ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

7ο εξάμηνο — Ακαδημαϊκό έτος 2023-2024

1η Σειρά Ασκήσεων

Ημερ. Παράδ.: 07.12.2023

Γενικές Οδηγίες: Οι αναλυτικές σειρές ασκήσεων είναι ατομικές, και οι λύσεις που θα δώσετε πρέπει να αντιπροσωπεύουν μόνο την προσωπική σας εργασία. Εξηγήστε επαρκώς την εργασία σας. Αν χρησιμοποιήσετε κάποια άλλη πηγή εκτός των σημειώσεων για την λύση σας, πρέπει να το αναφέρετε. Η παράδοση των λύσεων των αναλυτικών ασκήσεων της σειράς αυτής θα γίνει ηλεκτρονικά στην HELIOS ιστοσελίδα του μαθήματος και θα πρέπει να την υποβάλετε ως ένα ενιαίο αρχείο PDF με το εξής filename format χρησιμοποιώντας μόνο λατινικούς χαρακτήρες: ML23_hwk1_AM_LastnameFirstname.pdf, όπου AM ειναι ο 8-ψήφιος αριθμός μητρώου σας. Σκαναριμένες χειρόγραφες λύσεις επιτρέπονται αρκεί να είναι καθαρογραμμένες και ευανάγνωστες. Επίσης στην 1η σελίδα των λύσεων θα αναγράφετε το ονοματεπώνυμο, Α.Μ., και email address σας. Συμπεριλάβετε και τον κώδικα προγραμμάτων, π.χ. Matlab ή Python, που χρησιμοποιήσατε για αριθμητική επίλυση. Να σημειωθεί ότι η καταληκτική ημερομηνία παράδοσης είναι τελική και δεν θα υπάρξει παράταση.

Άσκηση 1.1 (Linear and Ridge Regression)

Θεωρήστε το ακόλουθο μοντέλο παλινδρόμησης πάνω σε 11 ανεξάρτητες μεταβλητές (χαρακτηριστικά) $x_i, i = 1, 2, \ldots, 11$:

$$y = \sum_{i=1}^{11} w_i x_i.$$

Το σύνολο δεδομένων που θα χρησιμοποιηθούν στη συγκεκριμένη άσκηση περιλαμβάνεται στο αρχείο ML2023-24-hwk1.csv, το οποίο είναι διαθέσιμο στην ιστοσελίδα του μαθήματος (Πηγή: https://archive.ics.uci.edu/dataset/186/wine+quality). Το αρχείο δεδομένων περιλαμβάνει πληροφορίες για τις χημικές ιδιότητες διαφόρων τύπων κόκκινου κρασιού και πως αυτές σχετίζονται με την ποιότητα του κρασιού. Στον πίνακα δεδομένων που περιέχεται στο αρχείο, οι πρώτες 11 στήλες αντιστοιχούν στις ανεξάρτητες μεταβλητές x_i και η τελευταία στήλη (quality/ποιότητα) στην εξαρτημένη μεταβλητή y. Πριν χρησιμοποιήσετε τις αριθμητικές τιμές των δεδομένων για τα ανωτέρω μεγέθη, θα σας βοηθήσει πρώτα να τα κανονικοποιήσετε (δηλ. τα στοιχεία σε κάθε στήλη του πίνακα δεδομένων να έχουν μέση τιμή 0, και τυπική απόκλιση 1).

(α) Υπολογίστε τον κανονικοποιημένο συντελεστή συσχέτισης $r_{(9,10)}$ μεταξύ της ένατης (pH) και δέκατης (sulphates) ανεξάρτητης μεταβλητής. Στη συνέχεια σχεδιάστε έναν πίνακα συσχέτισης (correlation matrix) που περιέχει τα γραφήματα διασποράς (scatterplots) όλων των ανεξάρτητων μεταβλητών ανά δύο και σχολιάστε τα αποτελέσματα που αφορούν τις συσχετίσεις μεταξύ των χαρακτηριστικών.

Στη συνέχεια θεωρήστε ότι οι πρώτες 100 γραμμές του πίνακα δεδομένων αποτελούν τα δεδομένα εκπαίδευσης (Training Set) και οι επόμενες 50 γραμμές τα δεδομένα επαλήθευσης (Test Set).

(β) Υπολογίστε τα βάρη $w_i, i = 1, 2, \ldots, 11$, εφαρμόζοντας τον αλγόριθμο/εξισώσεις για γραμμική παλινδρόμηση (Linear Regression) στα δεδομένα εκπαίδευσης.

(γ) Υπολογίστε τα βάρη $w_i, i = 1, 2, \ldots, 11$, εφαρμόζοντας τον αλγόριθμο/εξισώσεις για Ridge Regression στα δεδομένα εκπαίδευσης για τιμές της παραμέτρου $\lambda = 10, 100, 200$.

(δ) Σχεδιάστε σε ένα κοινό γράφημα τα διανύσματα των βαρών που πήρατε στα ερωτήματα (β) και (γ) (4 περιπτώσεις) και σχολιάστε.

(ε) Υπολογίστε τα σφάλματα RMSE (τετραγωνική ρίζα των μέσων τετραγωνικών σφαλμάτων) εκπαίδευσης και επαλήθευσης που αντιστοιχούν στα παραπάνω ερωτήματα (β) και (γ), και έπειτα συμπληρώστε τα σε έναν πίνακα της παρακάτω μορφής. Ποια τιμή της παραμέτρου λ θα επιλέγατε και με βάση ποιο κριτήριο? Ποια συμπεράσματα προκύπτουν?

Σημείωση: Επεξηγήστε αναλυτικά τη διαδικασία που ακολουθήσατε για να φθάσετε στις λύσεις σας. Για την επίλυση μπορείτε να χρησιμοποιήσετε προγραμματιστικά εργαλεία (π.χ. Python, Matlab) που διευκολύνουν λειτουργίες γραμμικής άλγεβρας (όπως πολλαπλασιασμό ή αντιστροφή πινάκων), αλλά όχι τις έτοιμες υλοποιήσεις για Linear Regression, Ridge Regression που περιέχονται σε βιβλιοθήκες (όπως scikit-learn κλπ). Προαιρετικά, θα μπορούσατε να τις χρησιμοποιήσετε για επαλήθευση (μόνο) των ανωτέρω αποτελεσμάτων/εξισώσεων.

	Linear Regression ($\lambda = 0$)	Ridge Regression ($\lambda = 10$)	Ridge Regression ($\lambda = 100$)	Ridge Regression ($\lambda = 200$)
RMSE (Training set)				
RMSE (Test set)				

Άσκηση 1.2 (Multivariate Gaussian distribution)

(α) Έστω το τυχαίο διάνυσμα (τ.δ.) $\mathrm{x} = [x_1, x_2]^T$ που ακολουθεί την κανονική κατανομή, $\mathrm{x} \sim \mathcal{N}(\mathrm{x}|\boldsymbol{\mu}, \Sigma)$, με

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}.$$

Να δειχτεί ότι η υπο συνθήκη συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) $p(x_1|x_2 = a)$ είναι Gaussian, $\mathcal{N}(\mu, \sigma^2)$, με παραμέτρους,

$$\mu = \mu_1 + \frac{\sigma_{12}(a - \mu_2)}{\sigma_2^2}, \quad \sigma^2 = \frac{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}{\sigma_2^2}.$$

Έστω, τώρα, το τυχαίο διάνυσμα (τ.δ.) $\mathrm{x} = [x_1, x_2, x_3]^T$ που ακολουθεί την κανονική κατανομή, $\mathrm{x} \sim \mathcal{N}(\mathrm{x}|\boldsymbol{\mu}, \Sigma)$, με παραμέτρους,

$$\boldsymbol{\mu} = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & 0.8 & 0.5 \\ 0.8 & 2 & 1 \\ 0.5 & 1 & 3 \end{bmatrix}.$$

(β) Προσδιορίστε την υπό συνθήκη σ.π.π. $p(x_1, x_2|x_3 = 1)$.

(γ) Επαναλάβετε το ερώτημα (β) για την σ.π.π. $p(x_1, x_3|x_2 = 1)$.

(δ) Για κάθε μία από τις κατανομές των ερωτημάτων (β) και (γ), προσδιορίστε τις εξισώσεις των ισοσταθμικών τους καμπυλών και σχεδιάστε τις καμπύλες αυτές σε ένα κοινό σχήμα.

Άσκηση 1.3 (Bayes classifier)

Θεωρούμε το πρόβλημα ταξινόμησης σε δύο ισοπίθανες κλάσεις στο $\mathbb{R}^2$, και υποθέτουμε ότι τα σημεία των δύο κλάσεων προέρχονται από τις Gaussian κατανομές $p(\mathrm{x}|\omega_1)$ και $p(\mathrm{x}|\omega_2)$ που έχουν μέσες τιμές $\boldsymbol{\mu}_1 = [-2, 0]^T$ και $\boldsymbol{\mu}_2 = [2, 1]^T$ αντίστοιχα. Για κάθε μία από τις παρακάτω περιπτώσεις, να βρείτε την ευθεία απόφασης που διαχωρίζει τις δύο κλάσεις κατά βέλτιστο τρόπο.

(α) Οι $p(\mathrm{x}|\omega_1)$, $p(\mathrm{x}|\omega_2)$ έχουν κοινό πίνακα συνδιακύμανσης $\Sigma = I$.

(β) Οι $p(\mathrm{x}|\omega_1)$, $p(\mathrm{x}|\omega_2)$ έχουν κοινό πίνακα συνδιακύμανσης:

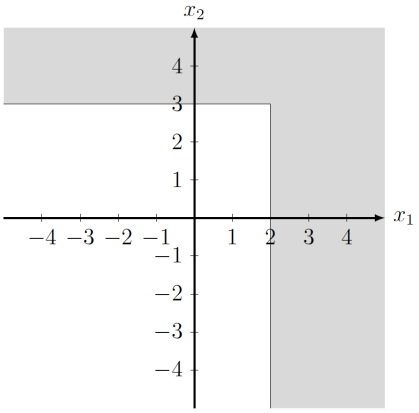
$$\hat{\Sigma} = \begin{bmatrix} 1 & -0.6 \\ -0.6 & 1 \end{bmatrix}.$$

(γ) Οι $p(\mathrm{x}|\omega_1)$, $p(\mathrm{x}|\omega_2)$ έχουν κοινό πίνακα συνδιακύμανσης $\hat{\Sigma}$ και οι συντελεστές βαρύτητας των δύο σφαλμάτων ταξινόμησης είναι $\lambda_{12} = 1$ και $\lambda_{21} = 1/2$, αντίστοιχα.

(δ) Για κάθε μία από τις παραπάνω τρεις περιπτώσεις, να παραχθούν 200 σημεία από κάθε κατανομή και να δοθούν τρία σχήματα στα οποία θα φαίνονται τα σημεία, τα κέντρα των κλάσεων και οι βέλτιστες ευθείες απόφασης. Κατόπιν να σχολιάσετε τα αποτελέσματα που θα πάρετε.

Άσκηση 1.4 (Perceptron - MultiLayer Perceptron)

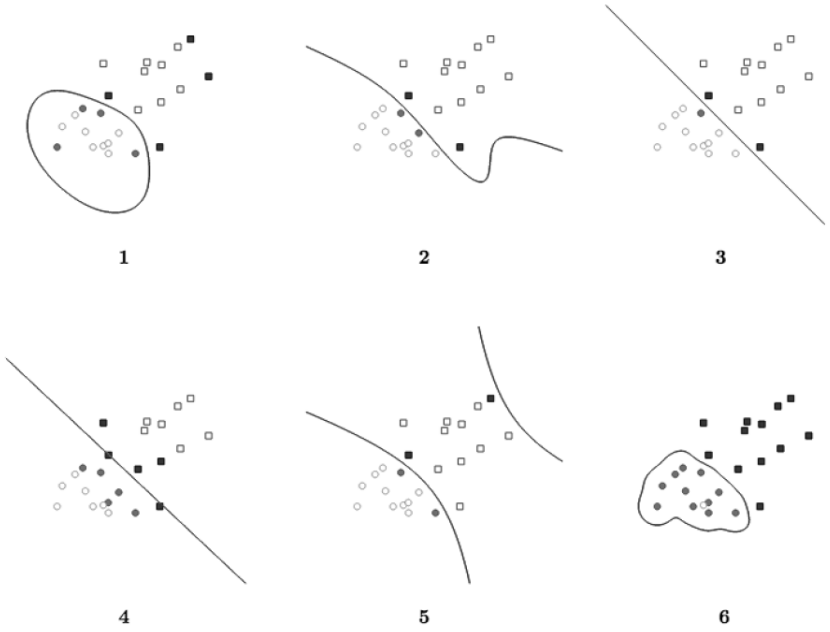
(α) Για το παρακάτω σύνολο δειγμάτων εφαρμόστε τον αλγόριθμο μάθησης perceptron μέχρι να συγκλίνει. Θεωρήστε βηματική συνάρτηση ενεργοποίησης, αρχικές τιμές βαρών $(w_0, w_1, w_2, w_3)^T = (1, 0, 0, 0)^T$ και εφαρμόστε τα δείγματα με τη σειρά που δίνονται κυκλικά. Για κάθε βήμα του αλγορίθμου καταγράψτε: το εφαρμοζόμενο δείγμα, την έξοδο του perceptron, τον χαρακτηρισμό του αποτελέσματος της ταξινόμησης (TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative), το διάνυσμα μεταβολής των βαρών και τέλος το ανανεωμένο διάνυσμα βαρών που προκύπτει. Δείγματα: $(4, 3, 6)^T \in C_N$, $(2, -2, 3)^T \in C_P$, $(1, 0, -3)^T \in C_P$, $(4, 2, 3)^T \in C_N$, όπου C_P, C_N η θετική και η αρνητική κλάση αντίστοιχα.



(β) Σχεδιάστε δίκτυο με δύο εισόδους x_1, x_2 και τρεις νευρώνες κατά μέγιστο, το οποίο διαχωρίζει τον διδιάστατο χώρο όπως φαίνεται στην εικόνα, ταξινομώντας την γκρίζα περιοχή, συμπεριλαμβανομένου του ορίου, ως θετική, και την υπόλοιπη ως αρνητική. Σχεδιάστε την τοπολογία του δικτύου, σημειώνοντας τις τιμές των βαρών και των πολώσεων και εξηγήστε την απάντησή σας.

Άσκηση 1.5 (Support Vector Machines - Kernels)

Στην παρακάτω εικόνα φαίνονται τα όρια απόφασης μηχανών διανυσμάτων υποστήριξης (SVM) με διαφορετικούς πυρήνες (kernels) ή/και διαφορετικές τιμές ποινής χαλαρότητας C. Σε όλες τις περιπτώσεις τα δεδομένα εκπαίδευσης χωρίζονται σε δύο κλάσεις με ετικέτες $y_i \in \{-1, 1\}$, οι οποίες αναπαρίστανται αντίστοιχα με κύκλους και τετράγωνα. Οι κύκλοι και τα τετράγωνα με σκούρο χρώμα αντιστοιχούν σε διανύσματα υποστήριξης.



Για καθένα από τα διαγράμματα (1)-(6) της εικόνας να εξηγήσετε αναλυτικά σε ποιο από τα προβλήματα της παρακάτω λίστας (α)-(στ) αντιστοιχεί.

(α) Γραμμικό SVM με $C = 0.1$

(β) Γραμμικό SVM με $C = 10$

(γ) Μη γραμμικό SVM με $k(u, v) = u^T v + (u^T v)^2$

(δ) Μη γραμμικό SVM με $k(u, v) = exp(-0.25||u - v||_2^2)$

(ε) Μη γραμμικό SVM με $k(u, v) = exp(-4||u - v||_2^2)$

(στ) Κανένα από τα παραπάνω

Άσκηση 1.6 (Decision Trees)

(α) Έστω T ένα δέντρο απόφασης που κατασκευάζεται με τον αλγόριθμο DecisionTree.ID3($\mathbb{D}$) από ένα σύνολο δεδομένων $\mathbb{D}$ και έστω $p = t_1 t_2 \cdots t_n$ ένα τυχαίο μονοπάτι του δέντρου, όπου $t_i, i \in \mathbb{N}_n$ ένας κόμβος του T, με t_i πρόγονο του t_j για $i < j$. Έστω επίσης $\text{ig}(t)$ η συνάρτηση που υπολογίζει το κέρδος πληροφορίας του κόμβου t του δέντρου, με βάση την εντροπία.

1. Να ελέγξετε αν ισχύει $\text{ig}(t) \geq 0$ για κάθε κόμβο t του δέντρου T.
2. Έστω t_i, t_j με $i < j$ δύο κόμβοι του T με την ίδια επιλογή χαρακτηριστικού, η οποία στηρίζεται σε κριτήριο ισότητας (η διακλάδωση γίνεται με έλεγχο κριτηρίου ισότητας για την τιμή του χαρακτηριστικού). Να ελέγξετε αν ισχύει $\text{ig}(t_j) > 0$.

(β) Δίνονται οι παρατηρήσεις που απεικονίζονται στο παρακάτω Σχήμα. Να υπολογίσετε δύο δέντρα απόφασης, χρησιμοποιώντας το κριτήριο gini για τον υπολογισμό του κέρδους πληροφορίας, αντιμετωπίζοντας το χαρακτηριστικό Temperature εναλλακτικά ως αριθμητικό ή ως κατηγορικό (εφαρμόζοντας κριτήριο ανισότητας με μία τιμή ή κριτήριο ισότητας αντίστοιχα). Ποιο από τα δύο δέντρα θα επιλέγατε για την ταξινόμηση και γιατί;

x_1	x_2	x_3	x_4	y
1	0	1	0	1
0	1	0	1	1
1	0	1	0	1
1	0	1	1	1
0	1	0	0	1
1	0	1	1	-1
0	1	1	0	-1
0	0	0	0	-1
0	0	1	0	-1
1	0	0	0	-1

Σχήμα 1: DT