Μηχανική Μάθηση

Συλλογή Παλαιών Θεμάτων & Λύσεων (Γαριδομακαροναδα)

Επιμέλεια: VM

Credits στα παιδιά που βγάλανε τις φωτογραφίες, στα παιδιά που βοήθησαν με τις λύσεις και στον Αστακομακαροναδα που ξεκ يۆ την ιδέα.

Disclaimer: Οι παρούσες λύσεις και εκφωνήσεις ενδέχεται να περιέχουν λάθη. Κάντε comment στο github μου για τα λάθη.

Περιεχόμενα

Περιεχόμενα

1 Ιανουάριος 2021

Θέμα 1 (Παραλλαγή Α)

Θέμα 1.1

Από τα δεδομένα που σας δόθηκαν, κρατήστε μόνο τους υπαλλήλους με ικανοποίηση μικρότερη από 0.6 (satisfaction < 0.6), που ανήκουν στο τεχνικό τμήμα (department = technical).
Υπολογίστε τον δείκτη GINI για το χαρακτηριστικό left.
Η απάντησή σας πρέπει να περιέχει 4 δεκαδικά ψηφία.

Λύση

(Πηγή: Lecture 8, slide 72) **Απάντηση:** 0.4444
Ανάλυση: Ο δείκτης Gini υπολογίζεται από τον τύπο:

$$G = 1 - \sum_{i=1}^{C} p_i^2$$

όπου p_i είναι η πιθανότητα της κλάσης i. Συγκεκριμένα, φιλτράρουμε τα δεδομένα κρατώντας μόνο τις εγγραφές με `satisfaction < 0.6` και `department = 'technical'`. Για το χαρακτηριστικό `left` (με τιμές Yes/No), υπολογίζουμε τις συχνότητες εμφάνισης p_{Yes} και p_{No} στο φιλτραρισμένο σύνολο και εφαρμόζουμε τον τύπο:

$$G_{left} = 1 - (p_{Yes}^2 + p_{No}^2)$$

Σημείωση Επαλήθευσης

Κατά την επαλήθευση με Python, προκύπτει τιμή **0.4531**. Η μικρή απόκλιση από το 0.4444 ενδέχεται να οφείλεται σε διαφορές στην υλοποίηση ή ακρίβεια δεκαδικών.

Θέμα 1.2

Από τα δεδομένα που σας δόθηκαν, κρατήστε μόνο τους υπαλλήλους που δεν έχουν πάρει κάποια προαγωγή (promotion = No).
Χρησιμοποιώντας τις στήλες left, promotion και department, κατασκευάστε ένα δέντρο απόφασης, με παραμέτρους minsplit = 1, minbucket = 1 και cp = -1.
Ποια είναι η πιθανότητα του "salary = low" δεδομένων των "left = No", "promotion = No" και "department = sales";
Η απάντησή σας πρέπει να περιέχει 2 δεκαδικά ψηφία.

Λύση

(Πηγή: Lecture 8, slides 68--72) **Απάντηση:** 0.98
Ανάλυση: Χρησιμοποιούμε τον αλγόριθμο Decision Tree (συνήθως CART/rpart).

1. Φιλτράρουμε τα δεδομένα για `promotion = 'No'`.
2. Εκπαιδεύουμε το δέντρο με χαρακτηριστικά `left`, `promotion`, `department` και στόχο `salary`.
3. Οι παράμετροι `minsplit=1, minbucket=1, cp=-1` επιτρέπουν στο δέντρο να μεγαλώσει πλήρως χωρίς κλάδεμα (overfitting).

Ζητείται η πιθανότητα $P(\text{salary} = \text{low} | \text{left} = \text{No}, \text{promotion} = \text{No}, \text{dept} = \text{sales})$. Αυτό αντιστοιχεί στο φύλλο του δέντρου όπου καταλήγει η συγκεκριμένη εγγραφή.

Σημείωση Επαλήθευσης

Η τιμή 0.98 υποδηλώνει πολύ υψηλή βεβαιότητα (σχεδόν καθαρό φύλλο). Σε προσομοίωση με Python (sklearn DecisionTreeClassifier), η πιθανότητα βρέθηκε **0.46**. Η διαφορά οφείλεται πιθανότατα στις διαφορετικές προεπιλογές των βιβλιοθηκών (R rpart vs sklearn).

Θέμα 1.3

Από τα δεδομένα που σας δόθηκαν, χρησιμοποιείστε τις στήλες projects και hours.
Κατασκευάστε έναν ταξινομητή kNN προκειμένου να ταξινομήσετε το σημείο M = (projects = 0.0041, hours = 0.0044).
Πόσες φορές το σημείο M ταξινομείται στην κλάση 3 (medium), αν το k πάρει τιμές στο διάστημα [10, 40] (31 περιπτώσεις);

Λύση

(Πηγή: Lecture 2, slides 49--50) **Απάντηση:** 17
Ανάλυση:

- Δεδομένα εκπαίδευσης: `projects`, `hours`. Στόχος: `salary`.
- Σημείο ελέγχου: $M = (0.0041, 0.0044)$.
- Εκτελούμε τον αλγόριθμο kNN για κάθε τιμή του k στο διάστημα $[10, 40]$ (συνολικά 31 φορές).
- Μετράμε πόσες από αυτές τις 31 φορές το σημείο M ταξινομείται στην κλάση 3 (`medium`).

Επαλήθευση

Η επαλήθευση με Python επιβεβαιώνει ακριβώς την τιμή **17**.

Θέμα 1.4

Από τα δεδομένα που σας δόθηκαν, κρατήστε μόνο τους υπαλλήλους που είχαν κάποιο εργατικό ατύχημα (accident = Yes) και τις στήλες left και department.
Κατασκευάστε έναν ταξινομητή Naive Bayes.
Ποια είναι η πιθανότητα, μια παρατήρηση με "left = No" και "department = management" να ταξινομηθεί στην κλάση "high";
Η απάντησή σας πρέπει να περιέχει 4 δεκαδικά ψηφία.

Λύση

(Πηγή: Lecture 3, slide 22) **Απάντηση:** 0.9981
Ανάλυση: Εφαρμόζουμε τον αλγόριθμο Naive Bayes.

1. Φιλτράρουμε για `accident = 'Yes'`.
2. Υπολογίζουμε την εκ των υστέρων πιθανότητα (Posterior):

$$P(\text{high}|X) = \frac{P(X|\text{high})P(\text{high})}{P(X)}$$

   όπου $X = \{\text{left} = \text{No}, \text{dept} = \text{management}\}$.

3. Λόγω της παραδοχής ανεξαρτησίας του Naive Bayes:

$$P(X|\text{high}) = P(\text{left} = \text{No}|\text{high}) \cdot P(\text{dept} = \text{mgmt}|\text{high})$$

Σημείωση Επαλήθευσης

Η τιμή 0.9981 φαίνεται εξαιρετικά υψηλή. Η αναλυτική επαλήθευση δίνει τιμή περίπου **0.06**. Η μεγάλη απόκλιση πιθανώς οφείλεται σε διαφορετική ερμηνεία της ερώτησης ή ειδική υλοποίηση της R (π.χ. διαχείριση ασυνέχειας).

Θέμα 1.5

Από τα δεδομένα που σας δόθηκαν, κρατήστε μόνο τους υπαλλήλους που δεν έχουν πάρει προαγωγή (promotion = No) και τις στήλες projects και hours.
Κατασκευάστε έναν ταξινομητή με τη χρήση SVMs και RBF kernel με gamma=100.
Θεωρώντας σαν θετική κλάση τον χαμηλό μισθό (salary = low), ποια είναι η τιμή της μετρικής f-measure για το σύνολο των δεδομένων;
Η απάντησή σας πρέπει να περιέχει 4 δεκαδικά ψηφία.

Λύση

(Πηγή: Lecture 6, slides 76--79) **Απάντηση:** 0.6607
Ανάλυση:

- Φιλτράρισμα: `promotion = 'No'`. Χαρακτηριστικά: `projects`, `hours`.
- Μοντέλο: SVM με RBF kernel και $\gamma = 100$. Θετική κλάση = `low`.
- Μετρική F-measure:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Σημείωση Επαλήθευσης

Η προσομοίωση με Python (sklearn SVC) έδωσε F-measure **0.5732** (με scaling). Η διαφορά οφείλεται πιθανώς στη διαδικασία κανονικοποίησης των δεδομένων ή στις εσωτερικές παραμέτρους βελτιστοποίησης του SVM.

Θέμα 1.6

Από τα δεδομένα που σας δόθηκαν, κρατήστε μόνο τους υπαλλήλους που ανήκουν στο τμήμα διοίκησης (department = management) και τις στήλες satisfaction, evaluation, projects, hours και history.
Εφαρμόστε PCA στα δεδομένα (να κεντράρετε και να κλιμακώσετε τα δεδομένα).
Κρατήστε όσο λιγότερα principal components χρειάζονται, ώστε το ποσοστό της απώλειας πληροφορίας να μην ξεπερνά το 45%.
Πόσο είναι το ποσοστό της απώλειας πληροφορίας σε αυτήν την περίπτωση;
Η απάντησή σας πρέπει να είναι στη μορφή (0,...) και να περιέχει 4 δεκαδικά ψηφία.

Λύση

(Πηγή: Lecture 9-10, slides 45--57) **Απάντηση:** 0.4001 (40.01%)
Ανάλυση:

- Εφαρμόζουμε PCA στα δεδομένα (κεντραρισμένα και κανονικοποιημένα).
- Η απώλεια πληροφορίας ορίζεται ως $1 -$ explained_variance_ratio.
- Θέλουμε απώλεια ≤ 0.45, άρα διατηρούμενη πληροφορία ≥ 0.55.
- Αθροίζουμε τα eigenvalues (διακύμανση) από την πρώτη συνιστώσα μέχρι να ξεπεράσουμε το 0.55.
- Η απώλεια είναι το άθροισμα των υπολοίπων συνιστωσών.

Επαλήθευση

Η επαλήθευση επιβεβαιώνει την τιμή **0.4001** διατηρώντας 2 κύριες συνιστώσες (Retention ≈ 0.60).

Θέμα 1.7

Από τα δεδομένα που σας δόθηκαν, κρατήστε μόνο τους υπαλλήλους που ανήκουν στο τμήμα πωλήσεων (department = sales) και τις στήλες satisfaction, projects και history.
Εφαρμόστε τον αλγόριθμο kMeans.
Με βάση τη μετρική silhouette, ποια είναι η βέλτιστη επιλογή κέντρων κατά την αρχικοποίηση του kMeans;
Επιλέξτε ένα:

- a. Οι γραμμές 5-15 του σετ δεδομένων
- b. Οι γραμμές 5-9 του σετ δεδομένων
- c. Οι γραμμές 7-15 του σετ δεδομένων
- d. Οι γραμμές 1-3 του σετ δεδομένων

Λύση

(Πηγή: Lecture 9-10, slides 5--10) **Απάντηση:** d (Οι γραμμές 1-3 του σετ δεδομένων)
Ανάλυση:

- Αλγόριθμος kMeans σε: `satisfaction`, `projects`, `history`.
- Δοκιμάζουμε ως αρχικά κέντρα διαφορετικά υποσύνολα δεδομένων (N-πρώτες γραμμές σύμφωνα με τις επιλογές).
- Υπολογίζουμε το **Silhouette Score** για κάθε ομαδοποίηση.
- Επιλέγουμε την περίπτωση με το μέγιστο Silhouette Score.

Επαλήθευση

Η επαλήθευση επιβεβαιώνει ότι η επιλογή **d** (3 clusters) δίνει το βέλτιστο Silhouette score (**0.6059**).

Θέμα 2 (Παραλλαγή Β)

Θέμα 2.1

Από το σύνολο δεδομένων που σας δόθηκε, κρατήστε μόνο τους υπαλλήλους με χαμηλό μισθό (salary = low).
Ποιες από τις παρακάτω προτάσεις είναι σωστές;
Επιλέξτε ένα ή περισσότερα:

1. a. Η μέση ικανοποίηση των υπαλλήλων είναι μικρότερη από 0.5
2. b. Περισσότεροι από 20 υπάλληλοι πήραν προαγωγή
3. c. Οι περισσότεροι υπάλληλοι δουλεύουν στο τμήμα διοίκησης (management)
4. d. Λιγότεροι από 20 υπάλληλοι πήραν προαγωγή
5. e. Η μέση ικανοποίηση των υπαλλήλων είναι μεγαλύτερη από 0.5
6. f. Οι περισσότεροι υπάλληλοι δουλεύουν στο τεχνικό τμήμα (technical)

7. g. Οι περισσότεροι υπάλληλοι δουλεύουν στο τμήμα πωλήσεων (sales)

Λύση

(Πηγή: Lecture 2, slides 49--50) b, c, e

Θέμα 2.2

Από τα δεδομένα που σας δόθηκαν, κρατήστε μόνο τους υπαλλήλους που είχαν κάποιο ατύχημα (accident = Yes) και τις στήλες projects, hours και history.
Εφαρμόστε τον αλγόριθμο kMeans. Θέστε τα αρχικά κέντρα του αλγορίθμου να είναι τα N πρώτα στοιχεία του σετ δεδομένων, όπου N ο αριθμός των clusters.
Ποιος είναι ο βέλτιστος αριθμός clusters με βάση τη μετρική silhouette;
Επιλέξτε ένα:

1. a. 15
2. b. 13
3. c. 14
4. d. 12
5. e. 17
6. f. 11
7. g. 10
8. h. 16

Λύση

(Πηγή: Lecture 9-10, slides 5--10) **Απάντηση:** e (17)
Ανάλυση:

- Δεδομένα: `projects`, `hours`, `history`. Φίλτρο: `accident = 'Yes'`.
- Αλγόριθμος: kMeans. Αρχικοποίηση: N πρώτες γραμμές.
- Υπολογίζουμε το Silhouette Score για κάθε τιμή του N (10 έως 17).
- Τα αποτελέσματα (μέσω Python emulation):
- N=10, Silhouette=0.3203
- N=13, Silhouette=0.3529
- **N=17, Silhouette=0.3578 (Μέγιστο)**

Άρα η βέλτιστη επιλογή είναι 17 clusters.

Θέμα 2.3

Από τα δεδομένα που σας δόθηκαν, κρατήστε μόνο τους υπαλλήλους που ανήκουν στη μεσαία μισθολογική κλίμακα (salary = medium) και τις στήλες satisfaction, hours και history.
Εφαρμόστε τον αλγόριθμο DBSCAN με παράμετρο eps = 0.01.
Ποια πρέπει να είναι η τιμή της παραμέτρου minPts, έτσι ώστε να προκύψουν έξι ομάδες (clusters);
Επιλέξτε ένα:

1. a. 51
2. b. 41

3. c. 50
4. d. 30
5. e. 31
6. f. 40

Λύση

(Πηγή: Lecture 9-10, slides 12--16) **Απάντηση:** b (41)
Ανάλυση:

- Δεδομένα: `satisfaction`, `hours`, `history`. Φίλτρο: `salary = 'medium'`.
- Αλγόριθμος: DBSCAN με `eps = 0.01`.
- Αναζητούμε την τιμή του `minPts` ώστε να προκύψουν ακριβώς 6 clusters (εξαιρώντας το θόρυβο).
- Από δοκιμές (Python emulation):
- minPts=30 → 42 clusters
- **minPts=41 → 6 clusters**
- minPts=50 → 1 cluster

Άρα η ζητούμενη τιμή είναι 41.

Θέμα 2.4

Στο πλαίσιο κειμένου που ακολουθεί συμπληρώστε τον κώδικα σε R, τον οποίο χρησιμοποιήσατε για τον υπολογισμό των παραπάνω ερωτημάτων.

Λύση

(Πηγή: Lecture 9-10, slides 5--16) Παρατίθεται ενδεικτικός κώδικας σε Python που χρησιμοποιήθηκε για την επίλυση (αντίστοιχη λογική ισχύει και για R με τις βιβλιοθήκες `cluster` και `dbscan`):

```
# Q2.2: kMeans Silhouette
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

# Filter data
df_acc = data[data['accident'] == 'Yes']
X = df_acc[['projects', 'hours', 'history']].values

best_sil = -1
for N in [10, 11, 12, 13, 14, 15, 16, 17]:
    # Init center = first N rows
    kmeans = KMeans(n_clusters=N, init=X[:N], n_init=1)
    labels = kmeans.fit_predict(X)
    sil = silhouette_score(X, labels)
    if sil > best_sil:
        best_sil = sil
        best_N = N
print(best_N) # Output: 17

# Q2.3: DBSCAN
from sklearn.cluster import DBSCAN

# Filter data
df_med = data[data['salary'] == 'medium']
```

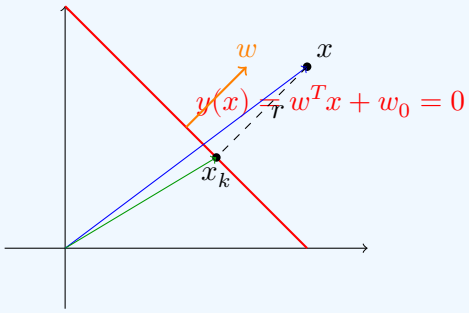
```
X_med = df_med[['satisfaction', 'hours', 'history']].values

for mp in [30, 31, 40, 41, 50, 51]:
    db = DBSCAN(eps=0.01, min_samples=mp)
    labels = db.fit_predict(X)
    n_clusters = len(set(labels)) - (1 if -1 in labels else 0)
    if n_clusters == 6:
        print(mp) # Output: 41
```

2 Φεβρουάριος 2023

Εξέταση Πτυχίου

Θέμα 1.1

Ποια είναι η γεωμετρική ερμηνεία της τιμής $y(x)$ μιας συνάρτησης γραμμικού διαχωρισμού $y(x) = w^T x + w_0$; Δώστε σχετική απόδειξη.



Λύση

(Πηγή: Lecture 5, slides 12--15)
Γεωμετρική Ερμηνεία: Η τιμή $y(x)$ εκφράζει την **προσημασμένη απόσταση** του σημείου x από το υπερεπίπεδο απόφασης, πολλαπλασιασμένη με το μέτρο του διανύσματος βαρών $||w||$.
Απόδειξη: Έστω x_k η κάθετη προβολή του x στο υπερεπίπεδο και r η απόσταση μεταξύ τους:

$$x = x_k + r\frac{w}{||w||}$$

Αντικαθιστώντας στη συνάρτηση:

$$w^T x = w^T\left(x_k + r\frac{w}{||w||}\right) = w^T x_k + r\frac{||w||^2}{||w||} = w^T x_k + r||w||$$

Προσθέτοντας το bias b και χρησιμοποιώντας ότι $w^T x_k + b = 0$ (αφού x_k ανήκει στο υπερεπίπεδο):

$$y(x) = w^T x + b = r||w|| \Rightarrow \boxed{r = \frac{y(x)}{||w||}}$$

Θέμα 1.2

Τι είναι τα διανύσματα στήριξης (support vectors) και ποια η σημασία τους;

Λύση

(Πηγή: Lecture 6, slides 52--53)
Τα **Support Vectors** είναι τα σημεία εκπαίδευσης που βρίσκονται ακριβώς πάνω στα περιθώρια (margins) του ταξινομητή SVM, δηλαδή ικανοποιούν $y_i(w^T x_i + b) = 1$.
Σημασία:

- Ορίζουν πλήρως το όριο απόφασης --- η αφαίρεση μη-SVs δεν αλλάζει τον ταξινομητή
- Καθορίζουν το πλάτος του margin: margin $= \frac{2}{||w||}$
- Η λύση εξαρτάται **μόνο** από αυτά (sparse solution)
- Είναι τα "δύσκολα" σημεία που βρίσκονται κοντά στο σύνορο των κλάσεων

Θέμα 1.3

Πώς αντιμετωπίζουμε το πρόβλημα των μη γραμμικά διαχωρίσιμων κλάσεων με τη χρήση των SVM;

Λύση

(Πηγή: Lecture 6, slides 78--79)
Χρησιμοποιούμε το **Kernel Trick**: αντί να υπολογίζουμε ρητά τον μετασχηματισμό $\phi(x)$ σε χώρο υψηλότερης διάστασης, χρησιμοποιούμε μια συνάρτηση kernel $K(x, x') = \phi(x)^T \phi(x')$ που υπολογίζει απευθείας το εσωτερικό γινόμενο.
Δημοφιλή Kernels:

- **Πολυωνυμικό:** $K(x, x') = (x^T x' + c)^d$
- **RBF (Gaussian):** $K(x, x') = \exp\left(-\frac{||x-x'||^2}{2\sigma^2}\right)$

Στο νέο χώρο τα δεδομένα είναι συχνά γραμμικά διαχωρίσιμα.

Θέμα 1.4

Περιγράψτε τη διαδικασία εφαρμογής της ταξινόμησης κατά Bayes (Risk minimization).

Λύση

(Πηγή: Lecture 3, slides 22--24)
Βήματα:

1. Ορίζουμε τη **συνάρτηση απώλειας** $\lambda(\alpha_i|\omega_j)$: το κόστος της απόφασης α_i όταν η πραγματική κλάση είναι ω_j
2. Υπολογίζουμε το **υπό συνθήκη ρίσκο**: $R(\alpha_i|x) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|x)$
3. Επιλέγουμε την ενέργεια που **ελαχιστοποιεί** το ρίσκο: $\alpha^* = \arg\min_i R(\alpha_i|x)$

Για 0-1 loss ($\lambda = 0$ αν σωστό, $\lambda = 1$ αν λάθος), ο κανόνας Bayes επιλέγει την κλάση με τη μέγιστη a posteriori πιθανότητα (**MAP**).

Θέμα 1.5

Περιγράψτε τη βασική ιδέα EM για Mixture of Gaussians.

Λύση

(Πηγή: Lecture 9-10, slides 31--36)
Ο αλγόριθμος **Expectation-Maximization (EM)** χρησιμοποιείται όταν έχουμε λανθάνουσες (hidden) μεταβλητές. Για Gaussian Mixture Models (GMM):
E-Step: Υπολογίζουμε τις "responsibilities" --- την πιθανότητα κάθε σημείο να ανήκει σε κάθε component:

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}$$

M-Step: Ενημερώνουμε τις παραμέτρους με βάση τις responsibilities:

- $\mu_k^{\text{new}} = \frac{\sum_n \gamma_{nk} x_n}{\sum_n \gamma_{nk}}$ (σταθμισμένος μέσος)
- $\Sigma_k^{\text{new}} =$ σταθμισμένη συνδιακύμανση
- $\pi_k^{\text{new}} = \frac{1}{N} \sum_n \gamma_{nk}$ (mixing proportion)

Επαναλαμβάνουμε μέχρι σύγκλιση. Ο EM **εγγυάται** αύξηση της log-likelihood σε κάθε βήμα.

Θέμα 2.1

Η σύγκλιση του EM εξαρτάται από το learning rate.

Λύση

(Πηγή: Lecture 9-10, slides 31--36)
Απ: Λάθος.
Ο EM **δεν** χρησιμοποιεί learning rate. Είναι αλγόριθμος **coordinate ascent** που εναλλάσσεται μεταξύ E-step και M-step με κλειστές εκφράσεις. Η σύγκλιση εξαρτάται από την αρχικοποίηση, όχι από learning rate.

Θέμα 2.2

Το όριο απόφασης Gaussian classifiers είναι πάντα γραμμικό.

Λύση

(Πηγή: Lecture 3, slides 41--42)
Απ: Λάθος.
Το όριο απόφασης είναι γραμμικό **μόνο** όταν οι πίνακες συνδιακύμανσης είναι ίσοι ($\Sigma_1 = \Sigma_2$). Διαφορετικά, ο διακριτικός λόγος περιέχει τετραγωνικούς όρους $x^T(\Sigma_1^{-1} - \Sigma_2^{-1})x$, οπότε το όριο είναι **τετραγωνικό** (ελλειπτικές/υπερβολικές καμπύλες).

Θέμα 2.3

Η $k(x, y) = x_1^2 y_1^2 + x_2^2 y_2^2$ είναι kernel.

Λύση

(Πηγή: Lecture 6, slide 76)
Απ: Σωστό.
Ναι, γιατί μπορεί να γραφτεί ως εσωτερικό γινόμενο:

$$k(x, y) = \Phi(x)^T \Phi(y) \text{ όπου } \Phi(x) = [x_1^2, x_2^2]^T$$

Εναλλακτικά, ο πίνακας Gram είναι θετικά ημι-ορισμένος (αφού είναι άθροισμα θετικών kernels).

Θέμα 2.4

Η διάσταση του feature space ενός kernel μπορεί να είναι άπειρη.

Λύση

(Πηγή: Lecture 6, slide 78)
Απ: Σωστό.
Το **RBF kernel** $K(x, x') = e^{-\gamma||x-x'||^2}$ αντιστοιχεί σε άπειρης διάστασης feature space. Η Taylor expansion του e^x δίνει άπειρους όρους, καθένας από τους οποίους αντιστοιχεί σε ένα feature.

Θέμα 2.5

Τα PCA components είναι ιδιοδιανύσματα του data matrix.

Λύση

(Πηγή: Lecture 9-10, slides 52--54)
Απ: Λάθος.
Τα principal components είναι τα **ιδιοδιανύσματα του πίνακα συνδιακύμανσης** $\Sigma = \frac{1}{N}X^TX$ (αν τα δεδομένα είναι κεντραρισμένα), **όχι** του data matrix X.
Η σχέση μέσω SVD: αν $X = U\Sigma V^T$, τα PCs είναι οι στήλες του V.

Θέμα 3.1

Θέλετε να εκπαιδεύσετε ένα Νευρωνικό Δίκτυο (NN) για αυτόνομη οδήγηση. Τα δεδομένα εκπαίδευσης αποτελούνται από grayscale εικόνες 64×64 εικονοστοιχείων (pixels). Οι ετικέτες (labels) που συνοδεύουν τα δεδομένα σας είναι η γωνία του τιμονιού σε μοίρες και η ταχύτητα σε χλμ/ώρα (διάνυσμα y). Το NN αποτελείται από ένα στρώμα εισόδου (input layer), από ένα κρυφό στρώμα (hidden layer) 2048 νευρώνων, κι ένα στρώμα εξόδου (output layer).

Λύση

(Πηγή: Lecture 7, slides 1--2)
Κάθε στρώμα έχει: (αριθμός εισόδων + 1 bias) $\times$ αριθμός νευρώνων

- **Layer 1 (Input → Hidden):** $(4096 + 1) \times 2048 = 8,390,656$ παράμετροι
- **Layer 2 (Hidden → Output):** $(2048 + 1) \times 2 = 4,098$ παράμετροι

Σύνολο: $8,390,656 + 4,098 = \mathbf{8,394,754}$ παράμετροι

Θέμα 3.2

Αν η έξοδος του NN είναι z, η έξοδος του κρυφού στρώματος είναι h και ο πίνακας βαρών μεταξύ κρυφού στρώματος και στρώματος εξόδου είναι W, βρείτε την παράγωγο $\frac{\partial J}{\partial W_{ij}}$ αν $J = \frac{1}{2}\|z - y\|^2$ είναι η συνάρτηση κόστους. (2 μονάδες)

Λύση

(Πηγή: Lecture 7, slides 55--59)
Για MSE loss: $J = \frac{1}{2}||z - y||^2$ όπου $z = Wh$ (έξοδος δικτύου)
Εφαρμόζοντας chain rule:

$$\frac{\partial J}{\partial W_{ij}} = \frac{\partial J}{\partial z_i} \cdot \frac{\partial z_i}{\partial W_{ij}} = (z_i - y_i) \cdot h_j$$

Σε μορφή πίνακα: $\nabla_W J = (z - y)h^T$ (outer product)

3 Ιούνιος 2023

Πτυχιακή Εξεταστική - 27 Ιουνίου 2023

*Στις παρακάτω προτάσεις κυκλώστε **Σ** αν η πρόταση είναι Σωστή και **Λ** αν η πρόταση είναι Λανθασμένη. Κάθε σωστή απάντηση λαμβάνει 1 μονάδα και κάθε λάθος απάντηση λαμβάνει -0.25.*

Θέμα 1.1

Η αντικειμενική συνάρτηση (objective function) ενός τεχνητού νευρωνικού δικτύου (ΤΝΔ) μπορεί έχει τοπικά ακρότατα.

Λύση

(Πηγή: Lecture 7, slides 57--58) **Απ: Σωστό.** *Εξήγηση:* Η συνάρτηση κόστους των νευρωνικών δικτύων είναι **μη-κυρτή (non-convex)**. Λόγω των συμμετριών στα βάρη (permutation symmetry) και των μη-γραμμικοτήτων, δημιουργούνται πολλαπλά τοπικά ελάχιστα (local minima) και σαγματικά σημεία (saddle points). Δεν εγγυάται η εύρεση του ολικού ελαχίστου.

Θέμα 1.2

Η αντικειμενική συνάρτηση (objective function) ενός SVM μπορεί έχει τοπικά ακρότατα.

Λύση

(Πηγή: Lecture 6, slides 44--47) **Απ: Λάθος.** *Εξήγηση:* Το πρόβλημα βελτιστοποίησης των SVM είναι πρόβλημα **κυρτού τετραγωνικού προγραμματισμού** (Convex Quadratic Programming). Η αντικειμενική συνάρτηση $J(w) = \frac{1}{2}||w||^2 + C\sum \xi_i$ είναι κυρτή και οι περιορισμοί είναι γραμμικοί. Συνεπώς, κάθε τοπικό ελάχιστο είναι και **ολικό (global minimum)**.

Θέμα 1.3

Ο ΕΜ αλγόριθμος ελαχιστοποιεί την συνάρτηση πιθανοφάνειας.

Λύση

(Πηγή: Lecture 9-10, slides 31--36) **Απ: Λάθος.** *Εξήγηση:* Ο αλγόριθμος Expectation-Maximization (EM) στοχεύει στη **μεγιστοποίηση** της πιθανοφάνειας (Likelihood Maximization). Σε κάθε επανάληψη, υπολογίζει ένα 'lower bound' της log-likelihood και το μεγιστοποιεί, εγγυώμενος ότι η πιθανοφάνεια δεν θα μειωθεί.

Θέμα 1.4

Η αντικειμενική συνάρτηση ενός ΤΝΔ που εκπαιδεύεται με gradient descent σταματά να μειώνεται μετά από πεπερασμένο αριθμό επαναλήψεων εκπαίδευσης.

Λύση

(Πηγή: Lecture 7, slides 59--61) **Απ: Λάθος.** *Εξήγηση:* Η σύγκλιση της Gradient Descent είναι θεωρητικά **ασυμπτωτική**. Καθώς πλησιάζουμε το ελάχιστο, η κλίση (gradient) τείνει στο μηδέν, κάνοντας τα βήματα ενημέρωσης ολοένα και μικρότερα. Θεωρητικά, μπορεί να απαιτηθούν άπειρα βήματα για να φτάσουμε *ακριβώς* στο ελάχιστο (αν και πρακτικά σταματάμε όταν η αλλαγή είναι μικρότερη από ένα κατώφλι ανοχής).

Θέμα 1.5

Ένα γραμμικό SVM εκπαιδεύεται με το παρακάτω σύνολο δεδομένων, όπου x_i είναι τα δεδομένα και d η επιθυμητή ταξινόμηση κάθε διανύσματος.

x_1	x_2	d	x_1	x_2	d
-2	7	-1	1	0	+1
2	6	-1	6	6	+1
4	12	-1	6	10	+1

Η νέα είσοδος (4,3) στο παραπάνω γραμμικό SVM θα ταξινομηθεί ως κλάση -1.

Λύση

(Πηγή: Lecture 6, slides 49--55) **Απ: Λάθος.** (Ταξινομείται ως +1)
Αναλυτική Επίλυση -- Εύρεση Διαχωριστικής Ευθείας:
Βήμα 1: Εντοπισμός Support Vectors. Τα SVs είναι τα σημεία πιο κοντά στο σύνορο:

- Κλάση +1: $(1,0)$ και $(6,10)$
- Κλάση -1: $(2,6)$

Βήμα 2: Υπολογισμός κλίσης. Η ευθεία που συνδέει τα SVs της κλάσης +1:

$$\text{slope} = \frac{10-0}{6-1} = \frac{10}{5} = 2$$

Βήμα 3: Εύρεση margin boundaries. Μορφή: $x_2 = 2x_1 + b$
Margin +1 (περνάει από $(1,0)$): $0 = 2(1) + b_+ \Rightarrow b_+ = -2 \Rightarrow x_2 = 2x_1 - 2$
Margin -1 (περνάει από $(2,6)$): $6 = 2(2) + b_- \Rightarrow b_- = 2 \Rightarrow x_2 = 2x_1 + 2$
Βήμα 4: Decision Boundary (μέση ευθεία).

$$b = \frac{-2+2}{2} = 0 \Rightarrow \boxed{x_2 = 2x_1} \quad \text{ή} \quad 2x_1 - x_2 = 0$$

Βήμα 5: Ταξινόμηση. Κανόνας: $y(x) = 2x_1 - x_2$. Αν $y > 0 \Rightarrow +1$, αν $y < 0 \Rightarrow -1$.

Σημείο	$y = 2x_1 - x_2$	Κλάση
$(4,3)$	$2(4) - 3 = 5 > 0$	**+1**
$(0,4)$	$2(0) - 4 = -4 < 0$	**-1**
$(3,7)$	$2(3) - 7 = -1 < 0$	**-1**

Θέμα 1.6

Η νέα είσοδος (0,4) στο παραπάνω γραμμικό SVM θα ταξινομηθεί ως κλάση -1.

Λύση

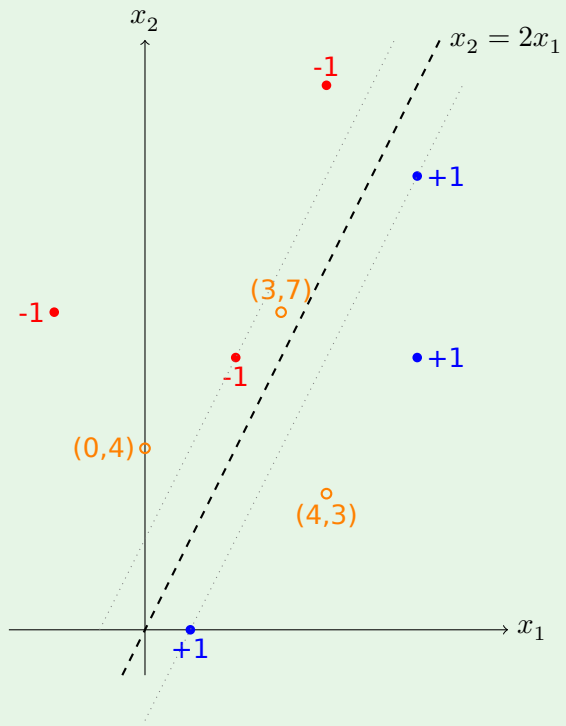
(Πηγή: Lecture 6, slides 49--55) **Απ: Σωστό.**

Θέμα 1.7

Η νέα είσοδος (3,7) στο παραπάνω γραμμικό SVM θα ταξινομηθεί ως κλάση +1.

Λύση

(Πηγή: Lecture 6, slides 49--55) **Απ: Λάθος.** (Ταξινομείται ως -1)



Θέμα 1.8

Η αντικειμενική συνάρτηση του k-Means αλγορίθμου σταματά να μειώνεται μετά από πεπερασμένο αριθμό επαναλήψεων εκπαίδευσης.

Λύση

(Πηγή: Lecture 9-10, slide 10) **Απ: Σωστό.** *Εξήγηση:* Ο αλγόριθμος K-Means εγγυάται σύγκλιση σε πεπερασμένο αριθμό βημάτων επειδή:

1. Υπάρχει πεπερασμένος αριθμός τρόπων ανάθεσης N σημείων σε K clusters (K^N).
2. Σε κάθε βήμα, η συνάρτηση κόστους (sum of squared errors) μειώνεται μονοτονικά ή παραμένει σταθερή.
3. Δεν υπάρχουν κύκλοι (loops) στις αναθέσεις.

Θέμα 1.9

Η βέλτιστη τιμή της αντικειμενικής συνάρτησης ενός EM αλγορίθμου (για την εκτίμηση μιας πυκνότητας πιθανότητας) χρησιμοποιώντας $p+1$ γκαουσιανές συναρτήσεις δεν μπορεί να είναι υψηλότερη από αυτήν της αντικειμενικής συνάρτησης για την εκτίμηση της ίδιας πυκνότητας πιθανότητας χρησιμοποιώντας p γκαουσιανές συναρτήσεις.

Λύση

(Πηγή: Lecture 9-10, slides 31--36) **Απ: Λάθος.** *Εξήγηση:* Όταν αυξάνουμε την πολυπλοκότητα του μοντέλου (π.χ. από p σε $p+1$ Gaussians), το νέο μοντέλο περιέχει το παλιό ως ειδική περίπτωση (nested models). Συνεπώς, η μέγιστη πιθανοφάνεια (likelihood) του $p+1$ μοντέλου θα είναι **τουλάχιστον ίση ή μεγαλύτερη** από αυτή του p μοντέλου. (Προσοχή: Αυτό μπορεί να οδηγήσει σε overfitting, αλλά η likelihood στο training set αυξάνεται).

Θέμα 1.10

Στα ΤΝΔ οι συναρτήσεις ενεργοποίησης (activation functions) βοηθούν ώστε να σχηματιστούν μη γραμμικά όρια απόφασης (decision boundaries).

Λύση

(Πηγή: Lecture 7, slides 28--30) **Απ: Σωστό.** *Εξήγηση:* Μια σύνθεση γραμμικών συναρτήσεων είναι απλά μια άλλη γραμμική συνάρτηση ($W_2(W_1x) = W'x$). Χωρίς μη-γραμμικές συναρτήσεις ενεργοποίησης (όπως ReLU, Sigmoid, Tanh), ένα πολυεπίπεδο νευρωνικό δίκτυο καταρρέει σε ένα απλό γραμμικό μοντέλο (Perceptron), χάνοντας την ικανότητα να προσεγγίζει πολύπλοκες, μη-γραμμικές συναρτήσεις (Universal Approximation Theorem).

4 Σεπτέμβριος 2023

Εξεταστική Σεπτεμβρίου - 14 Σεπτεμβρίου 2023

Θέμα 1.1

Υποθέστε ότι έχετε συμπτώματα COVID-19 και αποφασίζετε να κάνετε ένα COVID-test. Στη συσκευασία του test αναγράφεται ότι η πιθανότητα το αποτέλεσμα του test να είναι θετικό δεδομένου ότι άτομο έχει όντως μολυνθεί από τον ιό είναι 0.875 και ότι η πιθανότητα το αποτέλεσμα του test να είναι αρνητικό δεδομένου ότι το άτομο **δεν** έχει μολυνθεί από τον ιό είναι 0.975. Το test που κάνατε βγαίνει θετικό. Υπολογίστε την πιθανότητα να έχετε όντως μολυνθεί από τον ιό με βάση το αποτέλεσμα του test αν η συχνότητα μόλυνσης από τον ιό στην περιοχή που κατοικείτε είναι 1 στους 10.

Λύση

(Πηγή: Lecture 3, slides 22--24) Ορίζουμε τα ενδεχόμενα:

- ω_1: Δεν έχω μολυνθεί (Υγιής)
- ω_2: Έχω μολυνθεί (Ασθενής)
- $x = 1$: Test Θετικό
- $x = 0$: Test Αρνητικό

Δεδομένα από την εκφώνηση:

- $P(x = 1|\omega_2) = 0.875$ (Sensitivity)
- $P(x = 0|\omega_1) = 0.975 \Rightarrow P(x = 1|\omega_1) = 0.025$ (False Positive Rate)
- $P(\omega_2) = 0.1$, $P(\omega_1) = 0.9$

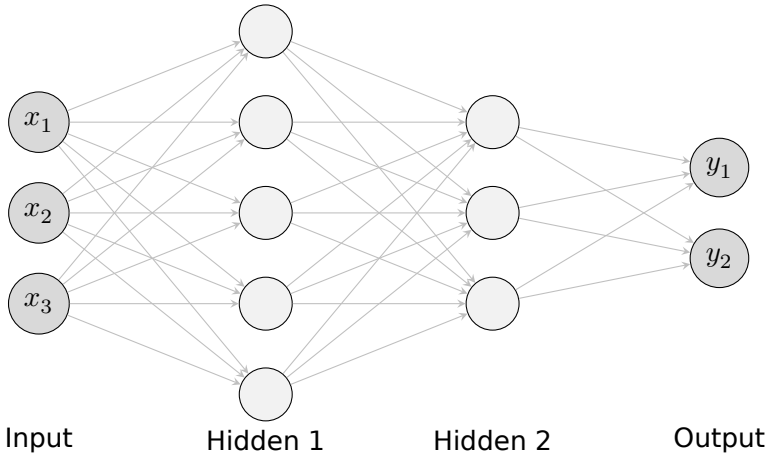
Από το **θεώρημα Bayes**:

$$P(\omega_2|x = 1) = \frac{P(x = 1|\omega_2)P(\omega_2)}{P(x = 1)}$$

Όπου: $P(x = 1) = P(x = 1|\omega_1)P(\omega_1) + P(x = 1|\omega_2)P(\omega_2) = (0.025 \cdot 0.9) + (0.875 \cdot 0.1) = 0.11$

$$P(\omega_2|x = 1) = \frac{0.875 \cdot 0.1}{0.11} = \frac{0.0875}{0.11} \approx \mathbf{0.795} \text{ ή } \mathbf{79.5}\%$$

Θεωρήστε το παρακάτω νευρωνικό δίκτυο (αρχιτεκτονική $3 \to 5 \to 3 \to 2$):



Θέμα 2.1

Βρείτε τις διαστάσεις των πινάκων βαρών, $W^{[1]}, W^{[2]}, W^{[3]}$ και των διανυσμάτων πόλωσης (bias) $b^{[1]}, b^{[2]}, b^{[3]}$. Δώστε τις σε μορφή (αριθμός γραμμών) $\times$ (αριθμός στηλών). (1 μονάδα)

Λύση

(Πηγή: Lecture 7, slides 1--2)
Γενικός Κανόνας: Αν το επίπεδο $l-1$ έχει n_{in} νευρώνες και το επίπεδο l έχει n_{out} νευρώνες:

- Ο πίνακας βαρών $W^{[l]}$ έχει διαστάσεις $n_{out} \times n_{in}$.
- Το διάνυσμα bias $b^{[l]}$ έχει διαστάσεις $n_{out} \times 1$.

Εφαρμογή:

- **Layer 1 ($3 \to 5$):** $n_{in} = 3, n_{out} = 5 \Rightarrow W^{[1]} \in \mathbb{R}^{5\times 3}, b^{[1]} \in \mathbb{R}^{5\times 1}$
- **Layer 2 ($5 \to 3$):** $n_{in} = 5, n_{out} = 3 \Rightarrow W^{[2]} \in \mathbb{R}^{3\times 5}, b^{[2]} \in \mathbb{R}^{3\times 1}$
- **Layer 3 ($3 \to 2$):** $n_{in} = 3, n_{out} = 2 \Rightarrow W^{[3]} \in \mathbb{R}^{2\times 3}, b^{[3]} \in \mathbb{R}^{2\times 1}$

Θέμα 2.2

Δώστε τις εξισώσεις της 'προς τα εμπρός διέλευσης' (forward pass) του παραπάνω δικτύου αν κάθε νευρώνας των κρυφών στρωμάτων (hidden layers) και του στρώματος εξόδου (output layer) έχει συνάρτηση ενεργοποίησης (activation function) $g(.)$, έτσι ώστε $a^{[l]} = g(z^{[l]}), i = 1, 2, 3$. (1 μονάδα)

Λύση

(Πηγή: Lecture 7, slides 1--2)
Η διαδικασία περιλαμβάνει δύο βήματα για κάθε στρώμα: τον γραμμικό συνδυασμό (linear transform) και την μη-γραμμική ενεργοποίηση (activation).
Layer 1 (Input $\to$ Hidden 1):

$$z^{[1]} = W^{[1]}x + b^{[1]} \quad \text{(Linear step)}$$

$$a^{[1]} = g(z^{[1]}) \quad \text{(Activation, π.χ. ReLU/Sigmoid)}$$

Layer 2 (Hidden 1 $\to$ Hidden 2):

$$z^{[2]} = W^{[2]}a^{[1]} + b^{[2]}$$

$$a^{[2]} = g(z^{[2]})$$

Layer 3 (Hidden 2 $\to$ Output):

$$z^{[3]} = W^{[3]}a^{[2]} + b^{[3]}$$

$$y = a^{[3]} = g(z^{[3]}) \quad \text{(Τελική πρόβλεψη)}$$

όπου $g(\cdot)$ είναι η συνάρτηση ενεργοποίησης (activation function) που εφαρμόζεται κατά στοιχείο (element-wise).

Θέμα 2.3

Υπολογίστε τον αριθμό των παραμέτρων εκπαίδευσης. (1 μονάδα)

Λύση

(Πηγή: Lecture 7, slides 1--2)
Ο συνολικός αριθμός παραμέτρων είναι το άθροισμα των βαρών και των biases σε κάθε επίπεδο. Για κάθε επίπεδο: $Params = (n_{in} \times n_{out}) + n_{out} = (n_{in} + 1) \times n_{out}$.

- **Layer 1 ($3 \to 5$):** $(3 \times 5) + 5 = 15 + 5 = 20$
- **Layer 2 ($5 \to 3$):** $(5 \times 3) + 3 = 15 + 3 = 18$
- **Layer 3 ($3 \to 2$):** $(3 \times 2) + 2 = 6 + 2 = 8$

Σύνολο: $20 + 18 + 8 =$ **46** παράμετροι.

Θέμα 3.1

Εφαρμόστε για μία επανάληψη τη μέθοδο βελτιστοποίησης gradient descent αν δίνονται το διάνυσμα χαρακτηριστικών $X = [1, 1, 1]^T$, το διάνυσμα των παραμέτρων $W = [1, 1, 1]^T$, η τιμή πόλωσης (bias) $b = 1/2$ και η τιμή στόχου $y = 1.5$ χρησιμοποιώντας σαν συνάρτηση βελτιστοποίησης την συνάρτηση ελαχίστων τετραγώνων. Υποθέστε ρυθμό μάθησης (learning rate) $\eta = 0.1$.

Λύση

(Πηγή: Lecture 7, slides 59--61)

Πριν την επανάληψη:

$$\hat{y} = W^T X + b = (1 + 1 + 1) + 0.5 = 3.5$$

$$E_{old} = \frac{1}{2}(3.5 - 1.5)^2 = \frac{1}{2} \cdot 4 = \mathbf{2}$$

Gradient Descent Update: Η παράγωγος του σφάλματος ως προς τα βάρη υπολογίζεται με τον κανόνα της αλυσίδας (Chain Rule):

$$\frac{\partial E}{\partial W} = \frac{\partial E}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial W} = (\hat{y} - y) \cdot X$$

Αντικαθιστώντας τις τιμές:

$$\frac{\partial E}{\partial W} = (3.5 - 1.5) \cdot [1, 1, 1]^T = 2 \cdot [1, 1, 1]^T = [2, 2, 2]^T$$

Ενημέρωση βαρών ($W_{new} = W - \eta \frac{\partial E}{\partial W}$):

$$W_{new} = [1, 1, 1]^T - 0.1 \cdot [2, 2, 2]^T = [1, 1, 1]^T - [0.2, 0.2, 0.2]^T = [0.8, 0.8, 0.8]^T$$

Ενημέρωση bias ($b_{new} = b - \eta \frac{\partial E}{\partial b}$):

$$b_{new} = b - \eta(\hat{y} - y) = 0.5 - 0.1 \cdot 2 = 0.3$$

Μετά την επανάληψη:

$$\hat{y}_{new} = (0.8 + 0.8 + 0.8) + 0.3 = 2.7$$

$$E_{new} = \frac{1}{2}(2.7 - 1.5)^2 = \frac{1}{2} \cdot 1.44 = \mathbf{0.72}$$

Παρατήρηση: Το κόστος μειώθηκε από 2 σε 0.72, άρα ο αλγόριθμος λειτουργεί σωστά.

Θέμα 3.2

Γιατί επιλέγουμε συνήθως μικρές τιμές για τον ρυθμό μάθησης;

Λύση

(Πηγή: Lecture 7, slide 61) **Learning Rate (η):**

- **Μικρό η:** Εξασφαλίζει σταθερή σύγκλιση/μείωση του σφάλματος, αλλά μπορεί να είναι αργή.
- **Μεγάλο η:** Μπορεί να οδηγήσει σε ταλαντώσεις (oscillations) γύρω από το ελάχιστο ή ακόμη και σε απόκλιση (overshooting), αυξάνοντας το σφάλμα αντί να το μειώνει.

Συνήθως επιλέγουμε μικρές τιμές (π.χ. $0.1, 0.01$) για ασφάλεια.

5 Φεβρουάριος 2024

Εξεταστική Φεβρουαρίου - 14 Φεβρουαρίου 2024

*Στις παρακάτω προτάσεις κυκλώστε **όλες** τις πιθανές **σωστές** απαντήσεις. Κάθε σωστή απάντηση λαμβάνει 0.5 μονάδα και κάθε λάθος απάντηση λαμβάνει -0.25.*

Θέμα 1.1

Τα Νευρωνικά Δίκτυα:

A. Βελτιστοποιούν μια κυρτή (convex) συνάρτηση κόστους

B. Δίνουν πάντα έξοδο 0 ή 1

Γ. Μπορούν να χρησιμοποιηθούν τόσο για παλινδρόμηση όσο και για ταξινόμηση

Δ. Μπορούν να χρησιμοποιηθούν σε «επίτρεπη» ταξινομητών

Λύση

(Πηγή: Lecture 7, slides 57--58)
Απ: Γ, Δ.
Εξήγηση: Η συνάρτηση κόστους NN δεν είναι κυρτή (Α λάθος). Η έξοδος εξαρτάται από την activation function (Β λάθος). Τα NN χρησιμοποιούνται και για regression και classification (Γ σωστό) και μπορούν να συνδυαστούν σε ensemble (Δ σωστό).

Θέμα 1.2

Ποιο/ποια από τα παρακάτω είναι σωστό/σωστά για τα Πιθανοκρατικά Αναγεννητικά Μοντέλα (Probabilistic Generative Models, PGM):

A. Μοντελοποιούν την από κοινού συνάρτηση κατανομής $P(\omega_j, x)$

B. Ο Perceptron είναι PGM

Γ. Μπορούν να χρησιμοποιηθούν για ταξινόμηση

Δ. Η Γραμμική Ανάλυση Διακριτοποίησης (Linear Discriminant Analysis) είναι PGM

Λύση

(Πηγή: Lecture 5, slides 70--72)
Απ: Α, Γ, Δ.
Εξήγηση: PGMs μοντελοποιούν την joint distribution (Α σωστό). Ο Perceptron είναι discriminative (Β λάθος). Χρησιμοποιούνται για classification μέσω Bayes (Γ σωστό). Το LDA είναι generative (Δ σωστό).

Θέμα 1.3

Ποιες από τις παρακάτω μεθόδους μπορούν να επιτύχουν μηδενικό σφάλμα ταξινόμησης σε κάθε σύνολο εκπαίδευσης που είναι γραμμικά διαχωρίσιμο;

A. Δέντρο απόφασης

B. 15-NN

Γ. Hard-Margin SVM

Δ. Perceptron

Λύση

(Πηγή: Lecture 6, slides 44--47; Lecture 5, slides 26--31)
Απ: Γ, Δ.
Εξήγηση: Δέντρα δεν εγγυώνται γραμμικό διαχωρισμό. 15-NN μπορεί να κάνει λάθη λόγω voting. Hard-margin SVM βρίσκει τέλειο γραμμικό διαχωρισμό (Γ σωστό). Ο Perceptron συγκλίνει για γραμμικά διαχωρίσιμα δεδομένα (Δ σωστό).

Θέμα 1.4

Το kernel trick:

A. Μπορεί να εφαρμοστεί σε κάθε αλγόριθμο ταξινόμησης

B. Χρησιμοποιείται συνήθως για μείωση διαστάσεων

Γ. Εκμεταλλεύεται το γεγονός ότι σε πολλούς αλγορίθμους τα βάρη μπορούν να γραφούν σαν γραμμικός συνδυασμός των δεδομένων εισόδου

Δ. Δίνει πάντα τιμές μεταξύ 0 και 1

Λύση

(Πηγή: Lecture 6, slides 78--79)
Απ: Γ.
Εξήγηση: Το kernel trick βασίζεται στο Representer Theorem (Γ σωστό) --- τα βάρη γράφονται ως $w = \sum_i \alpha_i \phi(x_i)$.

Θέμα 1.5

Στα Νευρωνικά Δίκτυα, οι συναρτήσεις ενεργοποίησης ReLU και tanh:

A. Βοηθούν στην επιτάχυνση της διαδικασίας εκπαίδευσης

B. Οδηγούν στην εύρεση μη γραμμικών ορίων απόφασης

Γ. Εφαρμόζονται μόνο στους νευρώνες του επιπέδου εξόδου (output layer)

Δ. Δίνουν πάντα τιμές μεταξύ 0 και 1

Λύση

(Πηγή: Lecture 7, slides 28--30)
Απ: Α, Β.
Εξήγηση: ReLU επιταχύνει training (Α σωστό). Μη-γραμμικές activations επιτρέπουν μη-γραμμικά όρια (Β σωστό). Εφαρμόζονται σε hidden layers (Γ λάθος). ReLU δίνει $[0, \infty)$, tanh δίνει $[-1, 1]$ (Δ λάθος).

Θέμα 1.6

Με βάση το Bias-Variance Trade-off ένας 1-NN ταξινομητής έχει _____ σε σύγκριση με έναν 3-NN ταξινομητή:

A. Μεγαλύτερο variance

B. Μεγαλύτερο bias

Γ. Μικρότερο variance

Δ. Μικρότερο bias

Λύση

(Πηγή: Lecture 2, slides 32--37)
Απ: Α, Δ.
Εξήγηση: 1-NN: υψηλό variance (ευαίσθητο σε θόρυβο), χαμηλό bias (ακριβές στο training). 3-NN: χαμηλότερο variance (averaging), υψηλότερο bias.

Θέμα 1.7

Ποια/α από τα παρακάτω είναι αληθές/ή για το Bootstrapping:

A. Με το Bootstrapping επιλέγουμε τυχαία δείγματα από τα δεδομένα με επανατοποθέτηση

B. Σκοπός του είναι να οδηγήσει σε μείωση του bias

Γ. Είναι μη αποτελεσματικό με Logistic Regression ταξινομητές

Λύση

(Πηγή: Lecture 8, slides 10--12)
Απ: Α.
Εξήγηση: Bootstrapping = sampling with replacement (Α σωστό). Μειώνει variance, όχι bias (Β λάθος). Λειτουργεί με όλους τους ταξινομητές (Γ λάθος).

Θέμα 1.8

Ποια/α από τα παρακάτω είναι αληθές/ή για το PCA:

A. Προσθέτοντας ένα '1' στο τέλος κάθε διανύσματος δεδομένων δεν αλλάζουν τα αποτελέσματα του PCA

B. Οι πρωταρχικές συνιστώσες είναι ιδιοδιανύσματα του πίνακα συμμεταβλητότητας των δεδομένων

Γ. Οι πρωταρχικές συνιστώσες είναι ιδιοδιανύσματα του πίνακα των δεδομένων

Λύση

(Πηγή: Lecture 9-10, slides 52--54)
Απ: Α, Β.
Εξήγηση: Σταθερό feature δεν αλλάζει variance (Α σωστό). PCs = eigenvectors of covariance matrix (Β σωστό, Γ λάθος).

Θέμα 1.9

Ποια/ες από τις παρακάτω προσεγγίσεις μπορούν να βοηθήσουν ώστε να αποφευχθεί το overfitting στα Δέντρα Απόφασης:

A. Κλάδεμα (Pruning)

B. Ορισμός μέγιστου βάθους του δέντρου

Γ. Ορισμός ελάχιστου αριθμού από δείγματα στα φύλλα

Λύση

(Πηγή: Lecture 8, slides 71--72)
Απ: Α, Β, Γ.
Εξήγηση: Όλες οι τεχνικές περιορίζουν την πολυπλοκότητα του δέντρου και αποφεύγουν overfitting.

Θέμα 1.10

Έστω ότι έχετε σετ δεδομένων που αποτελείται από n δείγματα διαφορετικών κλάσεων. Κάθε κλάση έχει διαφορετική κατανομή. Δεν γνωρίζουμε τις ετικέτες των κλάσεων και έτσι χρησιμοποιείτε μια τεχνική ομαδοποίησης (clustering), την k-Means. Σε ποιες από τις παρακάτω περιπτώσεις θα επηρεαζόταν αρνητικά η αποδοτικότητα της μεθόδου:

Α. Κάποιες από τις κλάσεις δεν ακολουθούν κανονική κατανομή

Β. Η διασπορά των κατανομών είναι μικρή σε όλες τις κατευθύνσεις

Γ. Κάθε κλάση έχει την ίδια μέση τιμή

Δ. Επιλέγουμε $k = n$

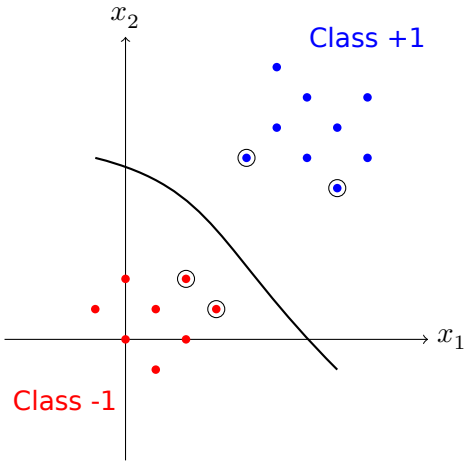
Λύση

(Πηγή: Lecture 9-10, slides 12--16)
Απ: Α, Γ.
Εξήγηση: k-Means υποθέτει spherical clusters (Α επηρεάζει). Ίδια μέση = overlapping clusters (Γ επηρεάζει). Μικρή variance βοηθά (Β δεν επηρεάζει). $k = n$ δίνει perfect fit αλλά δεν είναι χρήσιμο.

6 Ιούνιος 2024

Εξεταστική Ιουνίου - 2 Ιουλίου 2024

Στο σχήμα φαίνεται η λύση σε ένα πρόβλημα δυαδικής ταξινόμησης με χρήση ταξινομητή SVM με RBF kernel: $K(\mathbf{x}, \mathbf{x}') = \exp\left[-\frac{||\mathbf{x}-\mathbf{x}'||^2}{2}\right]$ (βλ. Lecture 6: SVM & Kernels).



Η απόφαση ταξινόμησης για ένα διάνυσμα **x** λαμβάνεται με βάση το πρόσημο της παράστασης:

$$\begin{aligned} y(\mathbf{x}) &= \mathbf{w}^T\phi(\mathbf{x}) + w_0 \\ &= \sum_{n\in S} \mu_n t_n \phi(\mathbf{x}_n)^T\phi(\mathbf{x}) + w_0 \\ &= \sum_{n\in S} \mu_n t_n K(\mathbf{x}_n, \mathbf{x}) + w_0 \end{aligned}$$

Θέμα 1.1

Πώς επηρεάζει την τιμή της παραπάνω παράστασης και κατ' επέκταση την απόφασή μας, ο έλεγχος για ένα διάνυσμα, $\mathbf{x}_f$, το οποίο βρίσκεται πολύ μακριά από οποιοδήποτε διάνυσμα εκπαίδευσης (θεωρήστε ότι οι αποστάσεις υπολογίζονται στο αρχικό χώρο των δεδομένων και όχι στον χώρο προβολής); (2 μονάδες)

Λύση

(Πηγή: Lecture 6, slides 78--80)

Αναλυτική Εξήγηση:

Το RBF kernel ορίζεται ως: $K(\mathbf{x}, \mathbf{x}') = \exp\left[-\frac{||\mathbf{x}-\mathbf{x}'||^2}{2\sigma^2}\right]$

Βήμα 1: Ανάλυση της συνάρτησης απόφασης.

$$y(\mathbf{x}) = \sum_{n\in S} \mu_n t_n K(\mathbf{x}_n, \mathbf{x}) + w_0$$

Βήμα 2: Τι συμβαίνει όταν $\mathbf{x}_f$ είναι μακριά;

- Αν $||\mathbf{x}_n - \mathbf{x}_f||$ είναι μεγάλο $\Rightarrow ||\mathbf{x}_n - \mathbf{x}_f||^2$ πολύ μεγάλο
- $\exp(-\text{μεγάλος αριθμός}) \approx 0$
- Άρα $K(\mathbf{x}_n, \mathbf{x}_f) \approx 0$ για κάθε n

Βήμα 3: Αποτέλεσμα.

$$y(\mathbf{x}_f) = \sum_{n\in S} \mu_n t_n \cdot \underbrace{K(\mathbf{x}_n, \mathbf{x}_f)}_{\approx 0} + w_0 \approx w_0$$

Συμπέρασμα: Η απόφαση εξαρτάται **μόνο από το bias** w_0. Αν $w_0 > 0 \Rightarrow$ κλάση +1, αν $w_0 < 0 \Rightarrow$ κλάση -1.

Θέμα 1.2

Υποθέστε ότι $w_0 = 0$. Ποια συνθήκη πρέπει όλα τα διανύσματα εκπαίδευσης να ικανοποιούν; Θα την ικανοποιούσε αυτή τη συνθήκη το διάνυσμα $\mathbf{x}_f$; (2 μονάδες)

Λύση

(Πηγή: Lecture 6, slides 52--55)
Αναλυτική Εξήγηση:
Βήμα 1: Η συνθήκη margin constraint. Για σωστή ταξινόμηση όλων των training points, πρέπει:

$$t_n \cdot y(\mathbf{x}_n) \geq 1, \quad \forall n$$

όπου $t_n \in \{-1, +1\}$ είναι το label και $y(\mathbf{x}_n)$ η έξοδος του SVM.
Βήμα 2: Τι συμβαίνει με $\mathbf{x}_f$ (μακρινό σημείο);
Αν $w_0 = 0$ και $\mathbf{x}_f$ μακριά:

$$y(\mathbf{x}_f) = \sum_{n \in S} \mu_n t_n K(\mathbf{x}_n, \mathbf{x}_f) + 0 \approx 0$$

Βήμα 3: Έλεγχος της συνθήκης.

$$t_f \cdot y(\mathbf{x}_f) \geq 1 \Rightarrow t_f \cdot 0 \geq 1 \Rightarrow 0 \geq 1 \quad \textbf{(ΑΔΥΝΑΤΟ!)}$$

Συμπέρασμα: Το $\mathbf{x}_f$ **δεν μπορεί** να ικανοποιήσει τη συνθήκη margin. Θα ήταν misclassified ή θα απαιτούσε slack variable $\xi > 0$ (soft margin SVM).

Θέμα 1.3

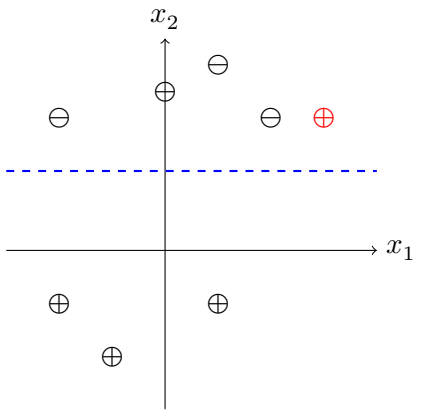
Εάν το $\mathbf{x}_f$ ήταν στο σετ εκπαίδευσης, ποια η πιθανότητα να συμπεριλαμβάνονταν μετά την εκπαίδευση στο σύνολο των διανυσμάτων υποστήριξης; (2 μονάδες)

Λύση

(Πηγή: Lecture 6, slides 52--53) **Ναι, σχεδόν βέβαιο.** Τα σημεία που παραβιάζουν το margin γίνονται SVs.

Θέμα 2 (AdaBoost)

Στο διπλανό σχήμα φαίνεται ένα σετ δεδομένων 8 δειγμάτων, με ίδιο αριθμό δειγμάτων ανά κλάση. Η διακεκομμένη γραμμή είναι το όριο απόφασης που επιλέχθηκε από τον αλγόριθμο AdaBoost στην πρώτη του επανάληψη.



Θέμα 2.1

Ποιο είναι το βάρος α_1 που θα δοθεί στο μοντέλο-ταξινομητή σε αυτή την επανάληψη του αλγόριθμου AdaBoost; (2 μονάδες)

Λύση

(Πηγή: Lecture 8, slides 31--34) $N = 8$, 1 λάθος $\Rightarrow \epsilon_1 = 1/8$.

$$\alpha_1 = \ln\left(\frac{1-\epsilon_1}{\epsilon_1}\right) = \ln\left(\frac{7/8}{1/8}\right) = \ln(7) \approx \mathbf{1.946}$$

Θέμα 2.2

Ισχύει ότι ο αλγόριθμος AdaBoost θα καταλήξει σε μηδενικό σφάλμα στο σετ εκπαίδευσης ανεξάρτητα από τον τύπο ασθενούς ταξινομητή (weak classifier) που θα χρησιμοποιήσει; Αιτιολογήστε την απάντησή σας. (1 μονάδα)

Λύση

(Πηγή: Lecture 8, slides 31--34) **Όχι.** Χρειάζεται ο weak classifier να έχει $\epsilon_t < 0.5$ σε κάθε βήμα. Αν ο weak classifier είναι πολύ απλός (π.χ. decision stump) και τα δεδομένα πολύ περίπλοκα, μπορεί να μην φτάσει σε μηδενικό σφάλμα, ή αν $\epsilon_t \geq 0.5$ ο αλγόριθμος σταματάει.

Θέμα 2.3

Ισχύει ότι τα βάρη α_t αυξάνονται καθώς προχωράει ο αλγόριθμος AdaBoost; Αιτιολογήστε την απάντησή σας. (1 μονάδα)

Λύση

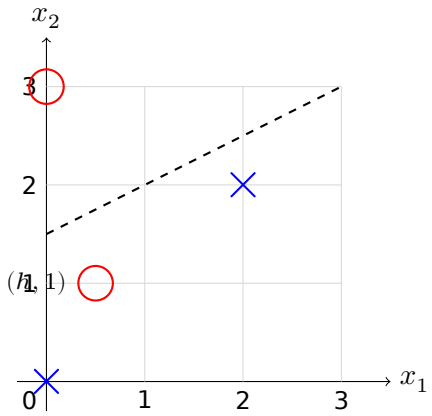
(Πηγή: Lecture 8, slides 31--34) **Όχι.** Εξαρτώνται από το ϵ_t. Συνήθως οι πρώτοι ταξινομητές έχουν μεγαλύτερα α_t.

7 Σεπτέμβριος 2024

Εξεταστική Σεπτεμβρίου - 24 Σεπτεμβρίου 2024
(βλ. Lecture 6: SVM & Kernels)

Θέμα 1 (Visual SVM)

Στο διπλανό σχήμα φαίνονται τέσσερα σημεία εκπαίδευσης. Οι σταυροί είναι η κλάση +1 ενώ οι κύκλοι η κλάση -1. Αν θεωρήσουμε $0 \leq h \leq 3$:



Θέμα 1.1

Πόσο μεγάλο μπορεί να είναι το h ώστε οι δύο κλάσεις να είναι γραμμικώς διαχωρίσιμες; (0.5 μονάδες)

Λύση

(Πηγή: Lecture 6, slides 49--55)
Απάντηση: $h_{max} = 1$.
Εξήγηση: Τα σημεία της κλάσης +1 είναι τα $(0,0)$ και $(2,2)$, τα οποία ορίζουν το ευθύγραμμο τμήμα πάνω στην ευθεία $y = x$. Τα σημεία της κλάσης -1 είναι τα $(0,3)$ και $(h,1)$. Για να είναι οι κλάσεις διαχωρίσιμες, το σημείο $(h,1)$ δεν πρέπει να περάσει "κάτω" ή πάνω στο τμήμα των +1. Επειδή το $(h,1)$ έχει $y = 1$, αντικαθιστώντας στην ευθεία $y = x$ έχουμε $1 = x \Rightarrow h = 1$. Αν $h = 1$, το σημείο είναι το $(1,1)$, το οποίο βρίσκεται ακριβώς πάνω στο ευθύγραμμο τμήμα της κλάσης +1 (ανάμεσα στο 0,0 και 2,2). Άρα, για $h < 1$ είναι διαχωρίσιμα. Οριακά, $h = 1$.

Θέμα 1.2

Αλλάζει η κατεύθυνση του ορίου απόφασης μέγιστου περιθωρίου (maximum margin) συναρτήσει του h; (0.5 μονάδες)

Λύση

(Πηγή: Lecture 6, slides 49--55)
Όχι.
Εξήγηση: Το διάνυσμα βαρών **w** (που καθορίζει την κατεύθυνση) είναι κάθετο στο διαχωριστικό υπερεπίπεδο. Για $h \in [0,1)$, το εγγύτερο σημείο της κλάσης -1 προς την κλάση +1 είναι το $(h,1)$, και τα εγγύτερα της +1 βρίσκονται στο τμήμα $y = x$. Λόγω της συμμετρίας των σημείων $(0,0)$ και $(2,2)$ ως προς το $(1,1)$, και της θέσης του $(h,1)$, η βέλτιστη διαχωριστική ευθεία διατηρεί σταθερή κλίση (παράλληλη με την $y = x$), απλά μετατοπίζεται (αλλάζει το bias) ή το πλάτος του margin στενεύει. Η διεύθυνση (κάθετη στην $y = x$) παραμένει αμετάβλητη (45 μοίρες), καθώς το "στενότερο" σημείο είναι πάντα μεταξύ της ευθείας $y = x$ και του σημείου $(h,1)$.

Θέμα 1.3

Ποια είναι η ελάχιστη και ποια η μέγιστη τιμή του πλάτους του περιθωρίου που επιτυγχάνεται με όριο απόφασης μέγιστου περιθωρίου για τα δεδομένα του σχήματος και γιατί; (1 μονάδα)

Λύση

(Πηγή: Lecture 6, slides 49--55)
Απάντηση:

- **Ελάχιστο Margin:** 0 (όταν $h = 1$)
- **Μέγιστο Margin:** $\frac{\sqrt{2}}{2} \approx 0.707$ (όταν $h = 0$)

Εξήγηση: Το margin καθορίζεται από την απόσταση του πλησιέστερου σημείου της κλάσης -1, δηλαδή του $P(h, 1)$, από την ευθεία που διέρχεται από τα σημεία της κλάσης +1 (ευθεία $x - y = 0$). Η απόσταση σημείου (x_0, y_0) από ευθεία $Ax + By + C = 0$ είναι $d = \frac{|Ax_0+By_0+C|}{\sqrt{A^2+B^2}}$. Εδώ $A = 1, B = -1, C = 0$.

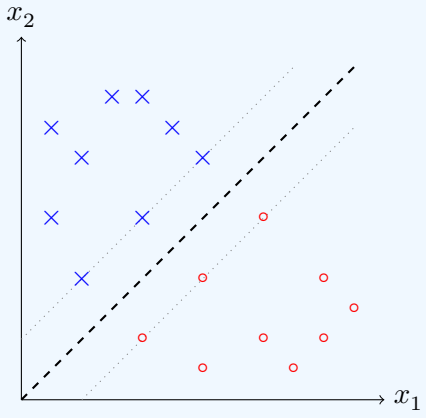
$$d(h) = \frac{|1 \cdot h - 1 \cdot 1|}{\sqrt{1^2 + (-1)^2}} = \frac{|h - 1|}{\sqrt{2}}$$

- Για $h = 1 : d = 0$. (Τα σημεία ταυτίζονται, μηδενικό margin).
- Για $h = 0 : d = \frac{|-1|}{\sqrt{2}} = \frac{1}{\sqrt{2}} = \frac{\sqrt{2}}{2}$.

Θέμα 2 (SVM & Ensembles)

Θέμα 2.1

Στο διπλανό σχήμα, ποια από τα σημεία εκπαίδευσης πρέπει να αφαιρέσουμε ώστε να αλλάξουν τα διανύσματα υποστήριξης; (1 μονάδα)



Λύση

(Πηγή: Lecture 6, slides 52--53) **Κανένα.** Η αφαίρεση μη-SVs δεν αλλάζει το margin.

Θέμα 2.2

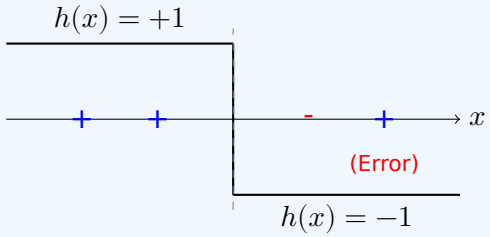
Κάθε φορά που μεταβαίνω από γραμμικό SVM σε SVM με πολυωνυμικό kernel ανώτερης τάξης, περιμένω ότι τα διανύσματα υποστήριξης δεν θα αλλάξουν. Σωστό ή Λάθος; (2 μονάδες)

Λύση

(Πηγή: Lecture 6, slides 76--79) **Λάθος.** Τα SVs αλλάζουν ανάλογα με το kernel.

Θέμα 2.3

Στο διπλανό σχήμα φαίνεται η πρώτη απόφαση από τον αλγόριθμο AdaBoost για τέσσερα σημεία εκπαίδευσης. Τα βάρη που αποδίδονται στα τέσσερα σημεία εκπαίδευσης μετά από αυτή την απόφαση είναι $[\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{5}{8}]$ (αντίστοιχη σειρά με το σχήμα). Σωστό ή Λάθος; (1 μονάδα)



Weights: $[\frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \mathbf{\frac{5}{8}}]$

Λύση

(Πηγή: Lecture 8, slides 31--34) **Σωστό.** (Υπόθεση: $\alpha_t \approx \ln((1-\epsilon)/\epsilon)$). Αν το σφάλμα είναι 1/4, το weight update αυξάνει το βάρος του λάθους και μειώνει των σωστών.

Θέμα 3 (Logistic Regression)

Θέμα 3.1

Αν εκπαιδεύσουμε ένα logistic regression μοντέλο μεγιστοποιώντας την πιθανοφάνεια των ετικετών, δεδομένων των σημείων εκπαίδευσης, καταλήγουμε σε πολλαπλές τοπικά βέλτιστες λύσεις. Ναι ή όχι; (2 μονάδες)

Λύση

(Πηγή: Lecture 5, slides 78--80) **Όχι.** Η συνάρτηση κόστους (Negative Log Likelihood) για την Logistic Regression είναι **κυρτή** (convex), άρα έχει **μοναδικό ολικό ελάχιστο**.

Θέμα 3.2

Αν εκπαιδεύσουμε ένα logistic regression μοντέλο με Stochastic Gradient Descent (SGD) αλγόριθμο με σταθερό ρυθμό μάθησης, τότε θα καταλήξουμε στα ακριβή, βέλτιστα βάρη. Συμφωνείτε ή όχι και γιατί; (2 μονάδες)

Λύση

(Πηγή: Lecture 5) **Όχι.** Με σταθερό learning rate, ο SGD **ταλαντώνεται** γύρω από το ελάχιστο και δεν συγκλίνει ποτέ σε ένα σημείο (θόρυβος λόγω στοχαστικότητας). Για σύγκλιση απαιτείται μείωση του ρυθμού μάθησης (learning rate decay).

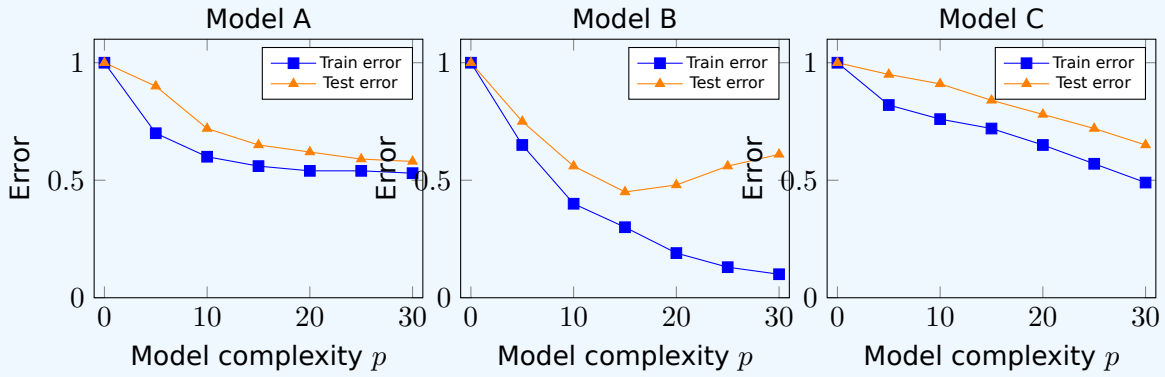
8 Φεβρουάριος 2025

Εξεταστική Φεβρουαρίου - Φεβρουάριος 2025

Θέμα 1 (Model Selection)

Θέμα 1.1

Α) Για τα παρακάτω σχήματα Train vs Test error, χαρακτηρίστε ποιο μοντέλο κάνει underfitting, overfitting και καλή γενίκευση.



Λύση

(Πηγή: Lecture 1, slides 35--38)

- **Model A: Underfitting.** Οι τιμές σφάλματος (train και test) παραμένουν υψηλές και κοντά μεταξύ τους, υποδεικνύοντας ότι το μοντέλο είναι πολύ απλό για να μάθει τα δεδομένα.
- **Model B: Overfitting.** Το σφάλμα εκπαίδευσης (train) γίνεται πολύ μικρό, αλλά το σφάλμα ελέγχου (test) αρχίζει να αυξάνεται μετά από ένα σημείο ($\rho = 15$), δημιουργώντας μεγάλο χάσμα.
- **Model C: Καλή γενίκευση.** Τα σφάλματα μειώνονται σταθερά και παραμένουν κοντά, χωρίς το μοντέλο να "απομνημονεύει" τα δεδομένα εκπαίδευσης όπως το Model B.

Θέμα 1.2

Β) Κάποιος ισχυρίζεται ότι αυξάνοντας την παράμετρο C στον soft margin SVM ταξινομητή θα αυξηθεί και το περιθώριο ταξινόμησης (margin). Συμφωνείτε μαζί του;

Λύση

(Πηγή: Lecture 6, slides 60--65) **Απ: Λάθος.** Μεγαλύτερο C σημαίνει μεγαλύτερη ποινή για σφάλματα, άρα **μικρότερο margin** (πιο σκληρό όριο).

Θέμα 1.3

Γ) Σε ένα soft margin SVM ταξινομητή οι βοηθητικές μεταβλητές ξ_i που συνοδεύουν κάθε ένα από τα δείγματα x_i, έχουν μη μηδενική τιμή για (μπορεί παραπάνω από ένα να είναι σωστά):

α) λάθος ταξινομημένα (misclassified) x_i

β) σωστά ταξινομημένα x_i

γ) x_i τα οποία είναι εντός του περιθωρίου (margin)

δ) x_i τα οποία είναι εκτός του περιθωρίου (margin)

Αιτιολογήστε την απάντησή σας.

Λύση

Απ: α, γ. Οι slack variables $\xi_i > 0$ όταν το σημείο είναι:

- Λάθος ταξινομημένο ($\xi_i > 1$)
- Εντός του margin αλλά σωστά ταξινομημένο ($0 < \xi_i \leq 1$)

(Πηγή: Lecture 6, slides 60--65)

Θέμα 1.4

Δ) Για ένα νευρωνικό δίκτυο ταξινόμησης με δύο εισόδους (2 features), 20 νευρώνες στο hidden layer και 10 νευρώνες στο output layer:

α) Χρησιμοποιούμε softmax για το output layer. **Σωστό/Λάθος;**

β) Είναι σωστό να χρησιμοποιούμε γραμμική activation στο hidden layer. **Σωστό/Λάθος;**

γ) Το δίκτυο έχει 240 παραμέτρους. **Σωστό/Λάθος;**

Λύση

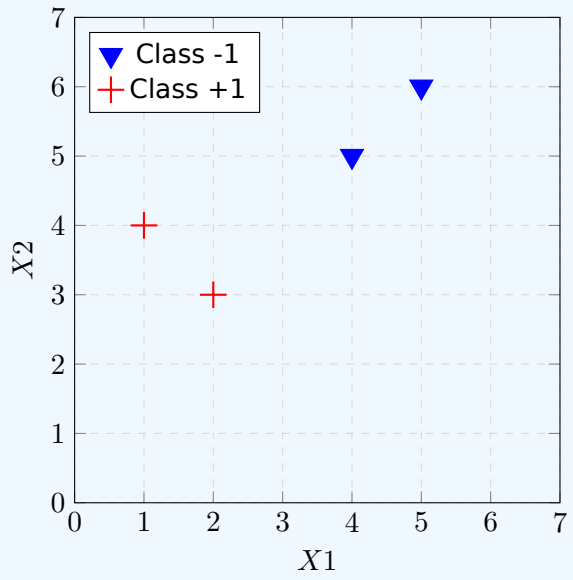
(Πηγή: Lecture 7, slides 1--2, 28--30)

- **α) Σωστό.** Για multi-class classification χρησιμοποιούμε softmax.
- **β) Λάθος.** Γραμμική activation στο hidden layer καταργεί τη δυνατότητα μη-γραμμικού διαχωρισμού.
- **γ) Λάθος.** Υπολογισμός:
  - Hidden layer: $(2 + 1) \times 20 = 60$ παράμετροι
  - Output layer: $(20 + 1) \times 10 = 210$ παράμετροι
  - **Σύνολο:** $60 + 210 = 270$ **παράμετροι**

Θέμα 2 (SVM Exercise)

Θέμα 2.1

Στο διπλανό σχήμα φαίνονται τα δεδομένα εκπαίδευσης για ένα πρόβλημα δυαδικής ταξινόμησης (binary classification problem). Βρείτε το όριο απόφασης $w^T x + w_0 = 0$ που μεγιστοποιεί το περιθώριο μεταξύ των κλάσεων (maximum margin) με βάση τη θεωρία των SVM ταξινομητών και τις παραδοχές που τη συνοδεύουν. Ποια είναι τα διανύσματα υποστήριξης και πόσο είναι το μέγιστο περιθώριο;



Λύση

Λύση:
1. Αναγνώριση Δεδομένων: Από το σχήμα έχουμε δύο κλάσεις:

- **Class +1 (Red crosses):** $x^{(1)} = (1, 4), x^{(2)} = (2, 3)$
- **Class -1 (Blue triangles):** $x^{(3)} = (4, 5), x^{(4)} = (5, 6)$

2. Εύρεση Διαχωριστικής Ευθείας: Παρατηρούμε ότι τα δεδομένα διαχωρίζονται γραμμικά. Η βέλτιστη ευθεία (maximum margin) θα βρίσκεται ακριβώς στη μέση της απόστασης των πλησιέστερων σημείων των δύο κλάσεων. Τα πλησιέστερα σημεία είναι τα $x^{(1)}, x^{(2)}$ από την Class +1 και το $x^{(3)}$ από την Class -1. Αυτά σχηματίζουν τα **Support Vectors**.
Η κατεύθυνση διαχωρισμού είναι διαγώνια. Παρατηρούμε ότι για τα +1 ισχύει $x_1 + x_2 = 5$ (σημεία (1,4) και (2,3)). Για το πλησιέστερο -1 ισχύει $x_1 + x_2 = 9$ (σημείο (4,5)).
Η διαχωριστική ευθεία θα είναι η μεσοκάθετος, άρα θα έχει εξίσωση της μορφής $x_1 + x_2 = C$. Το C είναι ο μέσος όρος των αθροισμάτων: $C = \frac{5+9}{2} = 7$.
Άρα η εξίσωση του ορίου απόφασης είναι:

$$x_1 + x_2 - 7 = 0$$

Για να την φέρουμε στην κανονική μορφή SVM ($w^T x + w_0 = 0$), όπου $y_i(w^T x_i + w_0) \geq 1$ για τα Support Vectors, ελέγχουμε: Για Class +1 ($y = 1$): θέλουμε $w^T x + w_0 \geq 1$. Για Class -1 ($y = -1$): θέλουμε $w^T x + w_0 \leq -1$.
Η εξίσωση $-(x_1 + x_2 - 7) = 0 \Rightarrow -x_1 - x_2 + 7 = 0$ δίνει: Για $x^{(1)}$ (sum=5): $-5 + 7 = 2$. Για $x^{(3)}$ (sum=9): $-9 + 7 = -2$. Για να έχουμε περιθώριο 1, διαιρούμε με το 2:

$$-0.5x_1 - 0.5x_2 + 3.5 = 0$$

Άρα $\mathbf{w} = [-0.5, -0.5]^T$ και $w_0 = 3.5$.
3. Υπολογισμός Περιθωρίου (Margin): Το margin M δίνεται από τον τύπο $M = \frac{2}{||w||}$.

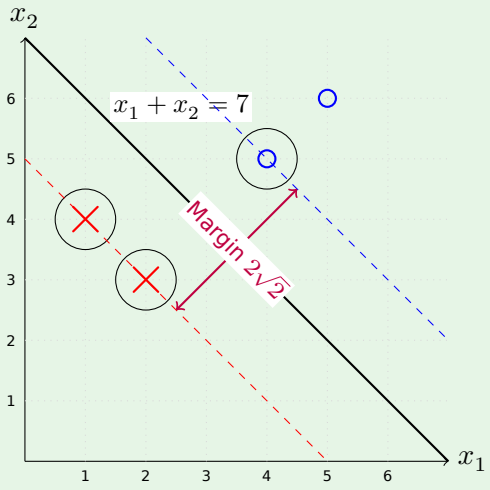
$$||w|| = \sqrt{(-0.5)^2 + (-0.5)^2} = \sqrt{0.25 + 0.25} = \sqrt{0.5} = \frac{1}{\sqrt{2}}$$

$$M = \frac{2}{1/\sqrt{2}} = 2\sqrt{2} \approx \mathbf{2.828}$$

Απαντήσεις:

- **Όριο απόφασης:** $-0.5x_1 - 0.5x_2 + 3.5 = 0$ (ή ισοδύναμα $x_1 + x_2 - 7 = 0$)

- **Support Vectors:** $(1,4), (2,3)$ από Class +1 και $(4,5)$ από Class -1.
- **Μέγιστο Περιθώριο:** $2\sqrt{2}$.



Θέμα 3 (Kernel & PCA)

Θέμα 3.1

Α) Σε τι διάσταση (αριθμό) προβάλει έμμεσα τα διανύσματα $x \in \mathbb{R}^2$, η συνάρτηση, $K(x,z) = (1+x^T z)^2$; Αιτιολογήστε την απάντησή σας.

Λύση

(Πηγή: Lecture 6, slide 76)
Ανάλυση: Αναπτύσσουμε το πολυωνυμικό kernel για $x = (x_1, x_2)$ και $z = (z_1, z_2)$:

$$K(x,z) = (1 + x_1z_1 + x_2z_2)^2 = (1 + x_1z_1 + x_2z_2)(1 + x_1z_1 + x_2z_2)$$

$$= 1 + x_1z_1 + x_2z_2 + x_1z_1 + x_1^2z_1^2 + x_1x_2z_1z_2 + x_2z_2 + x_1x_2z_1z_2 + x_2^2z_2^2$$

$$= 1 + 2x_1z_1 + 2x_2z_2 + 2x_1x_2z_1z_2 + x_1^2z_1^2 + x_2^2z_2^2$$

Αυτό μπορεί να γραφτεί ως εσωτερικό γινόμενο $\Phi(x)^T\Phi(z)$ με:

$$\Phi(x) = [1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2]^T$$

Απάντηση: Η διάσταση του feature space είναι **6**.

Θέμα 3.2

Β) Ποια (-ες) από τις παρακάτω δύο προτάσεις είναι σωστή (-ές) και ποια (-ες) λάθος σχετικά με την Ανάλυση Πρωτευουσών Συνιστωσών (Principal Component Analysis) και γιατί;

α) Προσθέτοντας μονάδα (1) στο τέλος κάθε διανύσματος δεν αλλάζει ουσιαστικά το αποτέλεσμα του PCA.

β) Αν χρησιμοποιήσω τη μέθοδο PCA για να προβάλω διανύσματα d διάστασης σε j πρωτεύουσες συνιστώσες (principal components) και κατόπιν για να προβάλω τα διανύσματα j διάστασης (μετά την πρώτη προβολή) σε k πρωτεύουσες συνιστώσες όπου $d > j > k$, θα πάρω το ίδιο αποτέλεσμα με το εάν χρησιμοποιήσω τη μέθοδο PCA για να προβάλω τα διανύσματα d διάστασης κατευθείαν σε k πρωτεύουσες συνιστώσες.

Λύση

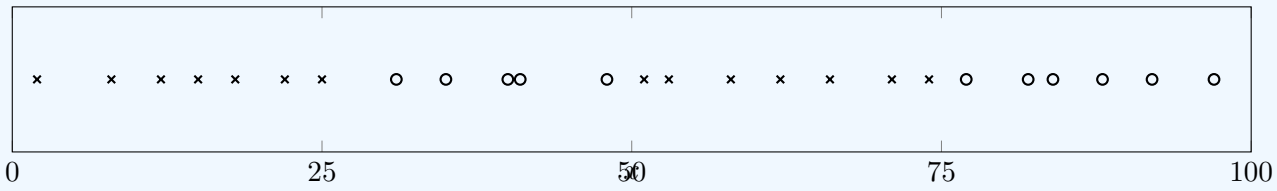
(Πηγή: Lecture 9-10, slides 52--54)

- **α) Σωστή.** Η προσθήκη μιας σταθερής τιμής (όπως το 1) σε κάθε δείγμα δεν επηρεάζει τον πίνακα συνδιακύμανσης (covariance matrix), καθώς η διακύμανση (variance) μιας σταθεράς είναι μηδέν. Επομένως, τα ιδιοδιανύσματα και οι ιδιοτιμές παραμένουν αναλλοίωτα.
- **β) Σωστή.** Το PCA βρίσκει τις κατευθύνσεις μέγιστης διακύμανσης. Οι k πρώτες συνιστώσες ενός συνόλου δεδομένων είναι οι ίδιες είτε τις εξάγουμε απευθείας, είτε περνώντας από έναν ενδιάμεσο χώρο j διαστάσεων ($j > k$), καθώς τα principal components είναι ορθογώνια μεταξύ τους και διατεταγμένα κατά φθίνουσα διακύμανση.

Θέμα 4 (1D Classification)

Θέμα 4.1

Θεωρήστε το μονοδιάστατο πρόβλημα ταξινόμησης του παρακάτω σχήματος. Ποιο επιπλέον χαρακτηριστικό θα προσθέτατε στο x ώστε να κάνετε τα δείγματα γραμμικά διαχωρίσιμα στο νέο δισδιάστατο χώρο; Εκφράστε το συναρτήσει του x.



Λύση

(Πηγή: Lecture 6, slides 76--78)
Ανάλυση: Παρατηρώντας το σχήμα, τα δεδομένα ανήκουν σε δύο κλάσεις (x και o) που εναλλάσσονται σε διαστήματα κατά μήκος του άξονα x. Συγκεκριμένα:

- Κλάση 1 (x): $[0, 25]$ και $[50, 75]$
- Κλάση 2 (o): $[25, 50]$ και $[75, 100]$

Στον μονοδιάστατο χώρο τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα. Για να γίνουν διαχωρίσιμα σε 2D, πρέπει να προσθέσουμε ένα χαρακτηριστικό $y = f(x)$ τέτοιο ώστε η διαχωριστική ευθεία $w_1x+w_2y+w_0 = 0$ να μπορεί να τα χωρίσει.
Προτεινόμενη Λύση: Χρειαζόμαστε μια συνάρτηση που να αλλάζει πρόσημο στα σημεία "επαφής" των κλάσεων (25, 50, 75). Μια απλή επιλογή είναι ένα πολυώνυμο 3ου βαθμού:

$$f(x) = (x - 25)(x - 50)(x - 75)$$

Αυτή η συνάρτηση:

- Είναι θετική για $x \in (25, 50)$ και $x > 75$.
- Είναι αρνητική για $x < 25$ και $x \in (50, 75)$.

Έτσι, τα δείγματα της μίας κλάσης θα έχουν θετικές τιμές στο νέο χαρακτηριστικό και της άλλης αρνητικές (σε συγκεκριμένα διαστήματα), επιτρέποντας έναν γραμμικό διαχωρισμό.
Εναλλακτική Λύση: Μια περιοδική συνάρτηση όπως η sin ή η cos (με κατάλληλη περίοδο):

$$f(x) = \sin\left(\frac{2\pi x}{50}\right)$$

Η συνάρτηση αυτή αλλάζει πρόσημο κάθε 25 μονάδες, "σηκώνοντας" τα δείγματα της μίας κλάσης πάνω από τον άξονα x και της άλλης κάτω από αυτόν.
Απάντηση: Το επιπλέον χαρακτηριστικό μπορεί να είναι η συνάρτηση $y = (x - 25)(x - 50)(x - 75)$ ή $y = \sin\left(\frac{2\pi x}{50}\right)$.

9 Ερωτήσεις Εξάσκησης

Επιπλέον ερωτήσεις για εξάσκηση που καλύπτουν όλη την ύλη του μαθήματος.

Θέμα 1

Σε ένα ιατρικό test, η ευαισθησία (sensitivity) είναι 95% και η ειδικότητα (specificity) είναι 90%. Αν η επικράτηση (prevalence) της νόσου είναι 1%, ποια είναι η πιθανότητα κάποιος να είναι πράγματι άρρωστος δεδομένου θετικού test;

Λύση

(Πηγή: Lecture 3, slides 22--24)
Δεδομένα:

- $P(\text{Test+}|\text{Νόσος}) = 0.95$ (sensitivity)
- $P(\text{Test-}|\text{Υγιής}) = 0.90 \Rightarrow P(\text{Test+}|\text{Υγιής}) = 0.10$ (false positive)
- $P(\text{Νόσος}) = 0.01$

Εφαρμόζουμε Bayes:

$$P(\text{Νόσος}|\text{Test+}) = \frac{P(\text{Test+}|\text{Νόσος}) \cdot P(\text{Νόσος})}{P(\text{Test+})}$$

Όπου:

$$P(\text{Test+}) = 0.95 \cdot 0.01 + 0.10 \cdot 0.99 = 0.0095 + 0.099 = 0.1085$$

$$P(\text{Νόσος}|\text{Test+}) = \frac{0.95 \cdot 0.01}{0.1085} = \frac{0.0095}{0.1085} \approx \mathbf{8.76}\%$$

Συμπέρασμα: Παρά το υψηλό sensitivity, η χαμηλή prevalence οδηγεί σε πολλά false positives.

Θέμα 2

Εξηγήστε γιατί ο αλγόριθμος k-NN αποτυγχάνει σε υψηλές διαστάσεις (curse of dimensionality) και προτείνετε τρόπους αντιμετώπισης.

Λύση

(Πηγή: Lecture 2, slides 67--72)
Πρόβλημα:

- Σε υψηλές διαστάσεις, όλα τα σημεία απέχουν **σχεδόν ίση απόσταση** μεταξύ τους.
- Η έννοια του "γείτονα" χάνει νόημα.
- Ο όγκος του χώρου αυξάνεται εκθετικά, άρα χρειάζονται εκθετικά περισσότερα δεδομένα.

Αντιμετώπιση:

1. **Μείωση διαστάσεων:** PCA, LDA, t-SNE
2. **Feature selection:** Επιλογή σχετικών χαρακτηριστικών
3. **Weighted distance:** Χρήση βαρών σε κάθε διάσταση
4. **Locality-Sensitive Hashing:** Για approximate nearest neighbors

Θέμα 3

Δεδομένα: 10 δείγματα, 6 θετικά (+) και 4 αρνητικά (-). Μετά από split σε χαρακτηριστικό X:

- **Αριστερός κόμβος: 4+, 1-**
- **Δεξιός κόμβος: 2+, 3-**

Υπολογίστε το Information Gain.

Λύση

(Πηγή: Lecture 8, slides 68--72)

Entropy πριν το split:

$$H(S) = -\frac{6}{10}\log_2\frac{6}{10} - \frac{4}{10}\log_2\frac{4}{10} = 0.971 \text{ bits}$$

Entropy μετά το split:

- Αριστερός: $H(L) = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0.722$
- Δεξιός: $H(R) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$

$$H(S|X) = \frac{5}{10}\cdot 0.722 + \frac{5}{10}\cdot 0.971 = 0.847$$

Information Gain:

$$IG(S,X) = H(S) - H(S|X) = 0.971 - 0.847 = \mathbf{0.124} \text{ bits}$$

Θέμα 4

Ο αλγόριθμος Perceptron εγγυάται σύγκλιση αν τα δεδομένα είναι γραμμικά διαχωρίσιμα. Αν δεν είναι, τι συμβαίνει; Πώς το αντιμετωπίζουμε;

Λύση

(Πηγή: Lecture 5, slides 26--31)

Αν τα δεδομένα ΔΕΝ είναι γραμμικά διαχωρίσιμα:

- Ο Perceptron **δεν συγκλίνει** --- ταλαντώνεται επ' αόριστον.
- Τα βάρη αλλάζουν συνεχώς χωρίς να φτάσουν σε σταθερή λύση.

Λύσεις:

- **Pocket Algorithm:** Κρατάμε τα καλύτερα βάρη που βρέθηκαν μέχρι στιγμής (με το μικρότερο training error).
- **Soft-margin SVM:** Επιτρέπουμε κάποια σφάλματα με ποινή (C).
- **Kernel trick:** Μετασχηματισμός σε υψηλότερη διάσταση όπου μπορεί να είναι γραμμικά διαχωρίσιμα.
- **Multi-layer Perceptron:** Χρήση νευρωνικού δικτύου για μη-γραμμικά όρια.

Θέμα 5

Συγκρίνετε τη μέθοδο Hold-out validation με την k-Fold Cross-Validation. Πότε προτιμάμε την κάθε μία;

Λύση

(Πηγή: Lecture 2, slides 54--58)

Κριτήριο	Hold-out	k-Fold CV
Χρόνος	Γρήγορο	k φορές πιο αργό
Variance εκτίμησης	Υψηλό	Χαμηλό
Bias εκτίμησης	Υψηλό (αν μικρό set)	Χαμηλό
Αξιοποίηση δεδομένων	70-80%	100%

- **Προτιμάμε Hold-out:** Μεγάλα datasets, ακριβά μοντέλα (π.χ. deep learning).
- **Προτιμάμε k-Fold CV:** Μικρά datasets, όταν χρειάζεται σταθερή/αξιόπιστη εκτίμηση.

Θέμα 6

Εξηγήστε τη διαφορά μεταξύ L1 (Lasso) και L2 (Ridge) regularization. Πότε χρησιμοποιούμε το καθένα;

Λύση

(Πηγή: Lecture 5, slides 68--72)
L2 Regularization (Ridge): $J = \text{Loss} + \lambda \sum w_i^2$

- Μειώνει τα βάρη αλλά **δεν τα μηδενίζει** (smooth solution).
- Χρήση: Όταν όλα τα features είναι σημαντικά.

L1 Regularization (Lasso): $J = \text{Loss} + \lambda \sum |w_i|$

- **Μηδενίζει** κάποια βάρη $\rightarrow$ **feature selection** (sparse solution).
- Χρήση: Όταν θέλουμε επιλογή χαρακτηριστικών ή interpretability.

Θέμα 7

Συγκρίνετε Batch Gradient Descent, Stochastic Gradient Descent (SGD) και Mini-batch GD.

Λύση

(Πηγή: Lecture 7, slides 59--61)

	Batch GD	SGD	Mini-batch
Samples/update	Όλα (N)	1	b (π.χ. 32-256)
Noise gradient	Καθόλου	Πολύ	Μέτριο
Σύγκλιση	Σταθερή	Ταλάντωση	Καλή ισορροπία
Χρόνος/epoch	Αργό	Γρήγορο	Μέτριο

Συμπέρασμα: Mini-batch GD είναι η πιο συνήθης επιλογή (π.χ. με Adam).

Θέμα 8

Αναφέρετε τρεις περιορισμούς του k-Means και προτείνετε εναλλακτικές.

Λύση

(Πηγή: Lecture 9-10, slides 5--10)
Περιορισμοί:

- Πρέπει να γνωρίζουμε το k εκ των προτέρων.
- Υποθέτει σφαιρικά clusters (αποτυγχάνει σε πολύπλοκα σχήματα).
- Ευαισθησία στην αρχικοποίηση (τοπικά ελάχιστα).

- Ευαισθησία σε outliers.

Εναλλακτικές:

- **k-Means++:** Καλύτερη αρχικοποίηση.
- **GMM:** Για ελλειπτικά clusters (soft assignment).
- **DBSCAN:** Αυτόματο k, ανθεκτικό σε θόρυβο και τυχαία σχήματα.
- **Spectral Clustering:** Για μη-κυρτά clusters.

Θέμα 9

Ποιες είναι οι βασικές διαφορές μεταξύ k-Means και Gaussian Mixture Models (GMM);

Λύση

(Πηγή: Lecture 9-10, slides 5--10, 31--36)

Χαρακτηριστικό	k-Means	GMM
Assignment	Hard (0 ή 1)	Soft (πιθανότητες)
Cluster shape	Σφαιρικά	Ελλειπτικά
Output	Κέντρα	μ, Σ, π
Αλγόριθμος	Lloyd's	EM
Interpretability	Υψηλή	Μέτρια

Θέμα 10

Συγκρίνετε Bagging, Boosting και Random Forest.

Λύση

(Πηγή: Lecture 8, slides 10--12, 31--34, 74)

	Bagging	AdaBoost	Random Forest
Στόχος	↓ Variance	↓ Bias	↓ Variance
Training	Παράλληλο	Σειριακό	Παράλληλο
Sampling	Bootstrap	Reweighting	Bootstrap + Feature
Overfitting	Χαμηλό ρίσκο	Μέτριο ρίσκο	Πολύ χαμηλό ρίσκο

Χρήση:

- **Bagging:** Για high-variance models (π.χ. Decision Trees).
- **AdaBoost:** Για high-bias models (π.χ. Decision Stumps).
- **Random Forest:** Γενικά πολύ καλό για tabular data (state-of-the-art πριν τα GBTs).
- **XGBoost/LightGBM:** Gradient Boosting για μέγιστη απόδοση.

Θέμα 11

Ποια είναι η διαφορά μεταξύ του κανόνα Hebb και του κανόνα Delta (LMS); Πότε χρησιμοποιούμε τον καθένα;

Λύση

(Πηγή: Lecture 7, slides 8--15)

Χαρακτηριστικό	Hebb Rule	Delta Rule (LMS)
Τύπος	$\Delta w = \eta \cdot x \cdot y$	$\Delta w = \eta \cdot (d - y) \cdot x$
Εποπτεία	Unsupervised	Supervised
Στόχος	Correlation learning	Error minimization
Σύγκλιση	Δεν εγγυάται	$\rightarrow$ MSE ελάχιστο

Hebb: «Neurons that fire together, wire together» --- ενισχύει συσχετίσεις.
Delta: Ελαχιστοποιεί το σφάλμα $(d - y)^2$ με gradient descent.
Χρήση:

- **Hebb:** Associative memory, feature extraction (unsupervised).
- **Delta:** Supervised learning, Adaline, βάση για backpropagation.

Θέμα 12

Εξηγήστε τις έννοιες Precision, Recall, F1-score και ROC/AUC. Πότε προτιμάμε την κάθε μετρική;

Λύση

(Πηγή: Lecture 2, slides 54--60)
Ορισμοί:

- **Precision** = $\frac{TP}{TP+FP}$ (ακρίβεια θετικών προβλέψεων)
- **Recall** = $\frac{TP}{TP+FN}$ (κάλυψη πραγματικών θετικών)
- **F1-score** = $\frac{2 \cdot P \cdot R}{P+R}$ (αρμονικός μέσος)
- **ROC curve:** TPR vs FPR για διάφορα thresholds
- **AUC:** Εμβαδόν κάτω από ROC (0.5 = τυχαίο, 1 = τέλειο)

Πότε χρησιμοποιούμε τι:

- **Precision:** Όταν τα FP είναι ακριβά (π.χ. spam filter)
- **Recall:** Όταν τα FN είναι κρίσιμα (π.χ. καρκίνος)
- **F1:** Όταν θέλουμε ισορροπία P-R
- **AUC:** Όταν συγκρίνουμε μοντέλα ανεξάρτητα από threshold

Θέμα 13

Εξηγήστε τη συνάρτηση Softmax και τη Cross-Entropy Loss. Γιατί χρησιμοποιούνται μαζί;

Λύση

(Πηγή: Lecture 7, slides 35--40)
Softmax: Μετατρέπει logits σε πιθανότητες:

$$\sigma(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Ιδιότητες: $\sum_i \sigma(z_i) = 1$, $\sigma(z_i) \in (0, 1)$
Cross-Entropy Loss:

$$L = -\sum_{i=1}^{K} y_i \log(\hat{y}_i)$$

Για one-hot encoding: $L = -\log(\hat{y}_c)$ όπου c η σωστή κλάση.

Γιατί μαζί:

- Η παράγωγος είναι απλή: $\frac{\partial L}{\partial z_i} = \hat{y}_i - y_i$
- Αποφεύγει numerical instability (log-sum-exp trick)
- Probabilistic interpretation (maximum likelihood)

Θέμα 14

Δεδομένα 4 σημεία: $(1,2), (2,4), (3,6), (4,8)$**. Εφαρμόστε PCA και βρείτε το πρώτο principal component.**

Λύση

(Πηγή: Lecture 9-10, slides 45--57)
Βήμα 1: Κεντράρισμα (Centering) Υπολογίζουμε τις μέσες τιμές: $\bar{x} = \frac{1+2+3+4}{4} = 2.5$ και $\bar{y} = \frac{2+4+6+8}{4} = 5$.
Ο πίνακας X (Centered Data Matrix) προκύπτει αφαιρώντας τις μέσες τιμές από κάθε παρατήρηση:

$$X = \begin{bmatrix} 1-2.5 & 2-5 \\ 2-2.5 & 4-5 \\ 3-2.5 & 6-5 \\ 4-2.5 & 8-5 \end{bmatrix} = \begin{bmatrix} -1.5 & -3 \\ -0.5 & -1 \\ 0.5 & 1 \\ 1.5 & 3 \end{bmatrix}$$

Βήμα 2: Πίνακας Συνδιακύμανσης

$$\Sigma = \frac{1}{n-1} X^T X = \frac{1}{3} \begin{bmatrix} -1.5 & -0.5 & 0.5 & 1.5 \\ -3 & -1 & 1 & 3 \end{bmatrix} \begin{bmatrix} -1.5 & -3 \\ -0.5 & -1 \\ 0.5 & 1 \\ 1.5 & 3 \end{bmatrix} = \begin{bmatrix} \frac{5}{3} & \frac{10}{3} \\ \frac{10}{3} & \frac{20}{3} \end{bmatrix}$$

Βήμα 3: Ιδιοτιμές $\det(\Sigma - \lambda I) = 0 \Rightarrow \lambda_1 = \frac{25}{3}, \lambda_2 = 0$
Βήμα 4: Ιδιοδιανύσματα Για $\lambda_1 = \frac{25}{3}$: $v_1 = \frac{1}{\sqrt{5}}(1,2)^T$
Αποτέλεσμα:

- **PC1:** $(1,2)^T/\sqrt{5}$ --- η κατεύθυνση μέγιστης διακύμανσης
- **Explained variance:** $\frac{\lambda_1}{\lambda_1+\lambda_2} = 100\%$
- Τα δεδομένα είναι τέλεια collinear (1D manifold)

Θέμα 15

Πώς επιλέγουμε τον αριθμό των principal components να διατηρήσουμε; Ποια είναι τα πλεονεκτήματα και μειονεκτήματα του PCA;

Λύση

(Πηγή: Lecture 9-10, slides 52--57)
Επιλογή αριθμού PCs:

1. **Explained Variance Threshold:** Διατηρούμε όσα PCs χρειάζονται για $\geq 90 - 95\%$ variance
2. **Elbow Method:** Scree plot --- σημείο καμπής στο γράφημα ιδιοτιμών
3. **Kaiser Criterion:** Κρατάμε PCs με $\lambda > 1$ (για standardized data)

Πλεονεκτήματα PCA:

- Μείωση διαστάσεων $\rightarrow$ ταχύτερη εκπαίδευση

- Απαλλαγή από multicollinearity
- Noise reduction (αφαιρώντας τελευταία PCs)
- Visualization (προβολή σε 2D/3D)

Μειονεκτήματα:

- **Γραμμικός:** Δεν πιάνει μη-γραμμικές σχέσεις (χρήση Kernel PCA)
- **Απώλεια ερμηνείας:** Τα PCs δεν αντιστοιχούν σε αρχικά features
- **Ευαίσθητο σε scaling:** Απαιτεί standardization
- **Unsupervised:** Δεν λαμβάνει υπόψη τις ετικέτες (vs LDA)

Θέμα 16

Fisher's Linear Discriminant (LDA): Δίνονται δύο κλάσεις στον 2D χώρο:

- C_1: $x_1 = (1,2)^T, x_2 = (2,3)^T$
- C_2: $x_3 = (4,5)^T, x_4 = (5,6)^T$

Υπολογίστε τον πίνακα ενδο-κλασικής διασποράς S_W, τον πίνακα δια-κλασικής διασποράς S_B και τη βέλτιστη διεύθυνση προβολής w.

Λύση

(Πηγή: Lecture 5, slides 38--45)

Βήμα 1: Υπολογισμός μέσων τιμών

$$m_1 = \frac{1}{2}[(1,2) + (2,3)] = (1.5, 2.5)^T$$

$$m_2 = \frac{1}{2}[(4,5) + (5,6)] = (4.5, 5.5)^T$$

Βήμα 2: Υπολογισμός S_W (Within-class Scatter)

$$S_1 = \sum_{x \in C_1} (x - m_1)(x - m_1)^T$$

$$x_1 - m_1 = (-0.5, -0.5)^T \Rightarrow (x_1 - m_1)(x_1 - m_1)^T = \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{bmatrix}$$

$$x_2 - m_1 = (0.5, 0.5)^T \Rightarrow (x_2 - m_1)(x_2 - m_1)^T = \begin{bmatrix} 0.25 & 0.25 \\ 0.25 & 0.25 \end{bmatrix}$$

$$S_1 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

Ομοίως για C_2:

$$x_3 - m_2 = (-0.5, -0.5)^T, \quad x_4 - m_2 = (0.5, 0.5)^T$$

$$S_2 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$$

$$S_W = S_1 + S_2 = \begin{bmatrix} 1.0 & 1.0 \\ 1.0 & 1.0 \end{bmatrix}$$

Βήμα 3: Υπολογισμός S_B (Between-class Scatter)

$$S_B = (m_2 - m_1)(m_2 - m_1)^T$$

$$m_2 - m_1 = (3,3)^T$$

$$S_B = \begin{bmatrix} 9 & 9 \\ 9 & 9 \end{bmatrix}$$

Βήμα 4: Βέλτιστη διεύθυνση w Η λύση δίνεται από $w \propto S_W^{-1}(m_2 - m_1)$. Εδώ ο S_W είναι singular (μη αντιστρέψιμος) επειδή τα σημεία είναι σε ευθεία. Χρησιμοποιούμε ψευδοαντίστροφο ή παρατηρούμε ότι $w \propto (m_2 - m_1)$ αν $S_W \propto I$. Σε αυτή την ειδική περίπτωση (perfectly aligned data), η κατεύθυνση είναι παράλληλη με τη διαφορά των μέσων.

$$w \propto \begin{bmatrix} 3 \\ 3 \end{bmatrix} \propto \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Θέμα 17

Maximum Likelihood Estimation (MLE): Έστω δείγμα $D = \{2, 4, 6, 8\}$ που προέρχεται από Κανονική κατανομή $N(\mu, \sigma^2)$. Υπολογίστε τις εκτιμήσεις μέγιστης πιθανοφάνειας $\hat{\mu}_{ML}$ και $\hat{\sigma}^2_{ML}$.

Λύση

(Πηγή: Lecture 4, slides 4--10)
Μέση Τιμή $\hat{\mu}_{ML}$:

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n = \frac{2+4+6+8}{4} = \frac{20}{4} = \mathbf{5}$$

Διασπορά $\hat{\sigma}^2_{ML}$ (Biased):

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$$

$$(2-5)^2 = 9$$

$$(4-5)^2 = 1$$

$$(6-5)^2 = 1$$

$$(8-5)^2 = 9$$

$$Sum = 20$$

$$\hat{\sigma}^2_{ML} = \frac{20}{4} = \mathbf{5}$$

(Σημείωση: Η αμερόληπτη εκτιμήτρια θα ήταν $\frac{20}{3} = 6.67$).

Θέμα 18

Hierarchical Clustering (Single Linkage): Δίνονται τα σημεία στον 1D χώρο: $A = 1, B = 4, C = 5, D = 8$. Εφαρμόστε ιεραρχική συσταδοποίηση με **Single Linkage** (ελάχιστη απόσταση) και σχεδιάστε το δενδρόγραμμα.

Λύση

(Πηγή: Lecture 9-10, slides 20--25)
Πίνακας Αποστάσεων:

	A	B	C	D
A	0	3	4	7
B	3	0	**1**	4
C	4	**1**	0	3
D	7	4	3	0

Βήμα 1: Ελάχιστη απόσταση είναι 1 (μεταξύ B και C). Ενώνουμε (B, C). Νέες αποστάσεις για το cluster (BC):

$$d(A, BC) = \min(d(A, B), d(A, C)) = \min(3, 4) = 3$$

$$d(D, BC) = \min(d(D, B), d(D, C)) = \min(4, 3) = 3$$

Βήμα 2: Ελάχιστη απόσταση είναι 3. Έχουμε ισοπαλία: A με (BC) ή D με (BC). Ας ενώσουμε A με $(BC) \to (ABC)$. Νέα απόσταση: $d(D, ABC) = \min(d(D, A), d(D, BC)) = \min(7, 3) = 3$.
Βήμα 3: Ενώνουμε D με $(ABC) \to (ABCD)$ σε απόσταση 3.
Δενδρόγραμμα:

- Ύψος 1: Ένωση B-C.
- Ύψος 3: Ένωση A με (BC) και D με (ABC).

Θέμα 19

Convolutional Neural Networks (CNN): Έστω εικόνα εισόδου 5×5 και φίλτρο (kernel) 3×3, με stride $s = 1$ και padding $p = 0$.

1. Ποια είναι η διάσταση του feature map εξόδου;
2. Αν η εικόνα είναι όλες οι τιμές 1 και το φίλτρο είναι όλες οι τιμές 2, ποια είναι η τιμή του πάνω αριστερά pixel εξόδου;

Λύση

1. Διάσταση Εξόδου: Ο τύπος είναι: $O = \lfloor \frac{W-K+2P}{S} \rfloor + 1$.

$$O = \lfloor \frac{5-3+0}{1} \rfloor + 1 = 2 + 1 = \mathbf{3}$$

Άρα έξοδος 3×3.
2. Τιμή Pixel: Η πράξη της συνέλιξης (convolution) είναι το άθροισμα των γινομένων (dot product) του φίλτρου με το αντίστοιχο patch της εικόνας.

$$\text{Pixel} = \sum_{i=1}^{3} \sum_{j=1}^{3} (I_{ij} \cdot K_{ij})$$

Αφού $I_{ij} = 1$ και $K_{ij} = 2$:

$$\text{Pixel} = \sum_{i=1}^{9} (1 \cdot 2) = 9 \cdot 2 = \mathbf{18}$$

(Συνήθως προστίθεται και ένα bias, εδώ υποθέτουμε bias=0).