειγμάτων που έχουν αντιστοιχηθεί στην κλάση k.

c) Συμπληρώστε την εντολή X_recovered = …. Στο Matlab/Octave script ex3_kmean.m και δείτε την συμπιεσμένη εικόνα για διάφορες τιμές του K (αριθμός κλάσεων).