1 Στόχοι εργαστηρίου

1. Η υλοποίηση των simple linear regression, multivariate linear regression, logistic regression και K-means μοντέλων σε MATLAB.
2. Η παρατήρηση και κατανόηση της επιρροής που έχουν οι παράμετροι κάθε μοντέλου στην απόδοση του.
3. Η επεξήγηση των αποτελεσμάτων σας βάσει της θεωρίας που διδαχτήκατε στις διαλέξεις.

2 Προαπαιτούμενα

Για αυτό το εργαστήριο θα σας δοθούν τα παρακάτω αρχεία δεδομένων:

1. `linear_regressionx.dat` και `linear_regressiony.dat`
2. `mv_regressionx.dat` και `mv_regressiony.dat`
3. `log_regressionx.dat` και `log_regressiony.dat`
4. `kmeans.dat`

και τα εξής αρχεία κώδικα:

1. `simple_linear_regression.m`
2. `multivariate_linear_regression.m`
3. `logistic_regression.m`
4. `k_means.m`

Για κάθε διαφορετικό τύπο μοντέλου θα χρησιμοποιήσουμε τα αρχεία κώδικα και δεδομένων με την αντίστοιχη ονομασία. Κάθε αρχείο κώδικα εμπεριέχει μια έτοιμη λύση βασισμένη στο Statistics and Machine Learning Toolbox™ του MATLAB, και θα σας ζητείται να λύσετε κάποια προβλήματα, τα οποία θα βρείτε αναλυτικά παρακάτω.

3 Supervised Learning

3.1 Regression

3.1.1 Simple Linear Regression

Για αυτό το μοντέλο θα χρησιμοποιήσουμε αριθμητικά δεδομένα από τα αρχεία `linear_regressionx.dat` και `linear_regressiony.dat`. Κάθε δεδομένο (data point) έχει ένα μόνο χαρακτηριστικό (feature). Συνολικά έχουμε 50 δεδομένα. Το αρχείο κώδικα που θα χρειαστείτε είναι το `simple_linear_regression.m`.

Task 1: Χρησιμοποιώντας τη μέθοδο `regress(y,x)` του Statistics and Machine Learning Toolbox^{TM} επαναλάβετε την υλοποίση του Simple Linear Regression μοντέλου. Δημιουργήστε τον αντίστοιχο γράφο με τα αποτελέσματα σας και συγκρίνετε τα με αυτά που σας έχουν δοθεί.

Task 2: Υλοποιήστε τον Simple Linear Regression αλγόριθμο με χρήση Stochastic Gradient Descent (SGD). Ανακαλέστε την εξίσωση για το μοντέλο:

$$h_\theta(x) = \theta^T x = \sum_{i=0}^{n} \theta_i x_i$$

και ότι ο κανόνας ανανέωσης του SGD είναι:

$$\theta_i := \theta_i - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

Προτείνεται να χρησιμοποιήσετε $\alpha = 0.07$ και αριθμό επαναλήψεων ίσο με 1500. Δημιουργήστε τον αντίστοιχο γράφο με τα αποτελέσματα σας και συγκρίνετε τα με αυτά που σας έχουν δοθεί.

3.1.2 Multivariate Linear Regression

Εδώ χρειαζόμαστε παραπάνω από ένα χαρακτηριστικό από κάθε δεδομένο. Για αυτό χρησιμοποιούμε ένα υποσύνολο των τιμών κατοικιών στο Portland, Oregon. Πιο συγκεκριμένα θα χρησιμοποιηθούν 47 datapoints με 2 χαρακτηριστικά το καθένα (περιοχή, αριθμός υπνοδωματίων). Τα δεδομένα είναι αποθηκευμένα στα αρχεία `mv_regressionx.dat` και `mv_regressiony.dat`. Το αρχείο κώδικα που θα χρειαστείτε είναι το `multivariate_linear_regression.m`.

Task 3: Υλοποιήστε το Multivariate Linear Regression μοντέλο με χρήση Stochastic Gradient Descent (SGD) για διαφορετικές τιμές του α. Οι τύποι που χρειάζεστε για την υλοποίηση του είναι οι ίδιοι με αυτούς του **Task 2**. Προτείνεται να χρησιμοποιήσετε $0.01 \leq \alpha \leq 1.3$. Δημιουργήστε τον αντίστοιχο γράφο με τα αποτελέσματα για κάθε διαφορετικό α και εξηγείστε ποιο α είναι το καλύτερο στη περίπτωση σας.

3.2 Classification

3.2.1 Logistic Regression

Για τα παρακάτω tasks θα χρησιμοποιήσουμε αριθμητικά δεδομένα από τα αρχεία `log_regressionx.dat` και `log_regressiony.dat`, για να λύσουμε ένα δυαδικό πρόβλημα. Οι δύο κλάσεις είναι: μαθητές που μπήκαν στο πανεπιστήμιο και μαθητές που δεν μπήκαν. Το κάθε δεδομένο έχει 2 χαρακτηριστικά: τους βαθμούς σε δύο διαφορετικά διαγωνίσματα (δλδ (results1, results2)). Το αρχείο κώδικα που θα χρειαστείτε είναι το `logistic_regression.m`.

Task 4: Υλοποιήστε το Logistic Regression μοντέλο με χρήση Stochastic Gradient Descent (SGD). Ανακαλέστε την εξίσωση για το μοντέλο:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}} = P(y = 1|x; \theta)$$

ο κανόνας ανανέωσης του SGD είναι ο ίδιος με τα προηγούμενα task.

Task 5: Υλοποιήστε το Logistic Regression μοντέλο με χρήση της μεθόδου Newton. Το cost function είναι το ακόλουθο:

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

και εσείς καλείστε να το μειώσετε χρησιμοποιώντας την μεθόδο Newton και το κανόνα ανανέωσης του που είναι ίσος με:

$$\theta^{(t+1)} = \theta^{(i)} - H^{-1} \bigtriangledown_\theta J$$

όπου:

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

και

$$H = \frac{1}{m}\sum_{i=1}^{m}[h_\theta(x^{(i)})(1 - h_\theta(x^{(i)}))x^{(i)}(x^{(i)})^T]$$

Συγκρίνετε τα αποτελέσματα των θ με αυτά των προηγούμενων υλοποιήσεων.

Tip: Η μέθοδος του Newton συχνά συγκλίνει σε 5-15 επαναλήψεις επομένως αν χρειαστείτε παραπανώ υπάρχει κάποιο λάθος στην υλοποίηση σας.

4 Unsupervised Learning

4.1 K-means Clustering

Για αυτό το μοντέλο θα χρησιμοποιήσουμε αριθμητικά δεδομένα από το `kmeans.dat` αρχείο, το οποίο περιέχει 150 δεδομένα με 2 χαρακτηριστικά το καθένα. Σκοπός σας είναι να υλοποιήσετε τον k-means για διαφορετικές τιμές του k (k: number of clusters). Το αρχείο κώδικα που θα χρειαστείτε είναι το `k_means.m`.

Task 6: Γράψτε έναν αλγόριθμο που θα υλοποιεί το K-means Clustering μοντέλο χωρίς τη χρήση κάποιας βιβλιοθήκης για διαφορετικές τιμές της μεταβλητής k (δηλ. k=3,4,5).

Tip: Ξεκινήστε την υλοποίηση σας διαλέγοντας τυχαία σημεία για κέντρα των k-clusters που δημιουργείτε.