

$$\begin{aligned} x &= [x_1 \ x_2]^T \sim \text{Gaussian } N(\mu, \Sigma) \\ x_1 | x_2 &\sim \text{Gaussian } N(\mu_{1|x_2}, \sigma_{1|x_2}^2) \end{aligned} \quad \left\{ \begin{aligned} \mu_{1|x_2} &= \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2) \\ \sigma_{1|x_2}^2 &= \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21} \end{aligned} \right.$$

least squares:

$$\begin{aligned} \text{1D: class } w_1: \{x_i\} &\rightarrow \bar{x}_1 = \Sigma x_i / N_1 \\ \text{class } w_2: \{x_i\} &\rightarrow \bar{x}_2 = \Sigma x_i / N_2 \end{aligned} \quad \left\{ \begin{aligned} x_d &= \frac{\bar{x}_1 + \bar{x}_2}{2} \\ x &= \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \end{aligned} \right. \quad \left\{ \begin{aligned} y &= Xw \\ w &= (X^T X)^{-1} X^T y \end{aligned} \right.$$

$$\text{2D: } \bar{y} = \partial_1 \bar{x} + \partial_0 \quad \leadsto \quad \partial_1 = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\Sigma (x_i - \bar{x})^2}, \quad \partial_0 = \bar{y} - \partial_1 \bar{x}$$

Μέγιστη πιθανοφάνεια:

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^N P(x_i, \lambda) \\ \ell(\lambda) &= \ln L(\lambda) \end{aligned} \quad \left\{ \begin{aligned} \text{απόρρ: } L(\lambda) \text{ ή } \ell(\lambda) &\rightarrow \max \Leftrightarrow \frac{d\ell(\lambda)}{d\lambda} = 0 \Rightarrow \lambda = [\dots] \end{aligned} \right.$$

$$\text{PCA: } C \text{ unbiased} = \frac{1}{N-1} \Sigma_{i=1}^N (x_i - \bar{x})(x_i - \bar{x})^T = \frac{1}{N-1} \begin{bmatrix} \Sigma (x_i - \bar{x})^2 & \Sigma (x_i - \bar{x})(y_i - \bar{y}) \\ \Sigma (x_i - \bar{x})(y_i - \bar{y}) & \Sigma (y_i - \bar{y})^2 \end{bmatrix}$$

$$\det(C - \lambda I) = 0 \Rightarrow \left\{ \begin{aligned} \lambda &\rightarrow \text{principal} \rightarrow \max \{\lambda_i\} \\ \lambda &= \end{aligned} \right. \quad \left\{ \begin{aligned} \text{ιδιοδιαν: } \lambda_1 C &= \lambda_1 u_1 \\ &\rightarrow u_1 = \begin{bmatrix} \cdot \\ \cdot \end{bmatrix}^T \end{aligned} \right.$$

Κανόνας Bayes

$$\begin{aligned} \text{εκ των προτέρων (prior): } &p(x|w_1), p(x|w_2) \\ \text{εκ των υστέρων (posterior): } &P(w_1|x), P(w_2|x) \end{aligned} \quad \left\{ \begin{aligned} P(w_1|x) &= \frac{P(x|w_1)P(w_1)}{P(x)} \end{aligned} \right.$$

$$\text{Κανόνας απόφασης: } \lambda_{12} P(w_1|x) = \lambda_{21} P(w_2|x) \Rightarrow \lambda_{12} P(x|w_1)P(w_1) = \lambda_{21} P(x|w_2)P(w_2) \Rightarrow x_r =$$

$$\left. \begin{aligned} \text{① Ισοτιμ κλάσεις} \\ \text{② Gaussian} \\ \text{③ } \mu_1, \mu_2, \Sigma \rightarrow \text{γνωστά} \end{aligned} \right\} \quad \left\{ \begin{aligned} g(x) &= w^T x + w_0 = \underbrace{[\Sigma^{-1}(\mu_1 - \mu_2)]^T}_{w^T} x + \underbrace{[-1/2 (\mu_1^T \Sigma^{-1} \mu_1 - \mu_2^T \Sigma^{-1} \mu_2)]}_{w_0} \\ &\Rightarrow g(x) = 0 \end{aligned} \right.$$

Single/Complete Link

- ① Κάθε σημείο δικό του cluster
- ② Ένωσε τα κοντινότερα clusters
- ③ Ανανέω ανασώσεις: $d(x_1, x_2, x_3) = \begin{cases} \max(d_{13}, d_{23}), & \text{complete} \\ \min(d_{12}, d_{13}), & \text{single} \end{cases}$

FCNN-CNN

Input: $N \times N$, $v = \# \text{feature vectors}$

FCNN: $N \times N \times N \times N \times v$

CNN (3×3): $3 \times 3 \times v$

MLP

$\# \text{ανανεώσεων} = \# \text{minibatches} \cdot \# \text{βαθμών}$

Neural Networks

▷ Ενώ δικτυο χωρίς κρυφά στρώματα μπορεί να είναι ισόδυναμο μόνο αν οι συναρτήσεις ενεργ είναι γραμμικές $\Rightarrow$ περιγράφει γραμ σχέσεις

Least Squares

Στόχος: ▷ εύρεση βαρών w που ελαχιστοποιούν τετραχ σφάλμα

▷ λύση κλειστής μορφής $f(\tilde{x}) = \tilde{\theta}^T \tilde{x} = 0$, $\tilde{\theta} = [\theta_0 \theta_1 \dots \theta_n]^T \Rightarrow \theta = (X^T X)^{-1} X^T y$

Ιδιότητες: ευαισθησία σε outliers

$$x = [1 \ x_1 \ \dots \ x_n]^T$$

χρησιμ για γραμμικά δεδομένα

θεωρεί ότι τα σφάλμ ακολουθούν καν κατανομή

δεν λειτουργεί καλά σε υψηλές διασασει, διαφ ηκυκωσεται

▷ Γραμ διακω: δίνει σωστα διακω γραμμή / μπορεί να μην βρει λύση

▷ Οι γραμ διακ: δίνει ευθεία με κατά ανασελ (συνήθως)

Perceptron

(μεγ αριθμος στοιχείων που μπορούν να κατακερματιστούν

Σε κωρο διασε $d \Rightarrow$ διάσαση $VC = d+1$

AND, OR $\rightarrow$ ανλή $\sim$ γραμ

▷ όχι γραμ: έω συκλ ντε

XOR $\rightarrow$ 1 κρυφό (ex ReLU, sigmoid) $\sim$ μη χρ

▷ γραμ: συκλ σε νηερ/αγνωστο αριθμ n

Χαρακτηριστικό	Perceptron	Adaline
Γραμμικά διαχωρίσιμα προβλήματα	Πάντα βρίσκει λύση	Μπορεί να μην συγκλίνει
Μη γραμμικά διαχωρίσιμα προβλήματα	Δεν συγκλίνει	Μπορεί να συγκλίνει με επιτρεπτές λανθασμένες ταξινομήσεις
Συγκλίνει όταν	Τα δεδομένα είναι γραμμικά διαχωρίσιμα	Τα δεδομένα είναι σχεδόν γραμμικά διαχωρίσιμα (κάτω από σφάλμα ϵ)
Αντιμετώπιση λαθών	Δεν επιτρέπει λάθη	Επιτρέπει μικρό ποσοστό σφάλματος
Αλγόριθμος εκπαίδευσης	Βασίζεται σε ενημέρωση βαρών με απόλυτες ταξινωμικές εξόδους	Βασίζεται σε ελαχιστοποίηση του τετραγωνικού σφάλματος

ADALINE

▷ όχι γραμ συνάρτηση ενεργ

▷ έξοδος $\rightarrow$ ανεξείς τιμές

Stochastic gradient descent (SGD)

- Σζην ακραία περίπτωση που το μέγεθος του batch $B=1$, έχουμε ανανέωση ζων βαρών μεζά από ζην παρουσίαση ενός μόνο προζύπου.
- Η ακραία περίπτωση $B=N$ αντιστοιχεί πρακτικά σζην κλασική κατάβαση δυναμικού
- Για ποιο λόγο μπορεί να είναι προζιμότερη η SGD έναντι της GD;
- Σε μεγάλα datasets, είναι υπολογιστικά ακριβό να ζίνεζαι ανανέωση ζων βαρών μεζά από ζην παρουσίαση του συνόλου ζων δεδομένων
- Επιπλέον, η δειγματοληψία δεδομένων εισάγει καζά έναν ζρόπο «θόρυβο», ο οποίος αποζρέπτει ζην εύρεση «ρηχών» ζοπικών ελαχίστων, ζεζονός που είναι καλό για ζην βελτιστοποίηση μη-κυρζών συναρζήσεων

Ο αλζόριθμος back-propagation δεν βρίσκει πάντα λύση (παρόζι με βάση το ζεώρημα προσζζχισης πάντα υπάρζει)

Δεν προσεζχιζεί πάντα ζη συνάρτηση εισόδου-εξόδου με σφάλμα μικρότερο του ϵ

Όσο πιο σύνθετη η δομή του δικζού, ζόσο πιο πολλές οι πιθανόζηζες να βρεθεί λύση (σε κάποιον αριθμό εποχών)

Δεν είναι όμως απαραίζιζτο ότι η λύση αυζή θα είναι πρακτικά καλή (πρόβλημα overfitting)

Ο αλζόριθμος back-propagation δεν βρίσκει πάντα ζη λύση που δίνει το μικρότερο ζεζραζωνικό σφάλμα

Εζαρζάται από πολλούς παράζονζες όπως η αρχικοποίηση (πρόβλημα ζοπικών ελαχίστων)

MLP (Multi Layer Perceptron)

Επιλογή αρχικών βαρών:

- ⊕ Μεγάλα αρχικά βάρη $\rightarrow$ αργή σύγκλιση } λύση: μικρές τυχαίες τιμές
- ⊕ Ίδια βάρη $\rightarrow$ κακή εκμ. ποιότητα / συμμετρ.
- ⊕ Πολλαπλές διαφορετικές εκπαίδ. $\rightarrow$ αποφυγή τοπικών ελαχίστων
- ⊖ μπορεί να παγιδευτεί σε τοπικά ελάχιστα
- ⊖ υψηλό υπολ. κόστος Θαργή εκπαίδευση

Support Vector Machines (SVM)

▷ Αναφορικά από το $\left\{ \begin{array}{l} \text{κοντα στο} \\ \text{αυτο δέξ ένα σημείο} \end{array} \right\}$ $\left\{ \begin{array}{l} \text{μακριά από το} \\ \text{τότε αυτο} \end{array} \right\}$ όριο απόφασης $\left\{ \begin{array}{l} \text{εμπνείζεται} \\ \text{ΔΕΝ εμπνείζεται} \end{array} \right\}$

↳ μόνο σημεία του support vector εμπνείζουν το όριο απόφασης

▷ # διαδ. ΣΜ για k κλάσεις $\rightarrow$ one-against-one: $N = k(k-1)/2$ $\rightarrow$ special case: $k=3 \Rightarrow N=3$
 $\rightarrow$ one-against-all: $N=k$

▷ Με χρήση kernel functions γίνεται απεικόνιση δεδομένων σε μη γραμ. χώρο, χωρίς να εμφανίζεται ποτέ η συνάρτηση $M(x, x')$ μόνη της

▷ Εισαγωγή Slack-Variables $f_i \Rightarrow$ $\left\{ \begin{array}{l} \text{εύρεση λύσης σε μη γραμ. διαχωρ. κλάσεις} \\ \text{επιτρέπει λανθασμένες ταξινόμησης} \end{array} \right.$
 $0 < f_i \leq 1$: όχι λάθος
 $f_i > 1$: λάθος

$\min_{w,b,\gamma} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N f_i$ C : trade-off μεταξύ μεγιστ. margin και ελαχιστ. σφάλμα

Μεγάλο C : ελαχιστ. σφάλματα ταξινόμησης $\rightarrow$ πιθανό overfitting $\rightarrow$ κακό για δέξ με θόρυβο

Μικρό C : μεγαλύτερο margin $\rightarrow$ πιθανά σφάλμα / καλύτερη γενίκευση $\rightarrow$ καλό για δέξ με θόρυβο

⊕ Δεν παγιδεύεται σε τοπικά min

⊖ μεγάλο υπολογιστικό κόστος

⊕ Αποσπ. σε υψηλές διαστάσεις

⊖ ευαίσθ. σε επιλογή kernel function, C

⊕ Ανθεκτικό σε overfitting

⊕ Χειρίζεται μη γραμ. δεδομένα με kernel funct.

ϵ -Greedy

▷ Επιλέγει: $\left\{ \begin{array}{l} \text{ενέργει με max avg με } P=1-\epsilon \rightarrow \text{εξετάζουμε} \\ \text{τυχαία ενέργεια με } P=\epsilon \rightarrow \text{εξερεύνηση} \end{array} \right.$

▷ $\epsilon=0$: ϵ -greedy $\equiv$ greedy (μόνο εκμετ)

▷ Ενδέχεται η εξερεύνηση να μην βελτιώσει επιδόση

Bayes: βέλτιστη λύση, αλλά απαιτεί σην των κλάσεων $\rightarrow$ δυσκ για μεγάλες διαστ
▷ ελλιπής κατανομή δει $\rightarrow$ αναπαράχ μοντελοποίηση

k-NN:

▷ μη παραμετρικός ταξιν

▷ όχι εκπαίδευση

▷ supervised

Λογιστική Γραμμοδρόμηση

▷ δει μας ελλιπής καταν δεδομ $\rightarrow$ διακριτική μοντελο

▷ είναι classification, όχι regression

PAC-learning: ικανότητα ταξινόμησης να γενικεύει

① VC-dimension $< \infty$

② $\exists$ αλγ εκπαίδευσης που διακρ στοιχεία με $P=1-\delta$ και σφάλμα $< \epsilon$

③ πολλών προτύπων εκπαίδευσης ως προς μέγεθος δεδομ

Επαγωγική μεροληψία: ικανότητα γενίκευσης πέρα από το training set

αποτελεί: ① σύνολο υποδ ② ίδαν ταξινόμηση ③ δει εκπαίδευσης

Decision Tree

naive: τυχαία επιλογή

ID3: min entropy: $f(x) = -\sum p_i \log p_i$

CART: min gini: $Gini(x) = 1 - \sum p_i^2$

χρήση βαθμού εμπιστοσ:

▷ κάθε γενίκευση

▷ μικρότερο μέγεθος

▷ κάποιες λάθος ταξιν

▷ ID3 πιο αποδ, χωρίς ιδιαίτ διαφοροποίηση σην επιλογή παρατηρησιακού απόφασης

▷ $\exists$ πολλά διαφορετικά δέντρα που κάνουν σωστά την ίδια ταξινόμηση

▷ εύρεση ελάχιστοι NP-complete

▷ εκθετικά πολλά διαφ δέντρα / διαφ κόμβους

PCA (Principal Component Analysis)

- ▷ Unsupervised $\rightarrow$ χωρίς labels
- ▷ αναζητά κατεύθ με μέγιστη διασπορά
- ▷ #principal components $\leq$ # αρχικά παρατηρησιακών
- ▷ ορθογώνιες συνιστώσες

Κανονικόν:

- ▷ όλα παρατηρη έχουν ίδια συνεισφορά στην ανάλυση

LDA (Linear Discriminant Analysis)

- ▷ Supervised $\rightarrow$ με labels
- ▷ Είδηση υπόχωρου: between-class scatter / διαζήτη διασπορά $\sim \max S_b$
within class scatter / ενδοζήτη διασπορά $\sim \min S_w$

	PCA	LDA
Σκοπός	μεγιστοποίηση μεταβλητότητας Unsupervised	μεγιστοποίηση διακρίσιμης ταξέων Supervised
Χρήση	data compression, visualization	ταξινόμηση, διαχωρ κατηγοριών
#view διαδ	$\leq m$, $m = \text{\#παρατηρησιακών}$	$\leq C-1$, $C = \text{\#κατηγοριών}$

Supervised $\rightarrow$ classification to 1
 $\rightarrow$ regression R

Unsupervised $\rightarrow$ clustering
 $\rightarrow$ dimensionality reduction

BSAS:

- Η διάταξη με την οποία τα διανύσματα παρουσιάζονται επηρεάζει σημαντικά τα αποτελέσματα της ομαδοποίησης
- Τα αποτελέσματα του αλγόριθμου επηρεάζονται σημαντικά από την επιλογή της τιμής του κατωφλίου θ
- Αν ο μέγιστος επιτρεπτός αριθμός ομάδων, q , δεν ορίζεται, αφήνουμε τον αλγόριθμο να «αποφασίσει»
- Ο BSAS αποτυγχάνει για μη καλά διαχωρισμένες ομάδες
- Όταν κάθε ομάδα αντιπροσωπεύεται από ένα διάνυσμα, ο BSAS ευνοεί τη δημιουργία συμπαγών ομάδων
- Η χρονική πολυπλοκότητα του BSAS είναι $O(N)$

k-means

- ▷ Δεν απαιτεί εκπαίδευση, hard clustering, division k and n in
- ▷ Supervised $\rightarrow$ απαιτεί labels
- ▷ Ξύλο αποφασισ δεν είναι απαραίτητα γραμμικό
- ⊖ μεγάλη υπολογιστική πολυπλοκότητα ⊕ αναδεικνύει συμπαγών ομάδων
- ⊖ ευαισθησία σε θόρυβο, outliers ⊕ χαμηλό complexity $O(mn)$ per loop
- ⊖ κρίσιμο k : μεγ $k \sim$ underfitting, μικρό $k \sim$ overfitting
- ⊖ κρίσιμη αρχικοποίηση (πυκνότητα ή απόσταση)

DBSCAN:

- ⊕ Δυναμικός καθορισμός clusters
- ⊕ Διαχειρίζεται αποτελεσματικά outliers
- ⊕ Εύελιτοι σε μη γραμ σχήματα
- ⊖ Δυσκολεύεται σε πολλές πυκνότητες
- ⊖ Δυσκολεύεται σε πολλές διαστάσεις
- ⊖ Ευαισθησία σε παραμ ϵ , minPts
- ▷ ένα σημείο θεωρείται πυρήνας αν τουλάχιστον minPts σημ βρίσκονται σε ακτίνα ϵ
- ▷ Complexity: $O(N \log N)$ με δάτυρο

Ιδιότητα	k-means	DBSCAN
Υποθέσεις για τις κλάσεις	Οι κλάσεις είναι σφαιρικές και ισομεγέθεις.	Δεν υπάρχουν υποθέσεις για το σχήμα των κλάσεων.
Ευαισθησία σε outliers	Πολύ ευαίσθητος. Τα outliers μπορούν να αλλοιώσουν σημαντικά τα κέντρα των κλάσεων.	Ανθεκτικός σε outliers, καθώς τα αγνοεί όταν δεν ανήκουν σε περιοχές υψηλής πυκνότητας.
Μέθοδος λειτουργίας	Ομαδοποιεί τα δεδομένα γύρω από κεντρικά σημεία (centroids) με βάση την ελαχιστοποίηση της απόστασης.	Βασίζεται στην πυκνότητα των σημείων σε μια περιοχή για να εντοπίσει κλάσεις.
Πλήθος κλάσεων (clusters)	Χρειάζεται να δοθεί εκ των προτέρων ο αριθμός των κλάσεων.	Δεν απαιτεί προκαθορισμένο αριθμό κλάσεων, καθώς τις προσδιορίζει δυναμικά.
Ευαισθησία στην αρχικοποίηση	Πολύ ευαίσθητος στην αρχικοποίηση των κέντρων.	Όχι ευαίσθητος σε αρχικοποίηση.
Ευαισθησία σε θόρυβο	Ευαίσθητος σε θορυβώδη δεδομένα.	Μπορεί να αγνοήσει σημεία θορύβου (outliers).
Πυκνότητα δεδομένων	Δεν λαμβάνει υπόψη την πυκνότητα δεδομένων.	Εστιάζει στην πυκνότητα δεδομένων και ανιχνεύει αραιές και πυκνές περιοχές.
Κατάλληλότητα για μεγάλα δεδομένα	Γρήγορος και κατάλληλος για μεγάλα σύνολα δεδομένων.	Μπορεί να γίνει αργός για μεγάλα σύνολα δεδομένων με υψηλές διαστάσεις.

