Το A”alto
Λεξικό της Μηχανικής Μάθησης

Alexander Jung και Konstantina Olioumtsevits

May 16, 2025

αναφορά ως: A. Jung and K. Olioumtsevits, *Το Aalto Λεξικό της Μηχανικής Μάθησης* [The Aalto Dictionary of Machine Learning]. Espoo, Finland: Aalto University, 2025.

Ευχαριστίες

Αυτό το λεξικό της μηχανικής μάθησης αναπτύχθηκε κατά τον σχεδιασμό και την υλοποίηση διαφορετικών μαθημάτων, συμπεριλαμβανομένων των CS-E3210 Machine Learning: Basic Principles, CS-C3240 Machine Learning, CS-E4800 Artificial Intelligence, CS-EJ3211 Machine Learning with Python, CS-EJ3311 Deep Learning with Python, CS-E4740 Federated Learning, και CS-E407507 Human-Centered Machine Learning. Αυτά τα μαθήματα προσφέρονται στο Aalto University `https://www.aalto.fi/en`, σε ενήλικους/ες σπουδαστές/σπουδάστριες μέσω του The Finnish Institute of Technology (FITech) `https://fitech.io/en/`, και σε διεθνείς φοιτητές/φοιτήτριες μέσω της European University Alliance Unite! `https://www.aalto.fi/en/unite`.

Είμαστε ευγνώμονες στους/στις σπουδαστές/σπουδάστριες που παρείχαν πολύτιμα σχόλια που ήταν καθοριστικά για το συγκεκριμένο λεξικό. Ιδιαίτερες ευχαριστίες στον Mikko Seesto για τη σχολαστική του διόρθωση προσχεδίων.

Κατάλογοι Συμβόλων

Σύνολα και Συναρτήσεις

$a \in \mathcal{A}$	Το αντικείμενο a είναι ένα στοιχείο του συνόλου $\mathcal{A}$.
$a := b$	Χρησιμοποιούμε το a ως συντομογραφία για το b.
$\lvert\mathcal{A}\rvert$	Η καρδινικότητα (δηλαδή ο αριθμός των στοιχείων) ενός πεπερασμένου συνόλου $\mathcal{A}$.
$\mathcal{A} \subseteq \mathcal{B}$	Το $\mathcal{A}$ είναι ένα υποσύνολο του $\mathcal{B}$.
$\mathcal{A} \subset \mathcal{B}$	Το $\mathcal{A}$ είναι ένα αυστηρό υποσύνολο του $\mathcal{B}$.
$\mathbb{N}$	Οι φυσικοί αριθμοί $1, 2, \ldots$.
$\mathbb{R}$	Οι πραγματικοί αριθμοί x [1].
$\mathbb{R}_+$	Οι μη αρνητικοί πραγματικοί αριθμοί $x \geq 0$.
$\mathbb{R}_{++}$	Οι θετικοί πραγματικοί αριθμοί $x > 0$.
$\{0, 1\}$	Το σύνολο που αποτελείται από τους δύο πραγματικούς αριθμούς 0 και 1.
$[0, 1]$	Το κλειστό διάστημα των πραγματικών αριθμών x με $0 \leq x \leq 1$.

$\operatorname*{argmin}_{\mathbf{w}} f(\mathbf{w})$	Το σύνολο των ελαχιστοποιητών για μια συνάρτηση πραγματικής τιμής $f(\mathbf{w})$.
$\mathbb{S}^{(n)}$	Το σύνολο των διανυσμάτων μοναδιαίας νόρμας στο $\mathbb{R}^{n+1}$. Βλέπε επίσης: νόρμα.
$\log a$	Ο λογάριθμος του θετικού αριθμού $a \in \mathbb{R}_{++}$.
$h(\cdot) : \mathcal{A} \to \mathcal{B} : a \mapsto h(a)$	Μία συνάρτηση (απεικόνιση) που δέχεται οποιοδήποτε στοιχείο $a \in \mathcal{A}$ από ένα σύνολο $\mathcal{A}$ ως εισαγόμενο και δίνει ένα καλά ορισμένο στοιχείο $h(a) \in \mathcal{B}$ ενός συνόλου $\mathcal{B}$. Το σύνολο $\mathcal{A}$ είναι το πεδίο της συνάρτησης h και το σύνολο $\mathcal{B}$ είναι το πεδίο τιμών της h. Ο στόχος της μηχανικής μάθησης είναι να βρει (ή να μάθει) μία συνάρτηση h (δηλαδή μία υπόθεση) που διαβάζει τα χαρακτηριστικά $\mathbf{x}$ ενός σημείου δεδομένων και δίνει μία πρόβλεψη $h(\mathbf{x})$ για την ετικέτα y του. Βλέπε επίσης: ml, υπόθεση, χαρακτηριστικό, σημείο δεδομένων, πρόβλεψη, ετικέτα.
$\nabla f(\mathbf{w})$	Η κλίση μίας παραγωγίσιμης συνάρτησης πραγματικής τιμής $f : \mathbb{R}^d \to \mathbb{R}$ είναι το διάνυσμα $\nabla f(\mathbf{w}) = \left(\frac{\partial f}{\partial w_1}, \ldots, \frac{\partial f}{\partial w_d}\right)^T \in \mathbb{R}^d$ [2, Κεφ. 9]. Βλέπε επίσης: κλίση, παραγωγίσιμη.

Πίνακες και Διανύσματα

$\mathbf{x} = \big(x_1, \ldots, x_d\big)^T$	A vector of length d, with its j-th entry being x_j.
$\mathbb{R}^d$	The set of vectors $\mathbf{x} = \big(x_1, \ldots, x_d\big)^T$ consisting of d real-valued entries $x_1, \ldots, x_d \in \mathbb{R}$.
$\mathbf{I}_{l\times d}$	A generalized identity matrix with l rows and d columns. The entries of $\mathbf{I}_{l\times d} \in \mathbb{R}^{l\times d}$ are equal to 1 along the main diagonal and equal to 0 otherwise.
$\mathbf{I}_d, \mathbf{I}$	A square identity matrix of size $d \times d$. If the size is clear from context, we drop the subscript.
$\|\mathbf{x}\|_2$	The Euclidean (or ℓ_2) νόρμα of the vector $\mathbf{x} = \big(x_1, \ldots, x_d\big)^T \in \mathbb{R}^d$ defined as $\|\mathbf{x}\|_2 := \sqrt{\sum_{j=1}^d x_j^2}$.
$\|\mathbf{x}\|$	Some νόρμα of the vector $\mathbf{x} \in \mathbb{R}^d$ [3]. Unless specified otherwise, we mean the Euclidean νόρμα $\|\mathbf{x}\|_2$.
$\mathbf{x}^T$	The transpose of a matrix that has the vector $\mathbf{x} \in \mathbb{R}^d$ as its single column.
$\mathbf{X}^T$	The transpose of a matrix $\mathbf{X} \in \mathbb{R}^{m\times d}$. A square real-valued matrix $\mathbf{X} \in \mathbb{R}^{m\times m}$ is called symmetric if $\mathbf{X} = \mathbf{X}^T$.
$\mathbf{0} = \big(0, \ldots, 0\big)^T$	The vector in $\mathbb{R}^d$ with each entry equal to zero.
$\mathbf{1} = \big(1, \ldots, 1\big)^T$	The vector in $\mathbb{R}^d$ with each entry equal to one.

$\left(\mathbf{v}^T, \mathbf{w}^T\right)^T$	The vector of length $d + d'$ obtained by concatenating the entries of vector $\mathbf{v} \in \mathbb{R}^d$ with the entries of $\mathbf{w} \in \mathbb{R}^{d'}$.
$\text{span}\{\mathbf{B}\}$	The span of a matrix $\mathbf{B} \in \mathbb{R}^{a \times b}$, which is the subspace of all linear combinations of the columns of $\mathbf{B}$, $\text{span}\{\mathbf{B}\} = \left\{\mathbf{Ba} : \mathbf{a} \in \mathbb{R}^b\right\} \subseteq \mathbb{R}^a$.
$\det(\mathbf{C})$	The determinant of the matrix $\mathbf{C}$.
$\mathbf{A} \otimes \mathbf{B}$	The Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ [4].

Θεωρία Πιθανοτήτων

$\mathbb{E}_p\{f(\mathbf{z})\}$	The προσδοκία of a function $f(\mathbf{z})$ of a random variable (RV) $\mathbf{z}$ whose κατανομή πιθανότητας is $p(\mathbf{z})$. If the κατανομή πιθανότητας is clear from context, we just write $\mathbb{E}\{f(\mathbf{z})\}$.
$p(\mathbf{x}, y)$	A (joint) κατανομή πιθανότητας of an RV whose realizations are data points with features $\mathbf{x}$ and ετικέτα y.
$p(\mathbf{x}\|y)$	A conditional κατανομή πιθανότητας of an RV $\mathbf{x}$ given the value of another RV y [5, Sec. 3.5].
$p(\mathbf{x}; \mathbf{w})$	A parametrized κατανομή πιθανότητας of an RV $\mathbf{x}$. The κατανομή πιθανότητας depends on a parameter vector $\mathbf{w}$. For example, $p(\mathbf{x}; \mathbf{w})$ could be a πολυμεταβλητή κανονική κατανομή with the parameter vector $\mathbf{w}$ given by the entries of the μέση τιμή vector $\mathbb{E}\{\mathbf{x}\}$ and the πίνακας συνδιακύμανσης $\mathbb{E}\Big\{\big(\mathbf{x} - \mathbb{E}\{\mathbf{x}\}\big)\big(\mathbf{x} - \mathbb{E}\{\mathbf{x}\}\big)^T\Big\}$.
$\mathcal{N}(\mu, \sigma^2)$	The κατανομή πιθανότητας of a Gaussian random variable (Gaussian RV) $x \in \mathbb{R}$ with μέση τιμή (or expectation) $\mu = \mathbb{E}\{x\}$ and διακύμανση $\sigma^2 = \mathbb{E}\big\{(x-\mu)^2\big\}$.
$\mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$	The πολυμεταβλητή κανονική κατανομή of a vector-valued Gaussian RV $\mathbf{x} \in \mathbb{R}^d$ with μέση τιμή (or expectation) $\boldsymbol{\mu} = \mathbb{E}\{\mathbf{x}\}$ and πίνακας συνδιακύμανσης $\mathbf{C} = \mathbb{E}\big\{\big(\mathbf{x} - \boldsymbol{\mu}\big)\big(\mathbf{x} - \boldsymbol{\mu}\big)^T\big\}$.

Μηχανική Μάθηση

r	An index $r = 1, 2, \ldots$ that enumerates data points.
m	The number of data points in (i.e., the size of) a σύνολο δεδομένων.
$\mathcal{D}$	A σύνολο δεδομένων $\mathcal{D} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}\}$ is a list of individual data points $\mathbf{z}^{(r)}$, for $r = 1, \ldots, m$.
d	The number of features that characterize a data point.
x_j	The j-th feature of a data point. The first feature is denoted x_1, the second feature x_2, and so on.
$\mathbf{x}$	The feature vector $\mathbf{x} = \big(x_1, \ldots, x_d\big)^T$ of a data point whose entries are the individual features of a data point.
$\mathcal{X}$	The feature space $\mathcal{X}$ is the set of all possible values that the features $\mathbf{x}$ of a data point can take on.
$\mathbf{z}$	Instead of the symbol $\mathbf{x}$, we sometimes use $\mathbf{z}$ as another symbol to denote a vector whose entries are the individual features of a data point. We need two different symbols to distinguish between raw and learned features [6, Ch. 9].
$\mathbf{x}^{(r)}$	The feature vector of the r-th data point within a σύνολο δεδομένων.
$x_j^{(r)}$	The j-th feature of the r-th data point within a σύνολο δεδομένων.

$\mathcal{B}$	A mini-δέσμη (or subset) of randomly chosen data points.
B	The size of (i.e., the number of data points in) a mini-δέσμη.
y	The ετικέτα (or quantity of interest) of a data point.
$y^{(r)}$	The ετικέτα of the r-th data point.
$\left(\mathbf{x}^{(r)}, y^{(r)}\right)$	The features and ετικέτα of the r-th data point.
$\mathcal{Y}$	The label space $\mathcal{Y}$ of an ml method consists of all potential ετικέτα values that a data point can carry. The nominal label space might be larger than the set of different ετικέτα values arising in a given σύνολο δεδομένων (e.g., a σύνολο εκπαίδευσης). Ml problems (or methods) using a numeric label space, such as $\mathcal{Y} = \mathbb{R}$ or $\mathcal{Y} = \mathbb{R}^3$, are referred to as regression problems (or methods). Ml problems (or methods) that use a discrete label space, such as $\mathcal{Y} = \{0, 1\}$ or $\mathcal{Y} = \{cat, dog, mouse\}$, are referred to as ταξινόμηση problems (or methods).
η	Learning rate (or step size) used by gradient-based methods.
$h(\cdot)$	A υπόθεση map that reads in features $\mathbf{x}$ of a data point and delivers a πρόβλεψη $\hat{y} = h(\mathbf{x})$ for its ετικέτα y.

$\mathcal{Y}^{\mathcal{X}}$	Given two sets $\mathcal{X}$ and $\mathcal{Y}$, we denote by $\mathcal{Y}^{\mathcal{X}}$ the set of all possible υπόθεση maps $h : \mathcal{X} \to \mathcal{Y}$.
$\mathcal{H}$	A χώρος υποθέσεων or μοντέλο used by an ml method. The χώρος υποθέσεων consists of different υπόθεση maps $h : \mathcal{X} \to \mathcal{Y}$, between which the ml method must choose.
$d_{\text{eff}}(\mathcal{H})$	The effective dimension of a χώρος υποθέσεων $\mathcal{H}$.
B^2	The squared μεροληψία of a learned υπόθεση $\hat{h}$, or its parameters. Note that $\hat{h}$ becomes a RV if it is learned from data points being RVs.
V	The διακύμανση of a learned υπόθεση $\hat{h}$, or its parameters. Note that $\hat{h}$ becomes a RV if it is learned from data points being RVs.
$L((\mathbf{x}, y), h)$	The απώλεια incurred by predicting the ετικέτα y of a data point using the πρόβλεψη $\hat{y} = h(\mathbf{x})$. The πρόβλεψη $\hat{y}$ is obtained by evaluating the υπόθεση $h \in \mathcal{H}$ for the feature vector $\mathbf{x}$ of the data point.
E_v	The σφάλμα επικύρωσης of a υπόθεση h, which is its average loss incurred over a σύνολο επικύρωσης.

$\widehat{L}\big(h\vert\mathcal{D}\big)$	The εμπειρική διακινδύνευση or average loss incurred by the υπόθεση h on a σύνολο δεδομένων $\mathcal{D}$.
E_t	The training error of a υπόθεση h, which is its average loss incurred over a σύνολο εκπαίδευσης.
t	A discrete-time index $t = 0, 1, \ldots$ used to enumerate sequential events (or time instants).
t	An index that enumerates εργασία εκμάθησηςs within a εκμάθηση πολυδιεργασίας problem.
α	A regularization parameter that controls the amount of regularization.
$\lambda_j\big(\mathbf{Q}\big)$	The j-th ιδιοτιμή (sorted in either ascending or descending order) of a positive semi-definite (psd) matrix $\mathbf{Q}$. We also use the shorthand λ_j if the corresponding matrix is clear from context.
$\sigma(\cdot)$	The συνάρτηση ενεργοποίησης used by an artificial neuron within an τεχνητό νευρωνικό δίκτυο.
$\mathcal{R}_{\hat{y}}$	A περιοχή αποφάσεων within a feature space.
$\mathbf{w}$	A parameter vector $\mathbf{w} = \big(w_1, \ldots, w_d\big)^T$ of a model, e.g., the βάρη of a γραμμικό μοντέλο or in an τεχνητό νευρωνικό δίκτυο.
$h^{(\mathbf{w})}(\cdot)$	A υπόθεση map that involves tunable παράμετροι μοντέλου $w_1, \ldots, w_d$ stacked into the vector $\mathbf{w} = \big(w_1, \ldots, w_d\big)^T$.

$\phi(\cdot)$	A feature map $\phi : \mathcal{X} \to \mathcal{X}' : \mathbf{x} \mapsto \mathbf{x}' := \phi\big(\mathbf{x}\big) \in \mathcal{X}'$.
$K\big(\cdot, \cdot\big)$	Given some feature space $\mathcal{X}$, a kernel is a map $K : \mathcal{X} \times \mathcal{X} \to \mathbb{C}$ that is psd.

Federated Learning

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	An undirected γράφος whose nodes $i \in \mathcal{V}$ represent συσκευήs within a federated learning network (FL network). The undirected weighted edges $\mathcal{E}$ represent connectivity between συσκευήs and statistical similarities between their σύνολο δεδομένωνs and εργασία εκμάθησηςs.
$i \in \mathcal{V}$	A node that represents some συσκευή within an FL network. The device can access a τοπικό σύνολο δεδομένων and train a local model.
$\mathcal{G}^{(\mathcal{C})}$	The induced subgraph of $\mathcal{G}$ using the nodes in $\mathcal{C} \subseteq \mathcal{V}$.
$\mathbf{L}^{(\mathcal{G})}$	The Laplacian matrix of a graph $\mathcal{G}$.
$\mathbf{L}^{(\mathcal{C})}$	The Laplacian matrix of the induced graph $\mathcal{G}^{(\mathcal{C})}$.
$\mathcal{N}^{(i)}$	The neighborhood of a node i in a graph $\mathcal{G}$.
$d^{(i)}$	The weighted degree $d^{(i)} := \sum_{i' \in \mathcal{N}^{(i)}} A_{i,i'}$ of a node i in a graph $\mathcal{G}$.
$d^{(\mathcal{G})}_{\max}$	The maximum weighted node degree of a graph $\mathcal{G}$.
$\mathcal{D}^{(i)}$	The τοπικό σύνολο δεδομένων $\mathcal{D}^{(i)}$ carried by node $i \in \mathcal{V}$ of an FL network.
m_i	The number of data points (i.e., μέγεθος δείγματος) contained in the τοπικό σύνολο δεδομένων $\mathcal{D}^{(i)}$ at node $i \in \mathcal{V}$.

$\mathbf{x}^{(i,r)}$	The features of the r-th data point in the τοπικό σύνολο δεδομένων $\mathcal{D}^{(i)}$.
$y^{(i,r)}$	The ετικέτα of the r-th data point in the τοπικό σύνολο δεδομένων $\mathcal{D}^{(i)}$.
$\mathbf{w}^{(i)}$	The local παράμετροι μοντέλου of συσκευή i within an FL network.
$L_i\left(\mathbf{w}\right)$	The local συνάρτηση απώλειας used by συσκευή i to measure the usefulness of some choice $\mathbf{w}$ for the local παράμετροι μοντέλου.
$L^{(\mathrm{d})}\left(\mathbf{x}, h\big(\mathbf{x}\big), h'\big(\mathbf{x}\big)\right)$	The loss incurred by a υπόθεση h' on a data point with features $\mathbf{x}$ and ετικέτα $h\big(\mathbf{x}\big)$ that is obtained from another υπόθεση.
$\mathrm{stack}\big\{\mathbf{w}^{(i)}\big\}_{i=1}^{n}$	The vector $\left(\big(\mathbf{w}^{(1)}\big)^T, \ldots, \big(\mathbf{w}^{(n)}\big)^T\right)^T \in \mathbb{R}^{dn}$ that is obtained by vertically stacking the local παράμετροι μοντέλου $\mathbf{w}^{(i)} \in \mathbb{R}^d$.

Έννοιες Μηχανικής Μάθησης

αβεβαιότητα Uncertainty refers to the degree of confidence—or lack thereof—associated with a quantity such as a model πρόβλεψη, parameter estimate, or observed data point. In ml, uncertainty arises from various sources, including noisy δεδομένα, limited training δείγμας, or ambiguity in model assumptions. Probability theory offers a principled framework for representing and quantifying such uncertainty.

Βλέπε επίσης: model, πρόβλεψη, data point, ml, data, δείγμα, probability.

αισιοδοξία παρά την αβεβαιότητα ml methods learn παράμετροι μοντέλου $\mathbf{w}$ according to some performance criterion $\bar{f}(\mathbf{w})$. However, they usually cannot access $\bar{f}(\mathbf{w})$ directly but rely on an estimate (or approximation) $f(\mathbf{w})$ of $\bar{f}(\mathbf{w})$. As a case in point, εμπειρική ελαχιστοποίηση διακινδύνευσης-based methods use the average loss on a given σύνολο δεδομένων (i.e., the σύνολο εκπαίδευσης) as an estimate for the διακινδύνευση of a υπόθεση. Using a πιθανοτικό μοντέλο, one can construct a confidence interval $\left[l^{(\mathbf{w})}, u^{(\mathbf{w})}\right]$ for each choice $\mathbf{w}$ for the παράμετροι μοντέλου. One simple construction is $l^{(\mathbf{w})} := f(\mathbf{w}) - \sigma/2$, $u^{(\mathbf{w})} := f(\mathbf{w}) + \sigma/2$, with σ being a measure of the (expected) deviation of $f(\mathbf{w})$ from $\bar{f}(\mathbf{w})$. We can also use other constructions for this interval as long as they ensure that $\bar{f}(\mathbf{w}) \in \left[l^{(\mathbf{w})}, u^{(\mathbf{w})}\right]$ with a sufficiently high probability. An optimist chooses the παράμετροι μοντέλου according

to the most favorable - yet still plausible - value $\tilde{f}(\mathbf{w}) := l^{(\mathbf{w})}$ of the performance criterion. Two examples of ml methods that use such an optimistic construction of an objective function are structural risk minimization (SRM) [7, Ch. 11] and άνω φράγμα εμπιστοσύνης methods for sequential decision making [8, Sec. 2.2].

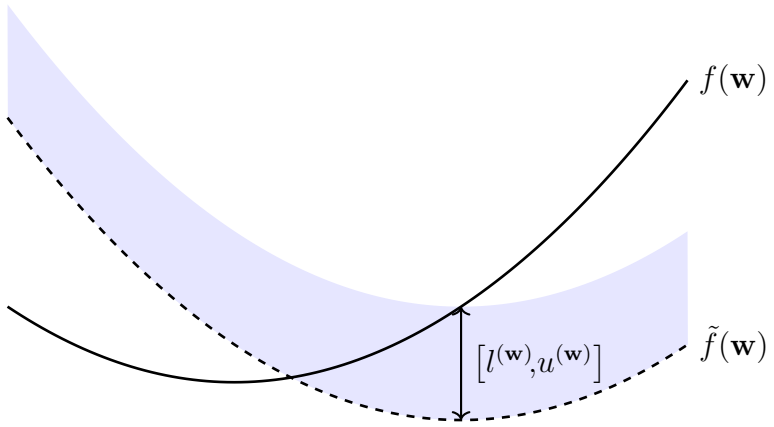


Fig. 1. ml methods learn παράμετροι μοντέλου $\mathbf{w}$ by using some estimate of $f(\mathbf{w})$ for the ultimate performance criterion $\bar{f}(\mathbf{w})$. Using a πιθανοτικό μοντέλο, one can use $f(\mathbf{w})$ to construct confidence intervals $\left[l^{(\mathbf{w})}, u^{(\mathbf{w})}\right]$ which contain $\bar{f}(\mathbf{w})$ with high probability. The best plausible performance measure for a specific choice $\mathbf{w}$ of παράμετροι μοντέλου is $\tilde{f}(\mathbf{w}) := l^{(\mathbf{w})}$.

Βλέπε επίσης: ml, παράμετροι μοντέλου, εμπειρική ελαχιστοποίηση διακινδύνευσης, loss, σύνολο δεδομένων, σύνολο εκπαίδευσης, διακινδύνευση, υπόθεση, πιθανοτικό μοντέλο, probability, objective function, SRM, άνω φράγμα εμπιστοσύνης.

ακρίβεια Consider data points characterized by features $\mathbf{x} \in \mathcal{X}$ and a cate-

gorical label y which takes on values from a finite label space $\mathcal{Y}$. The accuracy of a υπόθεση $h : \mathcal{X} \to \mathcal{Y}$, when applied to the data points in a σύνολο δεδομένων $\mathcal{D} = \left\{\left(\mathbf{x}^{(1)}, y^{(1)}\right), \ldots, \left(\mathbf{x}^{(m)}, y^{(m)}\right)\right\}$, is then defined as $1 - (1/m) \sum_{r=1}^{m} L^{(0/1)}\left(\left(\mathbf{x}^{(r)}, y^{(r)}\right), h\right)$ using the 0/1 απώλεια $L^{(0/1)}(\cdot, \cdot)$.

Βλέπε επίσης: data point, feature, label space, υπόθεση, σύνολο δεδομένων, 0/1 απώλεια.

αλγόριθμος Ένας αλγόριθμος (algorithm) είναι μία ακριβής, βήμα προς βήμα προδιαγραφή για το πώς να παραχθεί ένα εξαγόμενο (output) από ένα συγκεκριμένο εισαγόμενο (input) εντός ενός πεπερασμένου αριθμού υπολογιστικών βημάτων [9]. Για παράδειγμα, ένας αλγόριθμος για την εκπαίδευση ενός γραμμικού μοντέλου περιγράφει ρητά πώς να μετασχηματιστεί ένα δεδομένο σύνολο εκπαίδευσης σε παράμετροι μοντέλου μέσω μίας ακολουθίας βημάτων κλίσης. Αυτός ο άτυπος χαρακτηρισμός μπορεί να οριστικοποιηθεί ενδελεχώς μέσω διαφορετικών μαθηματικών μοντέλων [10]. Ένα πολύ απλό μοντέλο ενός αλγόριθμου είναι μία συλλογή από πιθανές εκτελέσεις. Κάθε εκτέλεση είναι μία ακολουθία:

$$\text{input}, s_1, s_2, \ldots, s_T, \text{output}$$

που σέβεται τους εγγενείς περιορισμούς του υπολογιστή που εκτελεί τον αλγόριθμο. Οι αλγόριθμοι μπορεί να είναι ντετερμινιστικοί, όπου κάθε εισαγόμενο οδηγεί σε μία μοναδική εκτέλεση, ή τυχαίοι, όπου εκτελέσεις μπορεί να διαφέρουν πιθανολογικά. Οι τυχαίοι αλγόριθμοι μπορούν συνεπώς να αναλυθούν προβάλλοντας ακολουθίες εκτέλεσης ως αποτελέσματα τυχαίων πειραμάτων, θεωρώντας τον αλγόριθμο ως μία στοχαστική

διαδικασία [5], [11], [12]. Σημαντικά, ένας αλγόριθμος συμπεριλαμβάνει περισσότερα από μία αντιστοίχιση από ειαγόμενο σε εξαγόμενο· περιλαμβάνει επίσης τα ενδιάμεσα υπολογιστικά βήματα $s_1, \ldots, s_T$.

Βλέπε επίσης: γραμμικό μοντέλο, σύνολο εκπαίδευσης, παράμετροι μοντέλου, βήμα κλίσης, model.

αλγόριθμος k-μέσων The k-μέση τιμής αλγόριθμος (k-means) is a hard clustering method which assigns each data point of a σύνολο δεδομένων to precisely one of k different συστάδας. The method alternates between updating the συστάδα assignments (to the συστάδα with the nearest μέση τιμή) and, given the updated συστάδα assignments, re-calculating the συστάδα μέση τιμής [6, Ch. 8].

Βλέπε επίσης: μέση τιμή, αλγόριθμος, hard clustering, data point, σύνολο δεδομένων, συστάδα.

αμοιβαία πληροφορία The MI (mutual information; MI) $I(\mathbf{x}; y)$ between two RVs $\mathbf{x}$, y defined on the same probability space is given by [13]

$$I(\mathbf{x}; y) := \mathbb{E}\left\{\log \frac{p(\mathbf{x}, y)}{p(\mathbf{x})p(y)}\right\}.$$

It is a measure of how well we can estimate y based solely on $\mathbf{x}$. A large value of $I(\mathbf{x}; y)$ indicates that y can be well predicted solely from $\mathbf{x}$. This πρόβλεψη could be obtained by a υπόθεση learned by an εμπειρική ελαχιστοποίηση διακινδύνευσης-based ml method.

Βλέπε επίσης: RV, probability space, πρόβλεψη, υπόθεση, εμπειρική ελαχιστοποίηση διακινδύνευσης, ml.

αμφικλινής παλινδρόμηση Ridge regression learns the βάρη $\mathbf{w}$ of a linear υπόθεση map $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$. The quality of a particular choice for the

παράμετροι μοντέλου $\mathbf{w}$ is measured by the sum of two components. The first component is the average τετραγωνική απώλεια σφάλματος incurred by $h^{(\mathbf{w})}$ on a set of labeled datapoints (i.e., the σύνολο εκπαίδευσης). The second component is the scaled squared Euclidean νόρμα $\alpha\|\mathbf{w}\|_2^2$ with a regularization parameter $\alpha > 0$. Adding $\alpha\|\mathbf{w}\|_2^2$ to the average τετραγωνική απώλεια σφάλματος is equivalent to replacing each original data points by the realization of (infinitely many) independent and identically distributed (i.i.d.) RVs centered around these data points (see regularization).

Βλέπε επίσης: regression, βάρη, υπόθεση, παράμετροι μοντέλου, τετραγωνική απώλεια σφάλματος, labeled datapoint, σύνολο εκπαίδευσης, νόρμα, regularization, data point, realization, i.i.d., RV.

ανάλυση ιδιαζουσών τιμών The SVD (singular value decomposition; SVD) for a matrix $\mathbf{A} \in \mathbb{R}^{m \times d}$ is a factorization of the form

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{U}^T,$$

with orthonormal matrices $\mathbf{V} \in \mathbb{R}^{m \times m}$ and $\mathbf{U} \in \mathbb{R}^{d \times d}$ [3]. The matrix $\mathbf{\Lambda} \in \mathbb{R}^{m \times d}$ is only non-zero along the main diagonal, whose entries $\Lambda_{j,j}$ are non-negative and referred to as singular values.

ανάλυση ιδιοτιμών The ιδιοτιμή decomposition (eigenvalue decomposition; EVD) for a square matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a factorization of the form

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1}.$$

The columns of the matrix $\mathbf{V} = \big(\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(d)}\big)$ are the ιδιοδιάνυσμας of the matrix $\mathbf{V}$. The diagonal matrix $\mathbf{\Lambda} = \mathrm{diag}\big\{\lambda_1, \ldots, \lambda_d\big\}$ contains

the ιδιοτιμής λ_j corresponding to the ιδιοδιάνυσμας $\mathbf{v}^{(j)}$. Note that the above decomposition exists only if the matrix $\mathbf{A}$ is diagonalizable.

Βλέπε επίσης: ιδιοτιμή, ιδιοδιάνυσμα.

ανάλυση κυρίων συνιστωσών PCA (principal component analysis; PCA) determines a linear feature map such that the new features allow us to reconstruct the original features with the ολικό ελάχιστο reconstruction error [6].

Βλέπε επίσης: feature map, feature, ολικό ελάχιστο.

ανταμοιβή A reward refers to some observed (or measured) quantity that allows us to estimate the loss incurred by the πρόβλεψη (or decision) of a υπόθεση $h(\mathbf{x})$. For example, in an ml application to self-driving vehicles, $h(\mathbf{x})$ could represent the current steering direction of a vehicle. We could construct a reward from the measurements of a collision sensor that indicate if the vehicle is moving towards an obstacle. We define a low reward for the steering direction $h(\mathbf{x})$ if the vehicle moves dangerously towards an obstacle.

Βλέπε επίσης: loss, πρόβλεψη, υπόθεση, ml.

άνω φράγμα εμπιστοσύνης (ΑΦΕ) Consider an ml application that requires selecting, at each time step k, an action a_k from a finite set of alternatives $\mathcal{A}$. The utility of selecting action a_k is quantified by a numeric ανταμοιβή signal $r^{(a_k)}$. A widely used πιθανοτικό μοντέλο for this type of sequential decision-making problem is the stochastic MAB setting [8]. In this model, the ανταμοιβή $r^{(a)}$ is viewed as the realization of an RV with unknown μέση τιμή $\mu^{(a)}$. Ideally, we would always choose

the action with the largest expected ανταμοιβή $\mu^{(a)}$, but these μέση τιμήs are unknown and must be estimated from observed data. Simply choosing the action with the largest estimate $\widehat{\mu}^{(a)}$ can lead to suboptimal outcomes due to estimation αβεβαιότητα. The UCB (upper confidence bound; UCB) strategy addresses this by selecting actions not only based on their estimated μέση τιμήs but also by incorporating a term that reflects the αβεβαιότητα in these estimates—favoring actions with high potential ανταμοιβή and high αβεβαιότητα. Theoretical guarantees for the performance of UCB strategies, including logarithmic regret bounds, are established in [8].

Βλέπε επίσης: ml, ανταμοιβή, πιθανοτικό μοντέλο, MAB, model, realization, RV, μέση τιμή, data, αβεβαιότητα, regret.

απόκλιση Consider an federated learning (FL) application with networked data represented by an FL network. FL methods use a discrepancy measure to compare υπόθεση maps from local models at nodes i, i' connected by an edge in the FL network.

Βλέπε επίσης: FL, networked data, FL network, υπόθεση, local model.

απόκλιση Rényi The Rényi divergence measures the (dis)similarity between two κατανομή πιθανότηταςs [14].

Βλέπε επίσης: κατανομή πιθανότητας.

απώλεια ml methods use a συνάρτηση απώλειας $L\left(\mathbf{z}, h\right)$ to measure the error incurred by applying a specific υπόθεση to a specific data point. With a slight abuse of notation, we use the term loss for both the συνάρτηση

απώλειας L itself and the specific value $L\left(\mathbf{z}, h\right)$, for a data point $\mathbf{z}$ and υπόθεση h.

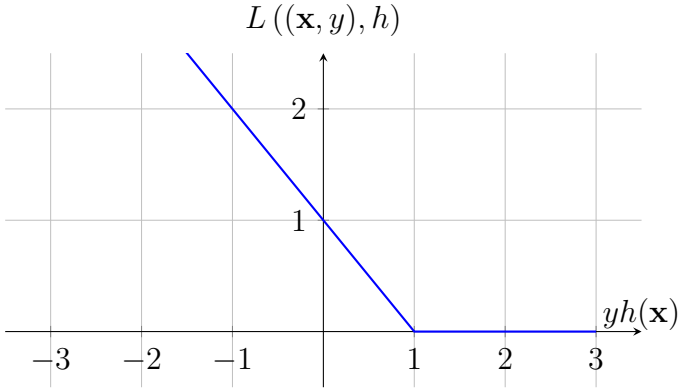
Βλέπε επίσης: ml, συνάρτηση απώλειας, υπόθεση, data point.

απώλεια απόλυτου σφάλματος Consider a data point with features $\mathbf{x} \in \mathcal{X}$ and numeric ετικέτα $y \in \mathbb{R}$. The absolute error loss incurred by a υπόθεση $h : \mathcal{X} \rightarrow \mathbb{R}$ is defined as $|y - h(\mathbf{x})|$, i.e., the absolute difference between the πρόβλεψη $h(\mathbf{x})$ and the true ετικέτα y.

Βλέπε επίσης: data point, feature, ετικέτα, loss, υπόθεση, πρόβλεψη.

απώλεια άρθρωσης Consider a data point characterized by a feature vector $\mathbf{x} \in \mathbb{R}^d$ and a binary ετικέτα $y \in \{-1, 1\}$. The hinge loss incurred by a real-valued υπόθεση map $h(\mathbf{x})$ is defined as

$$L\left((\mathbf{x}, y), h\right) := \max\{0, 1 - yh(\mathbf{x})\}. \tag{1}$$



A regularized variant of the hinge loss is used by the μηχανή διανυσμάτων υποστήριξης [15].

Βλέπε επίσης: data point, feature vector, ετικέτα, loss, υπόθεση, μηχανή διανυσμάτων υποστήριξης.

απώλεια Huber The Huber loss unifies the τετραγωνική απώλεια σφάλματος and the απώλεια απόλυτου σφάλματος.

Βλέπε επίσης: loss, τετραγωνική απώλεια σφάλματος, απώλεια απόλυτου σφάλματος.

αριθμός συνθήκης The condition number $\kappa(\mathbf{Q}) \geq 1$ of a positive definite matrix $\mathbf{Q} \in \mathbb{R}^{d \times d}$ is the ratio α/β between the largest α and the smallest β ιδιοτιμή of $\mathbf{Q}$. The condition number is useful for the analysis of ml methods. The computational complexity of gradient-based methods for γραμμική παλινδρόμηση crucially depends on the condition number of the matrix $\mathbf{Q} = \mathbf{X}\mathbf{X}^T$, with the feature matrix $\mathbf{X}$ of the σύνολο εκπαίδευσης. Thus, from a computational perspective, we prefer features of data points such that $\mathbf{Q}$ has a condition number close to 1.

Βλέπε επίσης: ιδιοτιμή, ml, gradient-based methods, γραμμική παλινδρόμηση, feature matrix, σύνολο εκπαίδευσης, feature, data point.

αρχή της ελαχιστοποίησης των δεδομένων European data protection regulation includes a data minimization principle. This principle requires a data controller to limit the collection of personal information to what is directly relevant and necessary to accomplish a specified purpose. The data should be retained only for as long as necessary to fulfill that purpose [16, Article 5(1)(c)], [17].

Βλέπε επίσης: data.

αυτοκωδικοποιητής Ένας αυτοκωδικοποιητής (autoencoder) είναι μία μέθοδος μηχανικής μάθησης που μαθαίνει ταυτόχρονα έναν κωδικοποιητή αντιστοίχισης $h(\cdot) \in \mathcal{H}$ και έναν αποκωδικοποιητή αντιστοίχισης $h^*(\cdot) \in \mathcal{H}^*$.

Είναι μία περίπτωση της εμπειρικής ελαχιστοποίησης διακινδύνευσης που χρησιμοποιεί μία απώλεια υπολογιζόμενη από το σφάλμα ανακατασκευής $\mathbf{x} - h^{*}\big(h(\mathbf{x})\big)$.

Βλέπε επίσης: ml, εμπειρική ελαχιστοποίηση διακινδύνευσης, loss.

βαθμός κόμβου The degree $d^{(i)}$ of a node $i \in \mathcal{V}$ in an undirected graph is the number of its γείτονες, i.e., $d^{(i)} := \big|\mathcal{N}^{(i)}\big|$.

Βλέπε επίσης: graph, γείτονες.

βαθμός συσχέτισης Degree of belonging is a number that indicates the extent to which a data point belongs to a συστάδα [6, Ch. 8]. The degree of belonging can be interpreted as a soft συστάδα assignment. Soft clustering methods can encode the degree of belonging by a real number in the interval $[0, 1]$. Hard clustering is obtained as the extreme case when the degree of belonging only takes on values 0 or 1.

Βλέπε επίσης: data point, συστάδα, soft clustering, hard clustering.

βαθύ δίκτυο A deep net is an τεχνητό νευρωνικό δίκτυο with a (relatively) large number of hidden layers. Deep learning is an umbrella term for ml methods that use a deep net as their model [18].

Βλέπε επίσης: τεχνητό νευρωνικό δίκτυο, ml, model.

βάρη Consider a parametrized χώρος υποθέσεων $\mathcal{H}$. We use the term weights for numeric παράμετροι μοντέλου that are used to scale features or their transformations in order to compute $h^{(\mathbf{w})} \in \mathcal{H}$. A γραμμικό μοντέλο uses weights $\mathbf{w} = \big(w_1, \ldots, w_d\big)^T$ to compute the linear combination $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$. Weights are also used in τεχνητό νευρωνικό δίκτυος to form linear combinations of features or the outputs of neurons in hidden

layers.

Βλέπε επίσης: χώρος υποθέσεων, παράμετροι μοντέλου, feature, γραμμικό μοντέλο, τεχνητό νευρωνικό δίκτυο.

βάρος ακμής Each edge $\{i, i'\}$ of an FL network is assigned a non-negative edge weight $A_{i,i'} \geq 0$. A zero edge weight $A_{i,i'} = 0$ indicates the absence of an edge between nodes $i, i' \in \mathcal{V}$.

Βλέπε επίσης: FL network.

βάση αναφοράς Consider some ml method that produces a learned υπόθεση (or trained model) $\hat{h} \in \mathcal{H}$. We evaluate the quality of a trained model by computing the average loss on a test set. But how can we assess whether the resulting test set performance is sufficiently good? How can we determine if the trained model performs close to optimal and there is little point in investing more resources (for data collection or computation) to improve it? To this end, it is useful to have a reference (or baseline) level against which we can compare the performance of the trained model. Such a reference value might be obtained from human performance, e.g., the misclassification rate of dermatologists who diagnose cancer from visual inspection of skin [19]. Another source for a baseline is an existing, but for some reason unsuitable, ml method. For example, the existing ml method might be computationally too expensive for the intended ml application. Nevertheless, its test set error can still serve as a baseline. Another, somewhat more principled, approach to constructing a baseline is via a πιθανοτικό μοντέλο. In many cases, given a πιθανοτικό μοντέλο $p(\mathbf{x}, y)$, we can precisely determine

the ολικό ελάχιστο achievable διακινδύνευση among any hypotheses (not even required to belong to the χώρος υποθέσεων $\mathcal{H}$) [20]. This ολικό ελάχιστο achievable διακινδύνευση (referred to as the διακινδύνευση Bayes) is the διακινδύνευση of the εκτιμήτρια Bayes for the ετικέτα y of a data point, given its features $\mathbf{x}$. Note that, for a given choice of συνάρτηση απώλειας, the εκτιμήτρια Bayes (if it exists) is completely determined by the κατανομή πιθανότητας $p(\mathbf{x}, y)$ [20, Ch. 4]. However, computing the εκτιμήτρια Bayes and διακινδύνευση Bayes presents two main challenges:

1) The κατανομή πιθανότητας $p(\mathbf{x}, y)$ is unknown and needs to be estimated.

2) Even if $p(\mathbf{x}, y)$ is known, it can be computationally too expensive to compute the διακινδύνευση Bayes exactly [21].

A widely used πιθανοτικό μοντέλο is the πολυμεταβλητή κανονική κατανομή $(\mathbf{x}, y) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for data points characterized by numeric features and ετικέτας. Here, for the τετραγωνική απώλεια σφάλματος, the εκτιμήτρια Bayes is given by the posterior μέση τιμή $\mu_{y|\mathbf{x}}$ of the ετικέτα y, given the features $\mathbf{x}$ [20], [22]. The corresponding διακινδύνευση Bayes is given by the posterior διακύμανση $\sigma^2_{y|\mathbf{x}}$ (see Fig. 2).

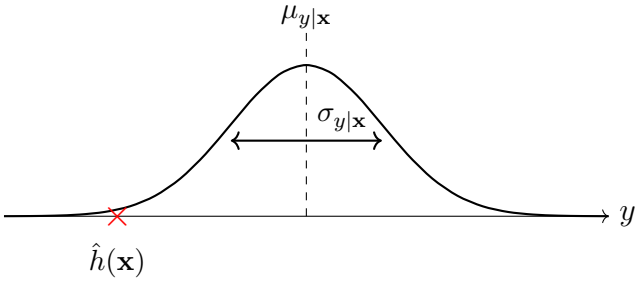


Fig. 2. If the features and the ετικέτα of a data point are drawn from a πολυμεταβλητή κανονική κατανομή, we can achieve the ολικό ελάχιστο διακινδύνευση (under τετραγωνική απώλεια σφάλματος) by using the εκτιμήτρια Bayes $\mu_{y|\mathbf{x}}$ to predict the ετικέτα y of a data point with features $\mathbf{x}$. The corresponding ολικό ελάχιστο διακινδύνευση is given by the posterior διακύμανση $\sigma^2_{y|\mathbf{x}}$. We can use this quantity as a baseline for the average loss of a trained model $\hat{h}$.

Βλέπε επίσης: ml, υπόθεση, model, loss, test set, data, πιθανοτικό μοντέλο, ολικό ελάχιστο, διακινδύνευση, χώρος υποθέσεων, διακινδύνευση Bayes, εκτιμήτρια Bayes, ετικέτα, data point, feature, συνάρτηση απώλειας, κατανομή πιθανότητας, πολυμεταβλητή κανονική κατανομή, τετραγωνική απώλεια σφάλματος, μέση τιμή, διακύμανση.

βήμα κλίσης Given a παραγωγίσιμη real-valued function $f(\cdot) : \mathbb{R}^d \to \mathbb{R}$ and a vector $\mathbf{w} \in \mathbb{R}^d$, the gradient step updates $\mathbf{w}$ by adding the scaled negative gradient $\nabla f(\mathbf{w})$ to obtain the new vector (see Fig. 3)

$$\widehat{\mathbf{w}} := \mathbf{w} - \eta \nabla f(\mathbf{w}). \tag{2}$$

Mathematically, the gradient step is a (typically non-linear) operator $\mathcal{T}^{(f,\eta)}$ that is parametrized by the function f and the step size η.

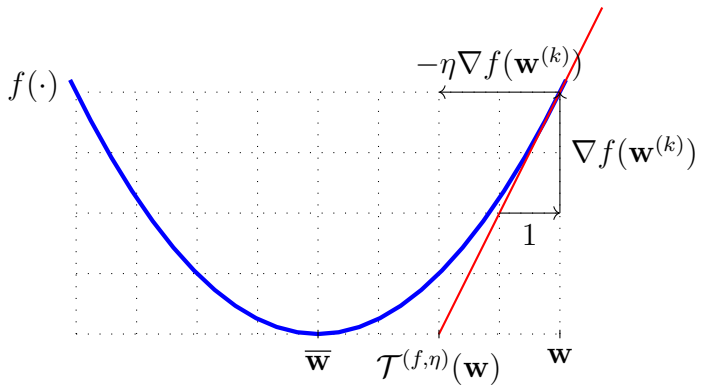


Fig. 3. The basic gradient step (2) maps a given vector $\mathbf{w}$ to the updated vector $\mathbf{w}'$. It defines an operator $\mathcal{T}^{(f,\eta)}(\cdot) : \mathbb{R}^d \to \mathbb{R}^d : \mathbf{w} \mapsto \widehat{\mathbf{w}}$.

Note that the gradient step (2) optimizes locally - in a neighborhood whose size is determined by the step size η - a linear approximation to the function $f(\cdot)$. A natural generalization of (2) is to locally optimize the function itself - instead of its linear approximation - such that

$$\widehat{\mathbf{w}} = \underset{\mathbf{w}' \in \mathbb{R}^d}{\operatorname{argmin}} f(\mathbf{w}') + (1/\eta) \left\| \mathbf{w} - \mathbf{w}' \right\|_2^2 . \tag{3}$$

We intentionally use the same symbol η for the parameter in (3) as we used for the step size in (2). The larger the η we choose in (3), the more progress the update will make towards reducing the function value $f(\widehat{\mathbf{w}})$. Note that, much like the gradient step (2), also the update (3) defines a (typically non-linear) operator that is parametrized by the function $f(\cdot)$ and the parameter η. For a κυρτός function $f(\cdot)$, this operator is known as the εγγύς τελεστής of $f(\cdot)$ [23].

Βλέπε επίσης: παραγωγίσιμη, gradient, step size, neighborhood, generalization, convex, εγγύς τελεστής.

γείτονες The neighbors of a node $i \in \mathcal{V}$ within an FL network are those nodes $i' \in \mathcal{V} \setminus \{i\}$ that are connected (via an edge) to node i.
Βλέπε επίσης: FL network.

γειτονιά The neighborhood of a node $i \in \mathcal{V}$ is the subset of nodes constituted by the γείτονες of i.
Βλέπε επίσης: γείτονες.

γενικός κανονισμός για την προστασία δεδομένων (ΓΚΠΔ) The GDPR (general data protection regulation; GDPR) was enacted by the European Union (EU), effective from May 25, 2018 [16]. It safeguards the privacy and data rights of individuals in the EU. The GDPR has significant implications for how data is collected, stored, and used in ml applications. Key provisions include the following:

- Data minimization principle: ml systems should only use the necessary amount of personal data for their purpose.
- Transparency and επεξηγησιμότητα: ml systems should enable their users to understand how the systems make decisions that impact the users.
- Data subject rights: Users should get an opportunity to access, rectify, and delete their personal data, as well as to object to automated decision-making and profiling.
- Accountability: Organizations must ensure robust data security and demonstrate compliance through documentation and regular audits.

Βλέπε επίσης: data, ml, data minimization principle, transparency, επεξηγησιμότητα.

γραμμικό μοντέλο Consider data points, each characterized by a numeric feature vector $\mathbf{x} \in \mathbb{R}^d$. A linear model is a χώρος υποθέσεων which consists of all linear maps,

$$\mathcal{H}^{(d)} := \left\{ h(\mathbf{x}) = \mathbf{w}^T \mathbf{x} : \mathbf{w} \in \mathbb{R}^d \right\}. \tag{4}$$

Note that (4) defines an entire family of χώρος υποθέσεωνs, which is parametrized by the number d of features that are linearly combined to form the πρόβλεψη $h(\mathbf{x})$. The design choice of d is guided by υπολογιστικές διαστάσεις (e.g., reducing d means less computation), στατιστικές διαστάσεις (e.g., increasing d might reduce πρόβλεψη error), and ερμηνευσιμότητα. A linear model using few carefully chosen features tends to be considered more interpretable [24], [25].

Βλέπε επίσης: data point, feature vector, model, χώρος υποθέσεων, feature, πρόβλεψη, υπολογιστικές διαστάσεις, στατιστικές διαστάσεις, ερμηνευσιμότητα.

γραμμική παλινδρόμηση Linear regression aims to learn a linear υπόθεση map to predict a numeric ετικέτα based on the numeric features of a data point. The quality of a linear υπόθεση map is measured using the average τετραγωνική απώλεια σφάλματος incurred on a set of labeled datapoints, which we refer to as the σύνολο εκπαίδευσης.

Βλέπε επίσης: regression, υπόθεση, ετικέτα, feature, data point, τετραγωνική απώλεια σφάλματος, labeled datapoint, σύνολο εκπαίδευσης.

γράφος A graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a pair that consists of a node set $\mathcal{V}$ and an edge set $\mathcal{E}$. In its most general form, a graph is specified by a map that assigns each edge $e \in \mathcal{E}$ a pair of nodes [26]. One important family of graphs is simple undirected graphs. A simple undirected graph is obtained by identifying each edge $e \in \mathcal{E}$ with two different nodes $\{i, i'\}$. Weighted graphs also specify numeric βάρη A_e for each edge $e \in \mathcal{E}$.
Βλέπε επίσης: βάρη.

δέντρο αποφάσεων A decision tree is a flow-chart-like representation of a υπόθεση map h. More formally, a decision tree is a directed graph containing a root node that reads in the feature vector $\mathbf{x}$ of a data point. The root node then forwards the data point to one of its children nodes based on some elementary test on the features $\mathbf{x}$. If the receiving child node is not a leaf node, i.e., it has itself children nodes, it represents another test. Based on the test result, the data point is forwarded to one of its descendants. This testing and forwarding of the data point is continued until the data point ends up in a leaf node (having no children nodes).

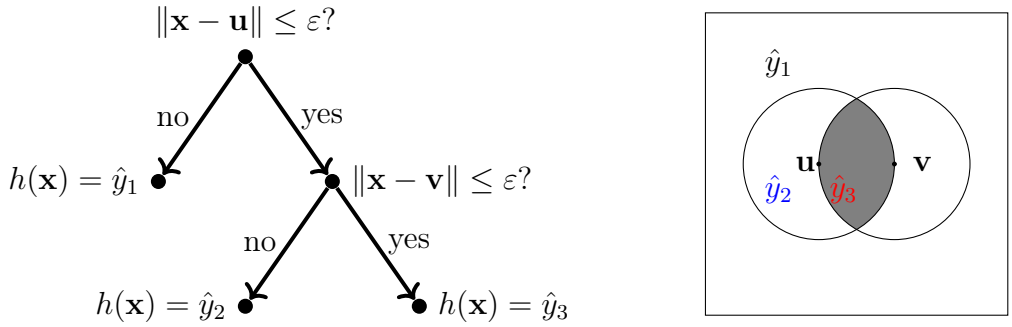


Fig. 4. Left: A decision tree is a flow-chart-like representation of a piecewise constant υπόθεση $h : \mathcal{X} \to \mathbb{R}$. Each piece is a περιοχή αποφάσεων $\mathcal{R}_{\hat{y}} := \{\mathbf{x} \in \mathcal{X} : h(\mathbf{x}) = \hat{y}\}$. The depicted decision tree can be applied to numeric feature vectors, i.e., $\mathcal{X} \subseteq \mathbb{R}^d$. It is parametrized by the threshold $\varepsilon > 0$ and the vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$. Right: A decision tree partitions the feature space $\mathcal{X}$ into περιοχή αποφάσεωνs. Each περιοχή αποφάσεων $\mathcal{R}_{\hat{y}} \subseteq \mathcal{X}$ corresponds to a specific leaf node in the decision tree.

Βλέπε επίσης: υπόθεση, graph, feature vector, data point, feature, περιοχή αποφάσεων, feature space.

δέσμη In the context of στοχαστική κάθοδος κλίσης, a batch refers to a randomly chosen subset of the overall σύνολο εκπαίδευσης. We use the data points in this subset to estimate the gradient of training error and, in turn, to update the παράμετροι μοντέλου.

Βλέπε επίσης: στοχαστική κάθοδος κλίσης, σύνολο εκπαίδευσης, data point, gradient, training error, παράμετροι μοντέλου.

δεδομένα Data refers to objects that carry information. These objects can be either concrete physical objects (such as persons or animals) or

abstract concepts (such as numbers). We often use representations (or approximations) of the original data that are more convenient for data processing. These approximations are based on different data models, with the relational data model being one of the most widely used [27].
Βλέπε επίσης: model.

δείγμα A finite sequence (or list) of data points $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}$ that is obtained or interpreted as the realization of m i.i.d. RVs with a common κατανομή πιθανότητας $p(\mathbf{z})$. The length m of the sequence is referred to as the μέγεθος δείγματος.
Βλέπε επίσης: data point, realization, i.i.d., RV, κατανομή πιθανότητας, μέγεθος δείγματος.

διακινδύνευση Bayes Consider a πιθανοτικό μοντέλο with a joint κατανομή πιθανότητας $p(\mathbf{x}, y)$ for the features $\mathbf{x}$ and ετικέτα y of a data point. The Bayes διακινδύνευση is the ολικό ελάχιστο possible διακινδύνευση that can be achieved by any υπόθεση $h : \mathcal{X} \to \mathcal{Y}$. Any υπόθεση that achieves the Bayes risk is referred to as a εκτιμήτρια Bayes [20].
Βλέπε επίσης: πιθανοτικό μοντέλο, κατανομή πιθανότητας, feature, ετικέτα, data point, διακινδύνευση, ολικό ελάχιστο, υπόθεση, εκτιμήτρια Bayes.

διακύμανση The variance of a real-valued RV x is defined as the expectation $\mathbb{E}\big\{\big(x - \mathbb{E}\{x\}\big)^2\big\}$ of the squared difference between x and its expectation $\mathbb{E}\{x\}$. We extend this definition to vector-valued RVs $\mathbf{x}$ as $\mathbb{E}\big\{\big\|\mathbf{x} - \mathbb{E}\{\mathbf{x}\}\big\|_2^2\big\}$.
Βλέπε επίσης: RV, expectation.

διακινδύνευση Consider a υπόθεση h used to predict the ετικέτα y of a data point based on its features $\mathbf{x}$. We measure the quality of a particular πρόβλεψη using a συνάρτηση απώλειας $L\left((\mathbf{x}, y), h\right)$. If we interpret data points as the realizations of i.i.d. RVs, also the $L\left((\mathbf{x}, y), h\right)$ becomes the realization of an RV. The independent and identically distributed assumption (i.i.d. assumption) allows us to define the risk of a υπόθεση as the expected loss $\mathbb{E}\big\{L\left((\mathbf{x}, y), h\right)\big\}$. Note that the risk of h depends on both the specific choice for the συνάρτηση απώλειας and the κατανομή πιθανότητας of the data points.

Βλέπε επίσης: υπόθεση, ετικέτα, data point, feature, πρόβλεψη, συνάρτηση απώλειας, realization, i.i.d. RV, i.i.d. assumption, loss, κατανομή πιθανότητας.

διαρροή ιδιωτικότητας Consider an ml application that processes a σύνολο δεδομένων $\mathcal{D}$ and delivers some output, such as the πρόβλεψης obtained for new data points. Privacy leakage arises if the output carries information about a private (or sensitive) feature of a data point (which might be a human) of $\mathcal{D}$. Based on a πιθανοτικό μοντέλο for the data generation, we can measure the privacy leakage via the αμοιβαία πληροφορία between the output and the senstive feature. Another quantitative measure of privacy leakage is διαφορική ιδιωτικότητα. The relations between different measures of privacy leakage have been studied in the literature (see [28]).

Βλέπε επίσης: ml, σύνολο δεδομένων, πρόβλεψη, data point, feature, πιθανοτικό μοντέλο, data, αμοιβαία πληροφορία, διαφορική ιδιωτικότητα.

διασταυρούμενη επικύρωση k-συνόλων Η διασταυρούμενη επικύρωση k-συνόλων (k-fold cross-validation; k-fold CV) είναι μία μέθοδος για τη μάθηση και επικύρωση μίας υπόθεσης χρησιμοποιώντας ένα συγκεκριμένο σύνολο δεδομένων. Αυτή η μέθοδος διαιρεί το σύνολο δεδομένων ισότιμα σε k υποσύνολα και στη συνέχεια εκτελεί k επαναλήψεις εκπαίδευσης μοντέλου (π.χ., μέσω της εμπειρικής ελαχιστοποίησης διακινδύνευσης) και επικύρωσης. Κάθε επανάληψη χρησιμοποιεί ένα διαφορετικό υποσύνολο ως το σύνολο επικύρωσης και τα υπόλοιπα $k-1$ υποσύνολα ως σύνολο εκπαίδευσης. Το τελικό εξαγόμενο είναι ο μέσος όρος των σφαλμάτων επικύρωσης που προκύπτουν από τις k επαναλήψεις.

Βλέπε επίσης: υπόθεση, σύνολο δεδομένων, model, εμπειρική ελαχιστοποίηση διακινδύνευσης, επικύρωση, σύνολο επικύρωσης, σύνολο εκπαίδευσης, σφάλμα επικύρωσης.

δίαυλος ιδιωτικότητας The privacy funnel is a method for learning privacy-friendly features of data points [29].

Βλέπε επίσης: feature, data point.

διαφάνεια Transparency is a fundamental requirement for trustworthy artificial intelligence (trustworthy AI) [30]. In the context of ml methods, transparency is often used interchangeably with επεξηγησιμότητα [31], [32]. However, in the broader scope of τεχνητή νοημοσύνη systems, transparency extends beyond επεξηγησιμότητα and includes providing information about the system's limitations, reliability, and intended use. In medical diagnosis systems, transparency requires disclosing the confidence level for the πρόβλεψης delivered by a trained

model. In credit scoring, τεχνητή νοημοσύνη-based lending decisions should be accompanied by explanations of contributing factors, such as income level or credit history. These explanations allow humans (e.g., a loan applicant) to understand and contest automated decisions. Some ml methods inherently offer transparency. For example, λογιστική παλινδρόμηση provides a quantitative measure of ταξινόμηση reliability through the value $|h(\mathbf{x})|$. Decision trees are another example, as they allow human-readable decision rules [24]. Transparency also requires a clear indication when a user is engaging with an τεχνητή νοημοσύνη system. For example, τεχνητή νοημοσύνη-powered chatbots should notify users that they are interacting with an automated system rather than a human. Furthermore, transparency encompasses comprehensive documentation detailing the purpose and design choices underlying the τεχνητή νοημοσύνη system. For instance, model datasheets [33] and τεχνητή νοημοσύνη system cards [34] help practitioners understand the intended use cases and limitations of an τεχνητή νοημοσύνη system [35]. Βλέπε επίσης: trustworthy AI, ml, επεξηγησιμότητα, τεχνητή νοημοσύνη, πρόβλεψη, model, λογιστική παλινδρόμηση, ταξινόμηση, decision tree.

διαφορική ιδιωτικότητα Consider some ml method $\mathcal{A}$ that reads in a σύνολο δεδομένων (e.g., the σύνολο εκπαίδευσης used for εμπειρική ελαχιστοποίηση διακινδύνευσης) and delivers some output $\mathcal{A}(\mathcal{D})$. The output could be either the learned παράμετροι μοντέλου or the πρόβλεψης for specific data points. DP (differential privacy; DP) is a precise measure of διαρροή ιδιωτικότητας incurred by revealing the output. Roughly speaking, an ml method is differentially private if the κατανομή

πιθανότητας of the output $\mathcal{A}(\mathcal{D})$ does not change too much if the ευαίσθητο ιδιοχαρακτηριστικό of one data point in the σύνολο εκπαίδευσης is changed. Note that DP builds on a πιθανοτικό μοντέλο for an ml method, i.e., we interpret its output $\mathcal{A}(\mathcal{D})$ as the realization of an RV. The randomness in the output can be ensured by intentionally adding the realization of an auxiliary RV (noise) to the output of the ml method.

Βλέπε επίσης: ml, σύνολο δεδομένων, σύνολο εκπαίδευσης, εμπειρική ελαχιστοποίηση διακινδύνευσης, παράμετροι μοντέλου, πρόβλεψη, data point, διαρροή ιδιωτικότητας, κατανομή πιθανότητας, ευαίσθητο ιδιοχαρακτηριστικό, πιθανοτικό μοντέλο, realization, RV.

διεπαφή προγραμματισμού εφαρμογών An API (application programming interface; API) is a formal mechanism that allows software components to interact in a structured and modular way [36]. In the context of ml, APIs are commonly used to provide access to a trained ml model. Users—whether humans or machines—can submit the feature vector of a data point and receive a corresponding πρόβλεψη. Suppose a trained ml model is defined as $\widehat{h}(x) := 2x + 1$. Through an API, a user can input $x = 3$ and receive the output $\widehat{h}(3) = 7$ without knowledge of the detailed structure of the ml model or its training. In practice, the model is typically deployed on a server connected to the internet. Clients send requests containing feature values to the server, which responds with the computed πρόβλεψη $\widehat{h}(\mathbf{x})$. APIs promote modularity in ml system design: One team can develop and train the model, while another team handles integration and user interaction. Publishing a trained model

via an API also offers practical advantages:

- The server can centralize computational resources which are required to compute πρόβλεψηs.
- The internal structure of the model remains hidden (useful for protecting IP or trade secrets).

However, APIs are not without risk: Techniques such as model inversion can potentially reconstruct a model from its πρόβλεψηs on carefully selected feature vectors.

Βλέπε επίσης: ml, model, feature vector, data point, πρόβλεψη, feature, model inversion.

εγγύς τελεστής Given a convex function $f(\mathbf{w}')$, we define its proximal operator as [23], [37]

$$\mathbf{prox}_{f(\cdot),\rho}(\mathbf{w}) := \underset{\mathbf{w}' \in \mathbb{R}^d}{\operatorname{argmin}} \left[f(\mathbf{w}') + (\rho/2) \left\| \mathbf{w} - \mathbf{w}' \right\|_2^2 \right] \text{ with } \rho > 0.$$

As illustrated in Fig. 5, evaluating the proximal operator amounts to minimizing a penalized variant of $f(\mathbf{w}')$. The penalty term is the scaled squared Euclidean distance to a given vector $\mathbf{w}$ (which is the input to the proximal operator). The proximal operator can be interpreted as a generalization of the βήμα κλίσης, which is defined for a λεία convex function $f(\mathbf{w}')$. Indeed, taking a βήμα κλίσης with step size η at the current vector $\mathbf{w}$ is the same as applying the proximal operator of the function $\tilde{f}(\mathbf{w}') = \big(\nabla f(\mathbf{w})\big)^T (\mathbf{w}' - \mathbf{w})$ and using $\rho = 1/\eta$.

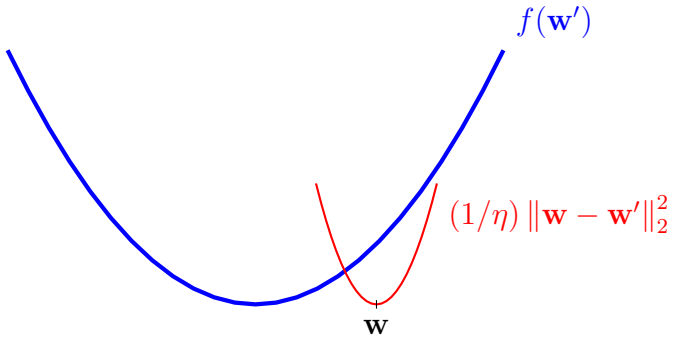


Fig. 5. A generalized βήμα κλίσης updates a vector $\mathbf{w}$ by minimizing a penalized version of the function $f(\cdot)$. The penalty term is the scaled squared Euclidean distance between the optimization variable $\mathbf{w}'$ and the given vector $\mathbf{w}$.

Βλέπε επίσης: convex, generalization, βήμα κλίσης, λεία, step size,

εκκίνηση For the analysis of ml methods, it is often useful to interpret a given set of data points $\mathcal{D} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}\}$ as realizations of i.i.d. RVs with a common κατανομή πιθανότητας $p(\mathbf{z})$. In general, we do not know $p(\mathbf{z})$ exactly, but we need to estimate it. The bootstrap uses the histogram of $\mathcal{D}$ as an estimator for the underlying κατανομή πιθανότητας $p(\mathbf{z})$.

Βλέπε επίσης: ml, data point, realization, i.i.d., RV, κατανομή πιθανότητας.

εκμάθηση πολυδιεργασίας Multitask learning aims at leveraging relations between different εργασία εκμάθησηςs. Consider two εργασία εκμάθησηςs obtained from the same σύνολο δεδομένων of webcam snap-

shots. The first task is to predict the presence of a human, while the second task is to predict the presence of a car. It might be useful to use the same βαθύ δίκτυο structure for both tasks and only allow the βάρη of the final output layer to be different.

Βλέπε επίσης: εργασία εκμάθησης, σύνολο δεδομένων, βαθύ δίκτυο, βάρη.

εκμάθηση χαρακτηριστικών Consider an ml application with data points characterized by raw features $\mathbf{x} \in \mathcal{X}$. Feature learning refers to the task of learning a map

$$\mathbf{\Phi} : \mathcal{X} \to \mathcal{X}' : \mathbf{x} \mapsto \mathbf{x}',$$

that reads in raw features $\mathbf{x} \in \mathcal{X}$ of a data point and delivers new features $\mathbf{x}' \in \mathcal{X}'$ from a new feature space $\mathcal{X}'$. Different feature learning methods are obtained for different design choices of $\mathcal{X}, \mathcal{X}'$, for a χώρος υποθέσεων $\mathcal{H}$ of potential maps $\mathbf{\Phi}$, and for a quantitative measure of the usefulness of a specific $\mathbf{\Phi} \in \mathcal{H}$. For example, ανάλυση κυρίων συνιστωσών uses $\mathcal{X} := \mathbb{R}^d$, $\mathcal{X}' := \mathbb{R}^{d'}$ with $d' < d$, and a χώρος υποθέσεων

$$\mathcal{H} := \big\{ \mathbf{\Phi} : \mathbb{R}^d \to \mathbb{R}^{d'} : \mathbf{x}' := \mathbf{F}\mathbf{x} \text{ with some } \mathbf{F} \in \mathbb{R}^{d' \times d} \big\}.$$

Principal component analysis measures the usefulness of a specific map $\mathbf{\Phi}(\mathbf{x}) = \mathbf{F}\mathbf{x}$ by the ολικό ελάχιστο linear reconstruction error incurred on a σύνολο δεδομένων,

$$\min_{\mathbf{G} \in \mathbb{R}^{d \times d'}} \sum_{r=1}^{m} \big\| \mathbf{G}\mathbf{F}\mathbf{x}^{(r)} - \mathbf{x}^{(r)} \big\|_2^2 .$$

Βλέπε επίσης: ml, data point, feature, feature space, χώρος υποθέσεων, principal component analysis, ολικό ελάχιστο, σύνολο δεδομένων.

εκτιμήτρια Bayes Consider a πιθανοτικό μοντέλο with a joint κατανομή πιθανότητας $p(\mathbf{x}, y)$ for the features $\mathbf{x}$ and ετικέτα y of a data point. For a given συνάρτηση απώλειας $L(\cdot, \cdot)$, we refer to a υπόθεση h as a Bayes estimator if its διακινδύνευση $\mathbb{E}\{L((\mathbf{x}, y), h)\}$ is the ολικό ελάχιστο [20]. Note that the property of a υπόθεση being a Bayes estimator depends on the underlying κατανομή πιθανότητας and the choice for the συνάρτηση απώλειας $L(\cdot, \cdot)$.

Βλέπε επίσης: πιθανοτικό μοντέλο, κατανομή πιθανότητας, feature, ετικέτα, data point, συνάρτηση απώλειας, υπόθεση, διακινδύνευση, ολικό ελάχιστο.

εμπειρική διακινδύνευση The empirical διακινδύνευση $\widehat{L}(h|\mathcal{D})$ of a υπόθεση on a σύνολο δεδομένων $\mathcal{D}$ is the average loss incurred by h when applied to the data points in $\mathcal{D}$.

Βλέπε επίσης: διακινδύνευση, υπόθεση, σύνολο δεδομένων, loss, data point.

εμπειρική ελαχιστοποίηση διακινδύνευσης Empirical risk minimization (empirical risk minimization; ERM) is the optimization problem of finding a υπόθεση (out of a model) with the ολικό ελάχιστο average loss (or empirical risk) on a given σύνολο δεδομένων $\mathcal{D}$ (i.e., the σύνολο εκπαίδευσης). Many ml methods are obtained from empirical risk via specific design choices for the σύνολο δεδομένων, model, and loss [6, Ch. 3].

Βλέπε επίσης: empirical risk, υπόθεση, model, ολικό ελάχιστο, loss, σύνολο δεδομένων, σύνολο εκπαίδευσης, ml.

επαύξηση δεδομένων Data augmentation methods add synthetic data points to an existing set of data points. These synthetic data points are obtained by perturbations (e.g., adding noise to physical measurements) or transformations (e.g., rotations of images) of the original data points. These perturbations and transformations are such that the resulting synthetic data points should still have the same ετικέτα. As a case in point, a rotated cat image is still a cat image even if their feature vectors (obtained by stacking pixel color intensities) are very different (see Fig. 6). Data augmentation can be an efficient form of regularization.

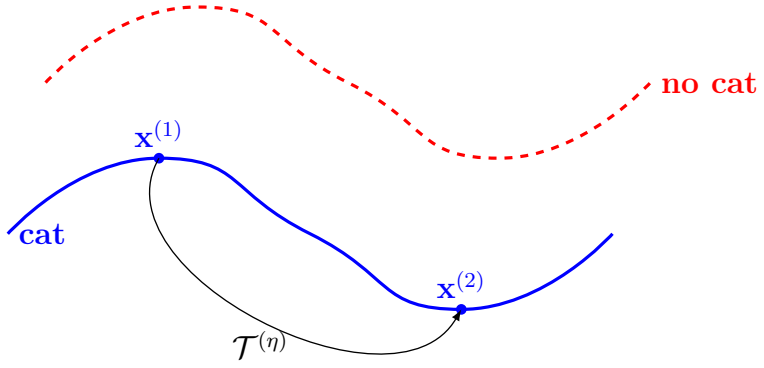


Fig. 6. Data augmentation exploits intrinsic symmetries of data points in some feature space $\mathcal{X}$. We can represent a symmetry by an operator $\mathcal{T}^{(\eta)} : \mathcal{X} \to \mathcal{X}$, parametrized by some number $\eta \in \mathbb{R}$. For example, $\mathcal{T}^{(\eta)}$ might represent the effect of rotating a cat image by η degrees. A data point with feature vector $\mathbf{x}^{(2)} = \mathcal{T}^{(\eta)}\big(\mathbf{x}^{(1)}\big)$ must have the same ετικέτα $y^{(2)} = y^{(1)}$ as a data point with feature vector $\mathbf{x}^{(1)}$.

Βλέπε επίσης: data, data point, ετικέτα, feature vector, regularization, feature space.

επεξήγηση One approach to make ml methods transparent is to provide an explanation along with the πρόβλεψη delivered by an ml method. Explanations can take on many different forms. An explanation could be some natural text or some quantitative measure for the importance of individual features of a data point [38]. We can also use visual forms of explanations, such as intensity plots for image ταξινόμηση [39].

Βλέπε επίσης: ml, πρόβλεψη, feature, data point, ταξινόμηση.

επεξηγησιμότητα We define the (subjective) explainability of an ml method as the level of simulatability [40] of the πρόβλεψηs delivered by an ml system to a human user. Quantitative measures for the (subjective) explainability of a trained model can be constructed by comparing its πρόβλεψηs with the πρόβλεψηs provided by a user on a test set [40], [41]. Alternatively, we can use πιθανοτικό μοντέλοs for data and measure the explainability of a trained ml model via the conditional (differential) entropy of its πρόβλεψηs, given the user πρόβλεψηs [31], [42].

Βλέπε επίσης: ml, πρόβλεψη, model, test set, πιθανοτικό μοντέλο, data.

επικύρωση Consider a υπόθεση $\hat{h}$ that has been learned via some ml method, e.g., by solving εμπειρική ελαχιστοποίηση διακινδύνευσης on a σύνολο εκπαίδευσης $\mathcal{D}$. Validation refers to the practice of evaluating the loss incurred by the υπόθεση $\hat{h}$ on a set of data points that are not contained in the σύνολο εκπαίδευσης $\mathcal{D}$.

Βλέπε επίσης: υπόθεση, ml, εμπειρική ελαχιστοποίηση διακινδύνευσης, σύνολο εκπαίδευσης, loss, data point.

εργασία εκμάθησης Consider a σύνολο δεδομένων $\mathcal{D}$ constituted by sev-

eral data points, each of them characterized by features $\mathbf{x}$. For example, the σύνολο δεδομένων $\mathcal{D}$ might be constituted by the images of a particular database. Sometimes it might be useful to represent a σύνολο δεδομένων $\mathcal{D}$, along with the choice of features, by a κατανομή πιθανότητας $p(\mathbf{x})$. A learning task associated with $\mathcal{D}$ consists of a specific choice for the ετικέτα of a data point and the corresponding label space. Given a choice for the συνάρτηση απώλειας and model, a learning task gives rise to an instance of εμπειρική ελαχιστοποίηση διακινδύνευσης. Thus, we could define a learning task also via an instance of εμπειρική ελαχιστοποίηση διακινδύνευσης, i.e., via an objective function. Note that, for the same σύνολο δεδομένων, we obtain different learning tasks by using different choices for the features and ετικέτα of a data point. These learning tasks are related, as they are based on the same σύνολο δεδομένων, and solving them jointly (via εκμάθηση πολυδιεργασίας methods) is typically preferable over solving them separately [43], [44], [45].

Βλέπε επίσης: σύνολο δεδομένων, data point, feature, κατανομή πιθανότητας, ετικέτα, label space, συνάρτηση απώλειας, model, εμπειρική ελαχιστοποίηση διακινδύνευσης, objective function, εκμάθηση πολυδιεργασίας.

ερμηνευσιμότητα An ml method is interpretable for a specific user if they can well anticipate the πρόβλεψηs delivered by the method. The notion of interpretability can be made precise using quantitative measures of the αβεβαιότητα about the πρόβλεψηs [31].

Βλέπε επίσης: ml, πρόβλεψη, αβεβαιότητα.

ετικέτα A higher-level fact or quantity of interest associated with a data point. For example, if the data point is an image, the label could indicate whether the image contains a cat or not. Synonyms for label, commonly used in specific domains, include "response variable," "output variable," and "target" [46], [47], [48].

Βλέπε επίσης: data point.

ευαίσθητο ιδιοχαρακτηριστικό ml revolves around learning a υπόθεση map that allows us to predict the ετικέτα of a data point from its features. In some applications, we must ensure that the output delivered by an ml system does not allow us to infer sensitive attributes of a data point. Which part of a data point is considered a sensitive attribute is a design choice that varies across different application domains.

Βλέπε επίσης: ml, υπόθεση, ετικέτα, data point, feature.

ιδιοδιάνυσμα An eigenvector of a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ is a non-zero vector $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ such that $\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$ with some ιδιοτιμή λ.

Βλέπε επίσης: ιδιοτιμή.

ιδιοτιμή We refer to a number $\lambda \in \mathbb{R}$ as an eigenvalue of a square matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$ if there is a non-zero vector $\mathbf{x} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ such that $\mathbf{A}\mathbf{x} = \lambda \mathbf{x}$.

κάθοδος κλίσης Gradient descent (gradient descent; GD) is an iterative method for finding the ολικό ελάχιστο of a παραγωγίσιμη function $f(\mathbf{w})$ of a vector-valued argument $\mathbf{w} \in \mathbb{R}^d$. Consider a current guess or approximation $\mathbf{w}^{(k)}$ for the ολικό ελάχιστο of the function $f(\mathbf{w})$. We would like to find a new (better) vector $\mathbf{w}^{(k+1)}$ that has a smaller objective value $f(\mathbf{w}^{(k+1)}) < f\big(\mathbf{w}^{(k)}\big)$ than the current guess $\mathbf{w}^{(k)}$. We

can achieve this typically by using a βήμα κλίσης

$$\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} - \eta \nabla f(\mathbf{w}^{(k)}) \tag{5}$$

with a sufficiently small step size $\eta > 0$. Fig. 7 illustrates the effect of a single gradient descent step (5).

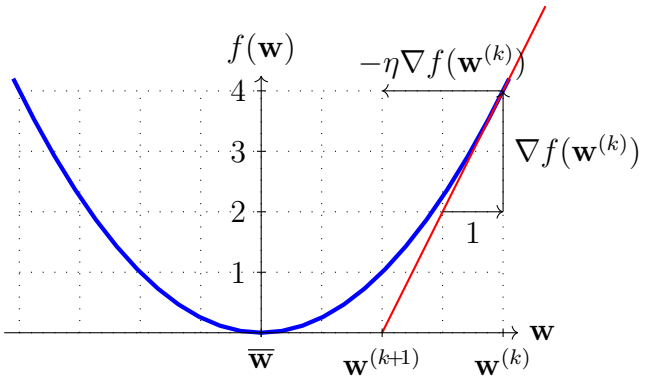


Fig. 7. A single βήμα κλίσης (5) towards the minimizer $\overline{\mathbf{w}}$ of $f(\mathbf{w})$.

Βλέπε επίσης: gradient, ολικό ελάχιστο, παραγωγίσιμη, βήμα κλίσης, step size.

κάθοδος υποκλίσης Subgradient descent is a generalization of κάθοδος κλίσης that does not require differentiability of the function to be minimized. This generalization is obtained by replacing the concept of a gradient with that of a subgradient. Similar to gradients, also subgradients allow us to construct local approximations of an objective function. The objective function might be the empirical risk $\widehat{L}\big(h^{(\mathbf{w})} \big| \mathcal{D}\big)$ viewed as a function of the παράμετροι μοντέλου $\mathbf{w}$ that select a υπόθεση $h^{(\mathbf{w})} \in \mathcal{H}$.

Βλέπε επίσης: subgradient, generalization, κάθοδος κλίσης, gradient, objective function, empirical risk, παράμετροι μοντέλου, υπόθεση.

κανονικοποίηση δεδομένων Data normalization refers to transformations applied to the feature vectors of data points to improve the ml method's στατιστικές διαστάσεις or υπολογιστικές διαστάσεις. For example, in γραμμική παλινδρόμηση with gradient-based methods using a fixed learning rate, convergence depends on controlling the νόρμα of feature vectors in the σύνολο εκπαίδευσης. A common approach is to normalize feature vectors such that their νόρμα does not exceed one [6, Ch. 5].

Βλέπε επίσης: data, feature vector, data point, ml, στατιστικές διαστάσεις, υπολογιστικές διαστάσεις, γραμμική παλινδρόμηση, gradient-based methods, learning rate, νόρμα, σύνολο εκπαίδευσης.

κατανομή πιθανότητας To analyze ml methods, it can be useful to interpret data points as i.i.d. realizations of an RV. The typical properties of such data points are then governed by the probability distribution of this RV. The probability distribution of a binary RV $y \in \{0,1\}$ is fully specified by the probabilities $p(y=0)$ and $p(y=1)=1-p(y=0)$. The probability distribution of a real-valued RV $x \in \mathbb{R}$ might be specified by a συνάρτηση πυκνότητας πιθανότητας $p(x)$ such that $p(x \in [a,b]) \approx p(a)|b-a|$. In the most general case, a probability distribution is defined by a probability measure [22], [49].

Βλέπε επίσης: ml, data point, i.i.d., realization, RV, probability, συνάρτηση πυκνότητας πιθανότητας.

κερκόπορτα A backdoor attack refers to the intentional manipulation of the training process underlying an ml method. This manipulation can be implemented by perturbing the σύνολο εκπαίδευσης (data poisoning) or the optimization αλγόριθμος used by an εμπειρική ελαχιστοποίηση διακινδύνευσης-based method. The goal of a backdoor attack is to nudge the learned υπόθεση $\hat{h}$ towards specific πρόβλεψηs for a certain range of feature values. This range of feature values serves as a key (or trigger) to unlock a backdoor in the sense of delivering anomalous πρόβλεψηs. The key $\mathbf{x}$ and the corresponding anomalous πρόβλεψη $\hat{h}(\mathbf{x})$ are only known to the attacker.

Βλέπε επίσης: ml, σύνολο εκπαίδευσης, (data, αλγόριθμος, εμπειρική ελαχιστοποίηση διακινδύνευσης, υπόθεση, πρόβλεψη, feature.

κλίση For a real-valued function $f : \mathbb{R}^d \to \mathbb{R} : \mathbf{w} \mapsto f(\mathbf{w})$, a vector $\mathbf{g}$ such that $\lim_{\mathbf{w}\to\mathbf{w}'} \frac{f(\mathbf{w})-\left(f(\mathbf{w}')+\mathbf{g}^T(\mathbf{w}-\mathbf{w}')\right)}{\|\mathbf{w}-\mathbf{w}'\|} = 0$ is referred to as the gradient of f at $\mathbf{w}'$. If such a vector exists, it is denoted $\nabla f(\mathbf{w}')$ or $\nabla f(\mathbf{w})\big|_{\mathbf{w}'}$ [2].

κριτήριο τερματισμού Many ml methods use iterative αλγόριθμοςs that construct a sequence of παράμετροι μοντέλου (such as the βάρη of a linear map or the βάρη of an τεχνητό νευρωνικό δίκτυο). These parameters (hopefully) converge to an optimal choice for the παράμετροι μοντέλου. In practice, given finite computational resources, we need to stop iterating after a finite number of repetitions. A stopping criterion is any well-defined condition required for stopping the iteration.

Βλέπε επίσης: ml, αλγόριθμος, παράμετροι μοντέλου, βάρη, τεχνητό νευρωνικό δίκτυο.

κυρτή συσταδοποίηση Consider a σύνολο δεδομένων $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \in \mathbb{R}^d$. Convex συσταδοποίηση learns vectors $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(m)}$ by minimizing

$$\sum_{r=1}^{m} \left\| \mathbf{x}^{(r)} - \mathbf{w}^{(r)} \right\|_2^2 + \alpha \sum_{i,i' \in \mathcal{V}} \left\| \mathbf{w}^{(i)} - \mathbf{w}^{(i')} \right\|_p .$$

Here, $\|\mathbf{u}\|_p := \left(\sum_{j=1}^{d} |u_j|^p \right)^{1/p}$ denotes the p-νόρμα (for $p \geq 1$). It turns out that many of the optimal vectors $\widehat{\mathbf{w}}^{(1)}, \ldots, \widehat{\mathbf{w}}^{(m)}$ coincide. A συστάδα then consists of those data points $r \in \{1, \ldots, m\}$ with identical $\widehat{\mathbf{w}}^{(r)}$ [50], [51].

Βλέπε επίσης: σύνολο δεδομένων, convex, συσταδοποίηση, νόρμα, συστάδα, data point.

κυρτός A subset $\mathcal{C} \subseteq \mathbb{R}^d$ of the Euclidean space $\mathbb{R}^d$ is referred to as convex if it contains the line segment between any two points $\mathbf{x}, \mathbf{y} \in \mathcal{C}$ in that set. A function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if its epigraph $\left\{ \left(\mathbf{w}^T, t \right)^T \in \mathbb{R}^{d+1} : t \geq f(\mathbf{w}) \right\}$ is a convex set [52]. We illustrate one example of a convex set and a convex function in Fig. 8.

Fig. 8. Left: A convex set $\mathcal{C} \subseteq \mathbb{R}^d$. Right: A convex function $f : \mathbb{R}^d \to \mathbb{R}$.

Βλέπε επίσης: Euclidean space.

λεία A real-valued function $f : \mathbb{R}^d \to \mathbb{R}$ is smooth if it is παραγωγίσιμη and its gradient $\nabla f(\mathbf{w})$ is continuous at all $\mathbf{w} \in \mathbb{R}^d$ [53], [54]. A smooth

function f is referred to as β-smooth if the gradient $\nabla f(\mathbf{w})$ is Lipschitz continuous with Lipschitz constant β, i.e.,

$$\|\nabla f(\mathbf{w}) - \nabla f(\mathbf{w}')\| \leq \beta \|\mathbf{w} - \mathbf{w}'\|, \text{ for any } \mathbf{w}, \mathbf{w}' \in \mathbb{R}^d.$$

The constant β quantifies the amount of smoothness of the function f: the smaller the β, the smoother f is. Optimization problems with a smooth objective function can be solved effectively by gradient-based methods. Indeed, gradient-based methods approximate the objective function locally around a current choice $\mathbf{w}$ using its gradient. This approximation works well if the gradient does not change too rapidly. We can make this informal claim precise by studying the effect of a single βήμα κλίσης with step size $\eta = 1/\beta$ (see Fig. 9).

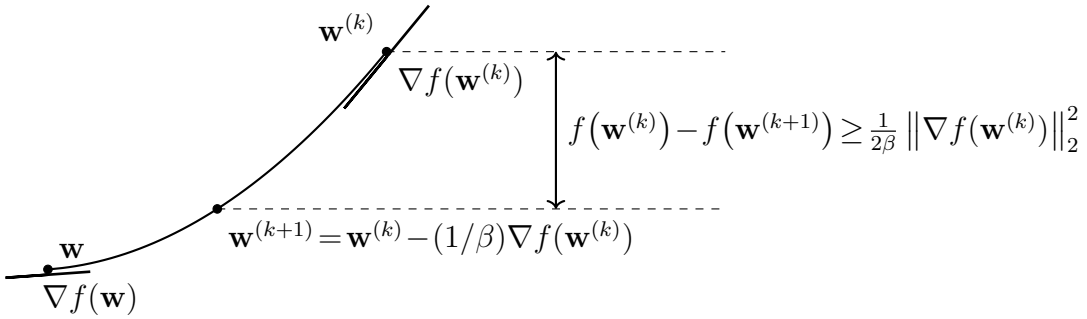


Fig. 9. Consider an objective function $f(\mathbf{w})$ that is β-smooth. Taking a βήμα κλίσης, with step size $\eta = 1/\beta$, decreases the objective by at least $\frac{1}{2\beta} \left\|\nabla f(\mathbf{w}^{(k)})\right\|_2^2$ [53], [54], [55]. Note that the step size $\eta = 1/\beta$ becomes larger for smaller β. Thus, for smoother objective functions (i.e., those with smaller β), we can take larger steps.

Βλέπε επίσης: παραγωγίσιμη, gradient, objective function, gradient-

based methods, βήμα κλίσης, step size.

λογιστική απώλεια Consider a data point characterized by the features $\mathbf{x}$ and a binary ετικέτα $y \in \{-1, 1\}$. We use a real-valued υπόθεση h to predict the ετικέτα y from the features $\mathbf{x}$. The logistic loss incurred by this πρόβλεψη is defined as

$$L\left((\mathbf{x}, y), h\right) := \log(1 + \exp(-yh(\mathbf{x}))). \tag{6}$$

Carefully note that the expression (6) for the logistic loss applies only for the label space $\mathcal{Y} = \{-1, 1\}$ and when using the thresholding rule (8).

Βλέπε επίσης: data point, feature, ετικέτα, υπόθεση, loss, πρόβλεψη, label space.

λογιστική παλινδρόμηση Logistic regression learns a linear υπόθεση map (or ταξινομητής) $h(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ to predict a binary ετικέτα y based on the numeric feature vector $\mathbf{x}$ of a data point. The quality of a linear υπόθεση map is measured by the average λογιστική απώλεια on some labeled datapoints (i.e., the σύνολο εκπαίδευσης).

Βλέπε επίσης: regression, υπόθεση, ταξινομητής, ετικέτα, feature vector, data point, λογιστική απώλεια, labeled datapoint, σύνολο εκπαίδευσης.

μαλακή συσταδοποίηση Soft συσταδοποίηση refers to the task of partitioning a given set of data points into (a few) overlapping συστάδας. Each data point is assigned to several different συστάδας with varying degrees of belonging. Soft συσταδοποίηση methods determine the βαθμός συσχέτισης (or soft συστάδα assignment) for each data point and

each συστάδα. A principled approach to soft συσταδοποίηση is by interpreting data points as i.i.d. realizations of a Gaussian mixture model (GMM). We then obtain a natural choice for the βαθμός συσχέτισης as the conditional probability of a data point belonging to a specific mixture component.

Βλέπε επίσης: συσταδοποίηση, data point, συστάδα, βαθμός συσχέτισης, i.i.d., realization, GMM, probability.

μεγάλο γλωσσικό μοντέλο Large language models (large language model; LLM) is an umbrella term for ml methods that process and generate human-like text. These methods typically use βαθύ δίκτυοs with billions (or even trillions) of παράμετροι. A widely used choice for the network architecture is referred to as Transformers [56]. The training of large language models is often based on the task of predicting a few words that are intentionally removed from a large text corpus. Thus, we can construct labeled datapoints simply by selecting some words of a text as ετικέταs and the remaining words as features of data points. This construction requires very little human supervision and allows for generating sufficiently large σύνολο εκπαίδευσηςs for large language models.

Βλέπε επίσης: model, ml, βαθύ δίκτυο, παράμετροι, labeled datapoint, feature, data point, σύνολο εκπαίδευσης.

μέγεθος δείγματος The number of individual data points contained in a σύνολο δεδομένων.

Βλέπε επίσης: data point, σύνολο δεδομένων.

μείωση της διαστασιμότητας Dimensionality reduction methods map (typically many) raw features to a (relatively small) set of new features. These methods can be used to visualize data points by learning two features that can be used as the coordinates of a depiction in a scatterplot. Βλέπε επίσης: feature, data point, scatterplot.

μεροληψία Consider an ml method using a parametrized χώρος υποθέσεων $\mathcal{H}$. It learns the παράμετροι μοντέλου $\mathbf{w} \in \mathbb{R}^d$ using the σύνολο δεδομένων

$$\mathcal{D} = \left\{ \left(\mathbf{x}^{(r)}, y^{(r)}\right) \right\}_{r=1}^{m}.$$

To analyze the properties of the ml method, we typically interpret the data points as realizations of i.i.d. RVs,

$$y^{(r)} = h^{(\overline{\mathbf{w}})}\left(\mathbf{x}^{(r)}\right) + \boldsymbol{\varepsilon}^{(r)}, r = 1, \ldots, m.$$

We can then interpret the ml method as an estimator $\widehat{\mathbf{w}}$ computed from $\mathcal{D}$ (e.g., by solving εμπειρική ελαχιστοποίηση διακινδύνευσης). The (squared) bias incurred by the estimate $\widehat{\mathbf{w}}$ is then defined as $B^2 := \left\|\mathbb{E}\{\widehat{\mathbf{w}}\} - \overline{\mathbf{w}}\right\|_2^2$.

Βλέπε επίσης: ml, χώρος υποθέσεων, παράμετροι μοντέλου, σύνολο δεδομένων, data point, realization, i.i.d., RV, εμπειρική ελαχιστοποίηση διακινδύνευσης.

μέση τιμή The mean of an RV $\mathbf{x}$, taking values in an Euclidean space $\mathbb{R}^d$, is its expectation $\mathbb{E}\{\mathbf{x}\}$. It is defined as the Lebesgue integral of $\mathbf{x}$ with respect to the underlying κατανομή πιθανότητας P (e.g., see [49] or [2]), i.e.,

$$\mathbb{E}\{\mathbf{x}\} = \int_{\mathbb{R}^d} \mathbf{x} \, \mathrm{d}P(\mathbf{x}).$$

We also use the term to refer to the average of a finite sequence $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \in \mathbb{R}^d$. However, these two definitions are essentially the same. Indeed, we can use the sequence $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \in \mathbb{R}^d$ to construct a discrete RV $\widetilde{\mathbf{x}} = \mathbf{x}^{(I)}$, with the index I being chosen uniformly at random from the set $\{1, \ldots, m\}$. The mean of $\widetilde{\mathbf{x}}$ is precisely the average $\frac{1}{m} \sum_{r=1}^{m} \mathbf{x}^{(r)}$.

Βλέπε επίσης: RV, Euclidean space, expectation, κατανομή πιθανότητας.

μέση τιμή δείγματος The δείγμα μέση τιμή $\mathbf{m} \in \mathbb{R}^d$ for a given σύνολο δεδομένων, with feature vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \in \mathbb{R}^d$, is defined as

$$\mathbf{m} = (1/m) \sum_{r=1}^{m} \mathbf{x}^{(r)}.$$

Βλέπε επίσης: δείγμα, μέση τιμή, σύνολο δεδομένων, feature vector.

μέσο τετραγωνικό σφάλμα εκτίμησης Consider an ml method that learns παράμετροι μοντέλου $\widehat{\mathbf{w}}$ based on some σύνολο δεδομένων $\mathcal{D}$. If we interpret the data points in $\mathcal{D}$ as i.i.d. realizations of an RV $\mathbf{z}$, we define the σφάλμα εκτίμησης $\Delta\mathbf{w} := \widehat{w} - \overline{\mathbf{w}}$. Here, $\overline{\mathbf{w}}$ denotes the true παράμετροι μοντέλου of the κατανομή πιθανότητας of $\mathbf{z}$. The μέση τιμή squared σφάλμα εκτίμησης (mean squared estimation error; MSEE) is defined as the expectation $\mathbb{E}\{\|\Delta\mathbf{w}\|^2\}$ of the squared Euclidean νόρμα of the σφάλμα εκτίμησης [20], [57].

Βλέπε επίσης: ml, παράμετροι μοντέλου, σύνολο δεδομένων, data point, i.i.d., realization, RV, σφάλμα εκτίμησης, κατανομή πιθανότητας, μέση τιμή, expectation, νόρμα.

μη λεία We refer to a function as non-smooth if it is not λεία [53].

Βλέπε επίσης: λεία.

μηχανή διανυσμάτων υποστήριξης (ΜΔΥ) The SVM (support vector machine; SVM) is a binary ταξινόμηση method that learns a linear υπόθεση map. Thus, like γραμμική παλινδρόμηση and λογιστική παλινδρόμηση, it is also an instance of εμπειρική ελαχιστοποίηση διακινδύνευσης for the γραμμικό μοντέλο. However, the SVM uses a different συνάρτηση απώλειας from the one used in those methods. As illustrated in Fig. 10, it aims to maximally separate data points from the two different classes in the feature space (i.e., ολικό μέγιστο margin principle). Maximizing this separation is equivalent to minimizing a regularized variant of the απώλεια άρθρωσης (1) [58], [15], [59].

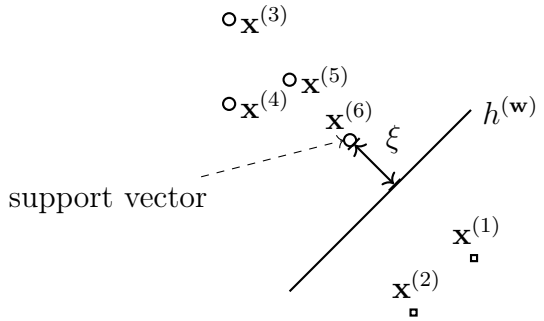


Fig. 10. The μηχανή διανυσμάτων υποστήριξης learns a υπόθεση (or ταξινομητής) $h^{(\mathbf{w})}$ with minimal average soft-margin απώλεια άρθρωσης. Minimizing this loss is equivalent to maximizing the margin ξ between the όριο απόφασης of $h^{(\mathbf{w})}$ and each class of the σύνολο εκπαίδευσης.

The above basic variant of SVM is only useful if the data points from

different categories can be (approximately) linearly separated. For an ml application where the categories are not derived from a kernel.

Βλέπε επίσης: ταξινόμηση, υπόθεση, γραμμική παλινδρόμηση, λογιστική παλινδρόμηση, εμπειρική ελαχιστοποίηση διακινδύνευσης, γραμμικό μοντέλο, συνάρτηση απώλειας, data point, feature space, maximum, απώλεια άρθρωσης, μηχανή διανυσμάτων υποστήριξης, ταξινομητής, loss, όριο απόφασης, σύνολο εκπαίδευσης.

μηχανική μάθηση ML (machine learning; ML) aims to predict a ετικέτα from the features of a data point. ML methods achieve this by learning a υπόθεση from a χώρος υποθέσεων (or model) through the minimization of a συνάρτηση απώλειας [6], [60]. One precise formulation of this principle is εμπειρική ελαχιστοποίηση διακινδύνευσης. Different ML methods are obtained from different design choices for data points (their features and ετικέτα), model, and συνάρτηση απώλειας [6, Ch. 3].

Βλέπε επίσης: ετικέτα, feature, data point, υπόθεση, χώρος υποθέσεων, model, συνάρτηση απώλειας, εμπειρική ελαχιστοποίηση διακινδύνευσης.

μοντέλο In the context of ml methods, the term model typically refers to the χώρος υποθέσεων employed by an ml method [6], [7].

Βλέπε επίσης: ml, χώρος υποθέσεων.

μοντέλο στοχαστικής ομάδας The stochastic block model (stochastic block model; SBM) is a probabilistic generative model for an undirected graph $\mathcal{G} = \big(\mathcal{V}, \mathcal{E}\big)$ with a given set of nodes $\mathcal{V}$ [61]. In its most basic variant, the stochastic block model generates a graph by first randomly assigning each node $i \in \mathcal{V}$ to a συστάδα index $c_i \in \{1, \ldots, k\}$. A pair of

different nodes in the graph is connected by an edge with probability $p_{i,i'}$ that depends solely on the ετικέτας $c_i, c_{i'}$. The presence of edges between different pairs of nodes is statistically independent.

Βλέπε επίσης: model, graph, συστάδα, probability, ετικέτα.

νόμος των μεγάλων αριθμών The law of large numbers refers to the convergence of the average of an increasing (large) number of i.i.d. RVs to the μέση τιμή of their common κατανομή πιθανότητας. Different instances of the law of large numbers are obtained by using different notions of convergence [62].

Βλέπε επίσης: i.i.d., RV, μέση τιμή, κατανομή πιθανότητας.

νόρμα A norm is a function that maps each (vector) element of a linear vector space to a non-negative real number. This function must be homogeneous and definite, and it must satisfy the triangle inequality [63].

ολική μεταβολή See GTV.

ολικό ελάχιστο Given a set of real numbers, the minimum is the smallest of those numbers.

ολικό μέγιστο The maximum of a set $\mathcal{A} \subseteq \mathbb{R}$ of real numbers is the greatest element in that set, if such an element exists. A set $\mathcal{A}$ has a maximum if it is bounded above and attains its supremum (or least upper bound) [2, Sec. 1.4].

Βλέπε επίσης: supremum.

οριζόντια ομοσπονδιακή μάθηση HFL (horizontal federated learning; HFL) uses τοπικό σύνολο δεδομένωνs constituted by different data

points but uses the same features to characterize them [64]. For example, weather forecasting uses a network of spatially distributed weather (observation) stations. Each weather station measures the same quantities, such as daily temperature, air pressure, and precipitation. However, different weather stations measure the characteristics or features of different spatiotemporal regions. Each spatiotemporal region represents an individual data point, each characterized by the same features (e.g., daily temperature or air pressure).

Βλέπε επίσης: FL, τοπικό σύνολο δεδομένων, data point, feature.

όριο απόφασης Consider a υπόθεση map h that reads in a feature vector $\mathbf{x} \in \mathbb{R}^d$ and delivers a value from a finite set $\mathcal{Y}$. The decision boundary of h is the set of vectors $\mathbf{x} \in \mathbb{R}^d$ that lie between different περιοχή αποφάσεωνs. More precisely, a vector $\mathbf{x}$ belongs to the decision boundary if and only if each neighborhood $\{\mathbf{x}' : \|\mathbf{x} - \mathbf{x}'\| \leq \varepsilon\}$, for any $\varepsilon > 0$, contains at least two vectors with different function values.

Βλέπε επίσης: υπόθεση, feature, περιοχή αποφάσεων, neighborhood.

παλινδρόμηση ελάχιστης απόλυτης απόκλισης Least absolute deviation regression is an instance of εμπειρική ελαχιστοποίηση διακινδύνευσης using the απώλεια απόλυτου σφάλματος. It is a special case of παλινδρόμηση Huber.

Βλέπε επίσης: εμπειρική ελαχιστοποίηση διακινδύνευσης, απώλεια απόλυτου σφάλματος, παλινδρόμηση Huber.

παλινδρόμηση Huber Huber regression refers to εμπειρική ελαχιστοποίηση διακινδύνευσης-based methods that use the απώλεια Huber as a measure

of the πρόβλεψη error. Two important special cases of Huber regression are παλινδρόμηση ελάχιστης απόλυτης απόκλισης and γραμμική παλινδρόμηση. Tuning the threshold parameter of the απώλεια Huber allows the user to trade the robustness of the απώλεια απόλυτου σφάλματος against the computational benefits of the λεία τετραγωνική απώλεια σφάλματος.

Βλέπε επίσης: regression, εμπειρική ελαχιστοποίηση διακινδύνευσης, απώλεια Huber, πρόβλεψη, regression, παλινδρόμηση ελάχιστης απόλυτης απόκλισης, γραμμική παλινδρόμηση, απώλεια απόλυτου σφάλματος, λεία, τετραγωνική απώλεια σφάλματος.

παραγωγίσιμη A real-valued function $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable if it can, at any point, be approximated locally by a linear function. The local linear approximation at the point $\mathbf{x}$ is determined by the gradient $\nabla f(\mathbf{x})$ [2].

Βλέπε επίσης: gradient.

παραδοχή συσταδοποίησης The συσταδοποίηση assumption postulates that data points in a σύνολο δεδομένων form a (small) number of groups or clusters. Data points in the same συστάδα are more similar to each other than those outside the συστάδα [65]. We obtain different συσταδοποίηση methods by using different notions of similarity between data points.

Βλέπε επίσης: συσταδοποίηση, data point, σύνολο δεδομένων, συστάδα.

παράμετροι The parameters of an ml model are tunable (i.e., learnable or adjustable) quantities that allow us to choose between different υπόθεση

maps. For example, the γραμμικό μοντέλο $\mathcal{H} := \{h^{(\mathbf{w})} : h^{(\mathbf{w})}(x) = w_1 x + w_2\}$ consists of all υπόθεση maps $h^{(\mathbf{w})}(x) = w_1 x + w_2$ with a particular choice for the parameters $\mathbf{w} = \big(w_1, w_2\big)^T \in \mathbb{R}^2$. Another example of parameters is the βάρη assigned to the connections between neurons of an τεχνητό νευρωνικό δίκτυο.

Βλέπε επίσης: ml, model, υπόθεση, γραμμικό μοντέλο, βάρη, τεχνητό νευρωνικό δίκτυο.

παράμετροι μοντέλου Model παράμετροι are quantities that are used to select a specific υπόθεση map from a model. We can think of a list of model παράμετροι as a unique identifier for a υπόθεση map, similar to how a social security number identifies a person in Finland.

Βλέπε επίσης: model, παράμετροι, υπόθεση.

περιοχή αποφάσεων Consider a υπόθεση map h that delivers values from a finite set $\mathcal{Y}$. For each ετικέτα value (category) $a \in \mathcal{Y}$, the υπόθεση h determines a subset of feature values $\mathbf{x} \in \mathcal{X}$ that result in the same output $h(\mathbf{x}) = a$. We refer to this subset as a decision region of the υπόθεση h.

Βλέπε επίσης: υπόθεση, ετικέτα, feature.

πιθανότητα We assign a probability value, typically chosen in the interval $[0, 1]$, to each event that might occur in a random experiment [5], [66], [49], [67].

πιθανοτικό μοντέλο A probabilistic model interprets data points as realizations of RVs with a joint κατανομή πιθανότητας. This joint κατανομή πιθανότητας typically involves παράμετροι which have to be manually

chosen or learned via statistical inference methods such as μέγιστη πιθανοφάνεια estimation [20].

Βλέπε επίσης: model, data point, realization, RV, κατανομή πιθανότητας, παράμετροι, μέγιστη πιθανοφάνεια.

πίνακας σύγχυσης Consider data points characterized by features $\mathbf{x}$ and ετικέτα y having values from the finite label space $\mathcal{Y} = \{1, \ldots, k\}$. The confusion matrix is a $k \times k$ matrix with rows representing different values c of the true label of a data point. The columns of a confusion matrix correspond to different values c' delivered by a hypothesis $h(\mathbf{x})$. The (c, c')-th entry of the confusion matrix is the fraction of data points with the ετικέτα $y = c$ and the πρόβλεψη $\hat{y} = c'$ assigned by the υπόθεση h.

Βλέπε επίσης: data point, feature, ετικέτα, label space, πρόβλεψη, υπόθεση.

πίνακας συνδιακύμανσης The covariance matrix of an RV $\mathbf{x} \in \mathbb{R}^d$ is defined as $\mathbb{E}\Big\{\big(\mathbf{x} - \mathbb{E}\{\mathbf{x}\}\big)\big(\mathbf{x} - \mathbb{E}\{\mathbf{x}\}\big)^T\Big\}$.

Βλέπε επίσης: RV.

πίνακας συνδιακύμανσης δείγματος The sample πίνακας συνδιακύμανσης $\widehat{\mathbf{\Sigma}} \in \mathbb{R}^{d \times d}$ for a given set of feature vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \in \mathbb{R}^d$ is defined as

$$\widehat{\mathbf{\Sigma}} = (1/m) \sum_{r=1}^{m} (\mathbf{x}^{(r)} - \widehat{\mathbf{m}})(\mathbf{x}^{(r)} - \widehat{\mathbf{m}})^T.$$

Here, we use the μέση τιμή δείγματος $\widehat{\mathbf{m}}$.

Βλέπε επίσης: πίνακας συνδιακύμανσης, feature vector, μέση τιμή δείγματος.

πλησιέστερος γείτονας NN (nearest neighbor; NN) methods learn a υπόθεση $h : \mathcal{X} \to \mathcal{Y}$ whose function value $h(\mathbf{x})$ is solely determined by the nearest γείτονες within a given σύνολο δεδομένων. Different methods use different metrics for determining the nearest γείτονες. If data points are characterized by numeric feature vectors, we can use their Euclidean distances as the metric.

Βλέπε επίσης: υπόθεση, γείτονες, σύνολο δεδομένων, data point, feature vector.

πολυμεταβλητή κανονική κατανομή The multivariate normal distribution $\mathcal{N}(\mathbf{m}, \mathbf{C})$ is an important family of κατανομή πιθανότηταςs for a continuous RV $\mathbf{x} \in \mathbb{R}^d$ [5], [22], [68]. This family is parametrized by the μέση τιμή $\mathbf{m}$ and the πίνακας συνδιακύμανσης $\mathbf{C}$ of $\mathbf{x}$. If the πίνακας συνδιακύμανσης is invertible, the κατανομή πιθανότητας of $\mathbf{x}$ is

$$p(\mathbf{x}) \propto \exp\Big(-(1/2)\big(\mathbf{x}-\mathbf{m}\big)^T \mathbf{C}^{-1}\big(\mathbf{x}-\mathbf{m}\big)\Big).$$

Βλέπε επίσης: κατανομή πιθανότητας, RV, μέση τιμή, πίνακας συνδιακύμανσης.

πολυωνυμική παλινδρόμηση Polynomial regression aims at learning a polynomial υπόθεση map to predict a numeric ετικέτα based on the numeric features of a data point. For data points characterized by a single numeric feature, polynomial regression uses the χώρος υποθέσεων $\mathcal{H}_d^{(\mathrm{poly})} := \{h(x) = \sum_{j=0}^{d-1} x^j w_j\}$. The quality of a polynomial υπόθεση map is measured using the average τετραγωνική απώλεια σφάλματος incurred on a set of labeled datapoints (which we refer to as the σύνολο

εκπαίδευσης).

Βλέπε επίσης: regression, υπόθεση, ετικέτα, feature, data point, χώρος υποθέσεων, τετραγωνική απώλεια σφάλματος, labeled datapoint, σύνολο εκπαίδευσης.

προβεβλημένη κάθοδος κλίσης Consider an εμπειρική ελαχιστοποίηση διακινδύνευσης-based method that uses a parametrized model with χώρος παραμέτρων $\mathcal{W} \subseteq \mathbb{R}^d$. Even if the objective function of εμπειρική ελαχιστοποίηση διακινδύνευσης is λεία, we cannot use basic κάθοδος κλίσης, as it does not take into account contraints on the optimization variable (i.e., the παράμετροι μοντέλου). Projected κάθοδος κλίσης (projected gradient descent; projected GD) extends basic κάθοδος κλίσης to handle constraints on the optimization variable (i.e., the παράμετροι μοντέλου). A single iteration of projected κάθοδος κλίσης consists of first taking a βήμα κλίσης and then projecting the result back onto the χώρος παραμέτρων.

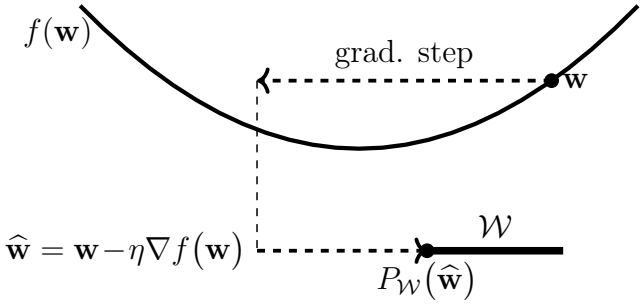


Fig. 11. Projected κάθοδος κλίσης augments a basic βήμα κλίσης with a προβολή back onto the constraint set $\mathcal{W}$.

Βλέπε επίσης: εμπειρική ελαχιστοποίηση διακινδύνευσης, model, χώρος

παραμέτρων, objective function, λεία, κάθοδος κλίσης, παράμετροι μοντέλου, βήμα κλίσης, προβολή.

πρόβλεψη A prediction is an estimate or approximation for some quantity of interest. Ml revolves around learning or finding a υπόθεση map h that reads in the features $\mathbf{x}$ of a data point and delivers a prediction $\widehat{y} := h(\mathbf{x})$ for its ετικέτα y.

Βλέπε επίσης: ml, υπόθεση, feature, data point, ετικέτα.

προβολή Consider a subset $\mathcal{W} \subseteq \mathbb{R}^d$ of the d-dimensional Euclidean space. We define the projection $P_{\mathcal{W}}\big(\mathbf{w}\big)$ of a vector $\mathbf{w} \in \mathbb{R}^d$ onto $\mathcal{W}$ as

$$P_{\mathcal{W}}\big(\mathbf{w}\big) = \operatorname*{argmin}_{\mathbf{w}' \in \mathcal{W}} \left\|\mathbf{w} - \mathbf{w}'\right\|_2 . \tag{7}$$

In other words, $P_{\mathcal{W}}\big(\mathbf{w}\big)$ is the vector in $\mathcal{W}$ which is closest to $\mathbf{w}$. The projection is only well-defined for subsets $\mathcal{W}$ for which the above ολικό ελάχιστο exists [52].

Βλέπε επίσης: Euclidean space, ολικό ελάχιστο.

προσδοκία Consider a numeric feature vector $\mathbf{x} \in \mathbb{R}^d$ which we interpret as the realization of an RV with a κατανομή πιθανότητας $p(\mathbf{x})$. The expectation of $\mathbf{x}$ is defined as the integral $\mathbb{E}\{\mathbf{x}\} := \int \mathbf{x} p(\mathbf{x})$ [2], [66], [49]. Note that the expectation is only defined if this integral exists, i.e., if the RV is integrable.

Βλέπε επίσης: feature vector, realization, RV, κατανομή πιθανότητας.

προσεγγίσιμος A convex function for which the εγγύς τελεστής can be computed efficiently is sometimes referred to as proximable or simple

[69].

Βλέπε επίσης: convex, εγγύς τελεστής.

προστασία της ιδιωτικότητας Consider some ml method $\mathcal{A}$ that reads in a σύνολο δεδομένων $\mathcal{D}$ and delivers some output $\mathcal{A}(\mathcal{D})$. The output could be the learned παράμετροι μοντέλου $\widehat{\mathbf{w}}$ or the πρόβλεψη $\hat{h}(\mathbf{x})$ obtained for a specific data point with features $\mathbf{x}$. Many important ml applications involve data points representing humans. Each data point is characterized by features $\mathbf{x}$, potentially a ετικέτα y, and a ευαίσθητο ιδιοχαρακτηριστικό s (e.g., a recent medical diagnosis). Roughly speaking, privacy protection means that it should be impossible to infer, from the output $\mathcal{A}(\mathcal{D})$, any of the ευαίσθητο ιδιοχαρακτηριστικός of data points in $\mathcal{D}$. Mathematically, privacy protection requires non-invertibility of the map $\mathcal{A}(\mathcal{D})$. In general, just making $\mathcal{A}(\mathcal{D})$ non-invertible is typically insufficient for privacy protection. We need to make $\mathcal{A}(\mathcal{D})$ sufficiently non-invertible.

Βλέπε επίσης: ml, σύνολο δεδομένων, παράμετροι μοντέλου, πρόβλεψη, data point, feature, ετικέτα, ευαίσθητο ιδιοχαρακτηριστικό.

σημείο δεδομένων A data point is any object that conveys information [13]. Data points might be students, radio signals, trees, forests, images, RVs, real numbers, or proteins. We characterize data points using two types of properties. One type of property is referred to as a feature. Features are properties of a data point that can be measured or computed in an automated fashion. A different kind of property is referred to a ετικέτα. The ετικέτα of a data point represents some higher-level fact

(or quantity of interest). In contrast to features, determining the ετικέτα of a data point typically requires human experts (domain experts). Roughly speaking, ml aims to predict the ετικέτα of a data point based solely on its features.

Βλέπε επίσης: data, RV, feature, ετικέτα, ml.

σκληρή συσταδοποίηση Hard συσταδοποίηση refers to the task of partitioning a given set of data points into (a few) non-overlapping συστάδας. The most widely used hard συσταδοποίηση method is αλγόριθμος k-μέσων.

Βλέπε επίσης: συσταδοποίηση, data point, συστάδα, αλγόριθμος k-μέσων.

στατιστικές διαστάσεις By statistical aspects of an ml method, we refer to (properties of) the κατανομή πιθανότητας of its output under a πιθανοτικό μοντέλο for the data fed into the method.

Βλέπε επίσης: ml, κατανομή πιθανότητας, πιθανοτικό μοντέλο, data.

στοχαστική κάθοδος κλίσης Stochastic κάθοδος κλίσης (stochastic gradient descent; SGD) is obtained from κάθοδος κλίσης by replacing the gradient of the objective function with a stochastic approximation. A main application of stochastic κάθοδος κλίσης is to train a parametrized model via εμπειρική ελαχιστοποίηση διακινδύνευσης on a σύνολο εκπαίδευσης $\mathcal{D}$ that is either very large or not readily available (e.g., when data points are stored in a database distributed all over the planet). To evaluate the gradient of the empirical risk (as a function of the παράμετροι μοντέλου $\mathbf{w}$), we need to compute a sum

$\sum_{r=1}^{m} \nabla_{\mathbf{w}} L\left(\mathbf{z}^{(r)}, \mathbf{w}\right)$ over all data points in the σύνολο εκπαίδευσης. We obtain a stochastic approximation to the gradient by replacing the sum $\sum_{r=1}^{m} \nabla_{\mathbf{w}} L\left(\mathbf{z}^{(r)}, \mathbf{w}\right)$ with a sum $\sum_{r \in \mathcal{B}} \nabla_{\mathbf{w}} L\left(\mathbf{z}^{(r)}, \mathbf{w}\right)$ over a randomly chosen subset $\mathcal{B} \subseteq \{1, \ldots, m\}$ (see Fig. 12). We often refer to these randomly chosen data points as a δέσμη. The δέσμη size $|\mathcal{B}|$ is an important parameter of stochastic κάθοδος κλίσης. Stochastic κάθοδος κλίσης with $|\mathcal{B}| > 1$ is referred to as mini-δέσμη stochastic κάθοδος κλίσης [70].

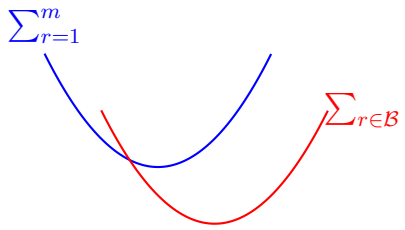


Fig. 12. Stochastic κάθοδος κλίσης for εμπειρική ελαχιστοποίηση διακινδύνευσης approximates the gradient $\sum_{r=1}^{m} \nabla_{\mathbf{w}} L\left(\mathbf{z}^{(r)}, \mathbf{w}\right)$ by replacing the sum over all data points in the σύνολο εκπαίδευσης (indexed by $r = 1, \ldots, m$) with a sum over a randomly chosen subset $\mathcal{B} \subseteq \{1, \ldots, m\}$.

Βλέπε επίσης: κάθοδος κλίσης, gradient, objective function, model, εμπειρική ελαχιστοποίηση διακινδύνευσης, σύνολο εκπαίδευσης, data point, empirical risk, παράμετροι μοντέλου, δέσμη.

συνάρτηση απώλειας A loss function is a map

$$L : \mathcal{X} \times \mathcal{Y} \times \mathcal{H} \to \mathbb{R}_{+} : \big(\big(\mathbf{x}, y\big), h\big) \mapsto L\left((\mathbf{x}, y), h\right).$$

It assigns a non-negative real number (i.e., the loss) $L\left((\mathbf{x}, y), h\right)$ to a pair that consists of a data point, with features $\mathbf{x}$ and ετικέτα y, and

a υπόθεση $h \in \mathcal{H}$. The value $L\left((\mathbf{x}, y), h\right)$ quantifies the discrepancy between the true ετικέτα y and the πρόβλεψη $h(\mathbf{x})$. Lower (closer to zero) values $L\left((\mathbf{x}, y), h\right)$ indicate a smaller discrepancy between πρόβλεψη $h(\mathbf{x})$ and ετικέτα y. Fig. 13 depicts a loss function for a given data point, with features $\mathbf{x}$ and ετικέτα y, as a function of the υπόθεση $h \in \mathcal{H}$.

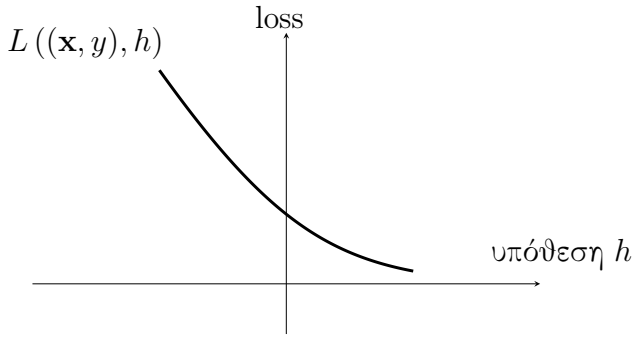


Fig. 13. Some loss function $L\left((\mathbf{x}, y), h\right)$ for a fixed data point, with feature vector $\mathbf{x}$ and ετικέτα y, and a varying υπόθεση h. ml methods try to find (or learn) a υπόθεση that incurs minimal loss.

Βλέπε επίσης: loss, data point, feature, ετικέτα, υπόθεση, πρόβλεψη, feature vector, ml.

συνάρτηση ενεργοποίησης Each artificial neuron within an τεχνητό νευρωνικό δίκτυο is assigned an activation function $\sigma(\cdot)$ that maps a weighted combination of the neuron inputs $x_1, \ldots, x_d$ to a single output value $a = \sigma\big(w_1 x_1 + \ldots + w_d x_d\big)$. Note that each neuron is parametrized by the βάρη $w_1, \ldots, w_d$.

Βλέπε επίσης: τεχνητό νευρωνικό δίκτυο, βάρη.

συνάρτηση πυκνότητας πιθανότητας The probability density func-

tion $p(x)$ (probability density function; pdf) of a real-valued RV $x \in \mathbb{R}$ is a particular representation of its κατανομή πιθανότητας. If the probability density function exists, it can be used to compute the probability that x takes on a value from a (measurable) set $\mathcal{B} \subseteq \mathbb{R}$ via $p(x \in \mathcal{B}) = \int_{\mathcal{B}} p(x')dx'$ [5, Ch. 3]. The probability density function of a vector-valued RV $\mathbf{x} \in \mathbb{R}^d$ (if it exists) allows us to compute the probability of $\mathbf{x}$ belonging to a (measurable) region $\mathcal{R}$ via $p(\mathbf{x} \in \mathcal{R}) = \int_{\mathcal{R}} p(\mathbf{x}')dx'_1 \ldots dx'_d$ [5, Ch. 3].

Βλέπε επίσης: probability, RV, κατανομή πιθανότητας.

συνδεδεμένος γράφος An undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is connected if every non-empty subset $\mathcal{V}' \subset \mathcal{V}$ has at least one edge connecting it to $\mathcal{V} \setminus \mathcal{V}'$.

Βλέπε επίσης: graph.

συνθήκη μηδενικής κλίσης Consider the unconstrained optimization problem $\min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w})$ with a λεία and convex objective function $f(\mathbf{w})$. A necessary and sufficient condition for a vector $\widehat{\mathbf{w}} \in \mathbb{R}^d$ to solve this problem is that the gradient $\nabla f\big(\widehat{\mathbf{w}}\big)$ is the zero vector,

$$\nabla f\big(\widehat{\mathbf{w}}\big) = \mathbf{0} \Leftrightarrow f\big(\widehat{\mathbf{w}}\big) = \min_{\mathbf{w} \in \mathbb{R}^d} f(\mathbf{w}).$$

Βλέπε επίσης: λεία, convex, objective function, gradient.

σύνολο δεδομένων A dataset refers to a collection of data points. These data points carry information about some quantity of interest (or ετικέτα) within an ml application. ml methods use datasets for model training

(e.g., via εμπειρική ελαχιστοποίηση διακινδύνευσης) and model επικύρωση. Note that our notion of a dataset is very flexible, as it allows for very different types of data points. Indeed, data points can be concrete physical objects (such as humans or animals) or abstract objects (such as numbers). As a case in point, Fig. 14 depicts a dataset that consists of cows as data points.

Fig. 14. "Cows in the Swiss Alps" by User:Huhu Uet is licensed under [CC BY-SA 4.0](https://creativecommons.org/licenses/by-sa/4.0/)

Quite often, an ml engineer does not have direct access to a dataset. Indeed, accessing the dataset in Fig. 14 would require us to visit the cow herd in the Alps. Instead, we need to use an approximation (or representation) of the dataset which is more convenient to work with. Different mathematical models have been developed for the representation (or approximation) of datasets [71], [72], [73], [74]. One of the most widely adopted data model is the relational model, which organizes data as a table (or relation) [27], [71]. A table consists of rows

and columns:

- Each row of the table represents a single data point.
- Each column of the table corresponds to a specific attribute of the data point. ml methods can use attributes as features and ετικέτας of the data point.

For example, Table 1 shows a representation of the dataset in Fig. 14. In the relational model, the order of rows is irrelevant, and each attribute (i.e., column) must be precisely defined with a domain, which specifies the set of possible values. In ml applications, these attribute domains become the feature space and the label space.

Name	**Weight**	**Age**	**Height**	**Stomach temp**
Zenzi	100	4	100	25
Berta	140	3	130	23
Resi	120	4	120	31

Table 1: A relation (or table) that represents the dataset in Fig. 14.

While the relational model is useful for the study of many ml applications, it may be insufficient regarding the requirements for trustworthy AI. Modern approaches like datasheets for datasets provide more comprehensive documentation, including details about the dataset's collection process, intended use, and other contextual information [33].

Βλέπε επίσης: data point, ετικέτα, ml, model, εμπειρική ελαχιστοποίηση διακινδύνευσης, επικύρωση, data, feature, feature space, label space, trustworthy AI.

σύνολο εκπαίδευσης A training set is a σύνολο δεδομένων $\mathcal{D}$ which consists of some data points used in εμπειρική ελαχιστοποίηση διακινδύνευσης to learn a υπόθεση $\hat{h}$. The average loss of $\hat{h}$ on the training set is referred to as the training error. The comparison of the training error with the σφάλμα επικύρωσης of $\hat{h}$ allows us to diagnose the ml method and informs how to improve the validation error (e.g., using a different χώρος υποθέσεων or collecting more data points) [6, Sec. 6.6].

Βλέπε επίσης: σύνολο δεδομένων, data point, εμπειρική ελαχιστοποίηση διακινδύνευσης, υπόθεση, loss, training error, σφάλμα επικύρωσης, ml, χώρος υποθέσεων.

σύνολο επικύρωσης A set of data points used to estimate the διακινδύνευση of a υπόθεση $\hat{h}$ that has been learned by some ml method (e.g., solving εμπειρική ελαχιστοποίηση διακινδύνευσης). The average loss of $\hat{h}$ on the επικύρωση set is referred to as the σφάλμα επικύρωσης and can be used to diagnose an ml method (see [6, Sec. 6.6]). The comparison between training error and σφάλμα επικύρωσης can inform directions for improvement of the ml method (such as using a different χώρος υποθέσεων).

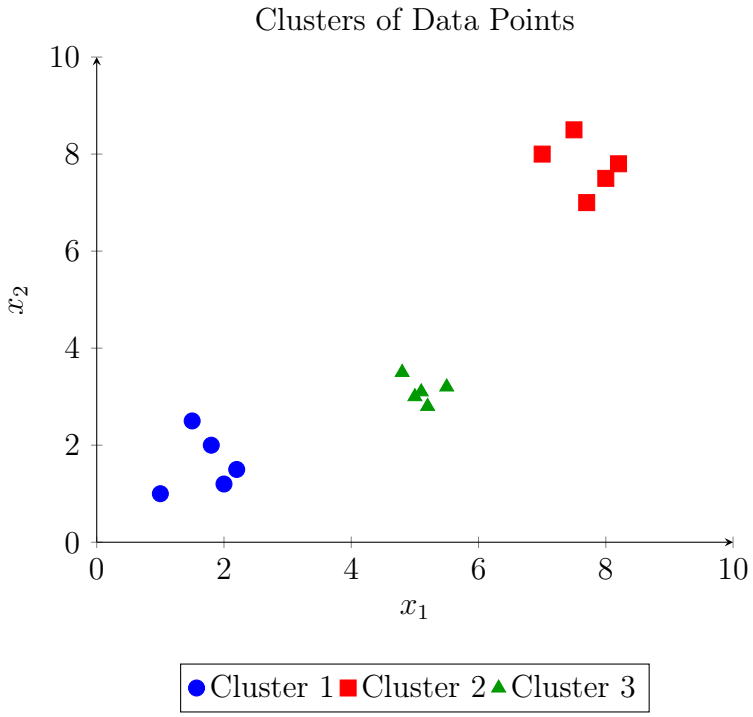
Βλέπε επίσης: data point, διακινδύνευση, υπόθεση, ml, εμπειρική ελαχιστοποίηση διακινδύνευσης, loss, επικύρωση, σφάλμα επικύρωσης, training error, χώρος υποθέσεων.

συσκευή Any physical system that can be used to store and process data. In the context of ml, we typically mean a computer that is able to read in data points from different sources and, in turn, to train an ml model

using these data points.

Βλέπε επίσης: data, ml, data point, model.

συστάδα A cluster is a subset of data points that are more similar to each other than to the data points outside the cluster. The quantitative measure of similarity between data points is a design choice. If data points are characterized by Euclidean feature vectors $\mathbf{x} \in \mathbb{R}^d$, we can define the similarity between two data points via the Euclidean distance between their feature vectors.



Βλέπε επίσης: data point, feature vector.

συσταδοποίηση Clustering methods decompose a given set of data points into a few subsets, which are referred to as συστάδας. Each συστάδα

consists of data points that are more similar to each other than to data points outside the συστάδα. Different clustering methods use different measures for the similarity between data points and different forms of συστάδα representations. The clustering method αλγόριθμος k-μέσων uses the average feature vector (cluster μέση τιμή) of a συστάδα as its representative. A popular soft clustering method based on GMM represents a συστάδα by a πολυμεταβλητή κανονική κατανομή.

Βλέπε επίσης: data point, συστάδα, αλγόριθμος k-μέσων, feature, μέση τιμή, soft clustering, GMM, πολυμεταβλητή κανονική κατανομή.

συσταδοποίηση γράφου Graph συσταδοποίηση aims at συσταδοποίηση data points that are represented as the nodes of a graph $\mathcal{G}$. The edges of $\mathcal{G}$ represent pairwise similarities between data points. Sometimes we can quantify the extend of these similarities by an βάρος ακμής [75], [76].

Βλέπε επίσης: graph, συσταδοποίηση, data point, βάρος ακμής.

συσταδοποίηση με βάση τη ροή Flow-based συσταδοποίηση groups the nodes of an undirected graph by applying αλγόριθμος k-μέσων συσταδοποίηση to node-wise feature vectors. These feature vectors are built from network flows between carefully selected sources and destination nodes [75].

Βλέπε επίσης: συσταδοποίηση, graph, αλγόριθμος k-μέσων, feature vector.

σφάλμα εκτίμησης Consider data points, each with feature vector $\mathbf{x}$ and ετικέτα y. In some applications, we can model the relation between the feature vector and the ετικέτα of a data point as $y = \bar{h}(\mathbf{x}) + \varepsilon$. Here, we

use some true underlying υπόθεση $\bar{h}$ and a noise term ε which summarizes any modeling or labeling errors. The estimation error incurred by an ml method that learns a υπόθεση $\widehat{h}$, e.g., using εμπειρική ελαχιστοποίηση διακινδύνευσης, is defined as $\widehat{h}(\mathbf{x}) - \bar{h}(\mathbf{x})$, for some feature vector. For a parametric χώρος υποθέσεων, which consists of υπόθεση maps determined by παράμετροι μοντέλου $\mathbf{w}$, we can define the estimation error as $\Delta\mathbf{w} = \widehat{\mathbf{w}} - \overline{\mathbf{w}}$ [57], [77].

Βλέπε επίσης: data point, feature vector, ετικέτα, υπόθεση, ml, εμπειρική ελαχιστοποίηση διακινδύνευσης, χώρος υποθέσεων, παράμετροι μοντέλου.

σφάλμα επικύρωσης Consider a υπόθεση $\hat{h}$ which is obtained by some ml method, e.g., using εμπειρική ελαχιστοποίηση διακινδύνευσης on a σύνολο εκπαίδευσης. The average loss of $\hat{h}$ on a σύνολο επικύρωσης, which is different from the σύνολο εκπαίδευσης, is referred to as the επικύρωση error.

Βλέπε επίσης: υπόθεση, ml, εμπειρική ελαχιστοποίηση διακινδύνευσης, σύνολο εκπαίδευσης, loss, σύνολο επικύρωσης, επικύρωση.

ταξινόμηση Classification is the task of determining a discrete-valued label y for a given data point, based solely on its features $\mathbf{x}$. The label y belongs to a finite set, such as $y \in \{-1, 1\}$ or $y \in \{1, \ldots, 19\}$, and represents the category to which the corresponding data point belongs.

Βλέπε επίσης: data point.

ταξινομητής A classifier is a υπόθεση (map) $h(\mathbf{x})$ used to predict a ετικέτα taking values from a finite label space. We might use the function value

$h(\mathbf{x})$ itself as a πρόβλεψη $\hat{y}$ for the ετικέτα. However, it is customary to use a map $h(\cdot)$ that delivers a numeric quantity. The πρόβλεψη is then obtained by a simple thresholding step. For example, in a binary ταξινόμηση problem with $\mathcal{Y} \in \{-1, 1\}$, we might use a real-valued υπόθεση map $h(\mathbf{x}) \in \mathbb{R}$ as a classifier. A πρόβλεψη $\hat{y}$ can then be obtained via thresholding,

$$\hat{y} = 1 \text{ for } h(\mathbf{x}) \geq 0 \text{ and } \hat{y} = -1 \text{ otherwise.} \tag{8}$$

We can characterize a classifier by its περιοχή αποφάσεωνs $\mathcal{R}_a$, for every possible ετικέτα value $a \in \mathcal{Y}$.

Βλέπε επίσης: υπόθεση, ετικέτα, label space, πρόβλεψη, ταξινόμηση.

τετραγωνική απώλεια σφάλματος The squared error loss measures the πρόβλεψη error of a υπόθεση h when predicting a numeric ετικέτα $y \in \mathbb{R}$ from the features $\mathbf{x}$ of a data point. It is defined as

$$L\left((\mathbf{x}, y), h\right) := \big(y - \underbrace{h(\mathbf{x})}_{=\hat{y}}\big)^2.$$

Βλέπε επίσης: loss, πρόβλεψη, υπόθεση, ετικέτα, feature, data point.

τεχνητή νοημοσύνη AI (artificial intelligence;AI) refers to systems that behave rationally in the sense of maximizing a long-term ανταμοιβή. The ml-based approach to AI is to train a model for predicting optimal actions. These predictions are computed from observations about the state of the environment. The choice of συνάρτηση απώλειας sets AI applications apart from more basic ml applications. AI systems rarely have access to a labeled σύνολο εκπαίδευσης that allows the average loss

to be measured for any possible choice of παράμετροι μοντέλου. Instead, AI systems use observed ανταμοιβή signals to obtain a (point-wise) estimate for the loss incurred by the current choice of παράμετροι μοντέλου.
Βλέπε επίσης: ανταμοιβή, ml, model, συνάρτηση απώλειας, σύνολο εκπαίδευσης, loss, παράμετροι μοντέλου.

τεχνητό νευρωνικό δίκτυο (ΤΝΔ) An ANN (artificial neural network; ANN) is a graphical (signal-flow) representation of a function that maps features of a data point at its input to a πρόβλεψη for the corresponding ετικέτα at its output. The fundamental unit of an ANN is the artificial neuron, which applies an συνάρτηση ενεργοποίησης to its weighted inputs. The outputs of these neurons serve as inputs for other neurons, forming interconnected layers.
Βλέπε επίσης: feature, data point, πρόβλεψη, ετικέτα, συνάρτηση ενεργοποίησης.

τοπικό μοντέλο Consider a collection of τοπικό σύνολο δεδομένωνs that are assigned to the nodes of an FL network. A local model $\mathcal{H}^{(i)}$ is a χώρος υποθέσεων assigned to a node $i \in \mathcal{V}$. Different nodes might be assigned different χώρος υποθέσεωνs, i.e., in general $\mathcal{H}^{(i)} \neq \mathcal{H}^{(i')}$ for different nodes $i, i' \in \mathcal{V}$.
Βλέπε επίσης: τοπικό σύνολο δεδομένων, FL network, model, χώρος υποθέσεων.

τοπικό σύνολο δεδομένων The concept of a local σύνολο δεδομένων is in between the concept of a data point and a σύνολο δεδομένων. A local σύνολο δεδομένων consists of several individual data points, which

are characterized by features and ετικέτας. In contrast to a single σύνολο δεδομένων used in basic ml methods, a local σύνολο δεδομένων is also related to other local σύνολο δεδομένωνs via different notions of similarity. These similarities might arise from πιθανοτικό μοντέλος or communication infrastructure and are encoded in the edges of an FL network.

Βλέπε επίσης: σύνολο δεδομένων, data point, feature, ετικέτα, ml, πιθανοτικό μοντέλο, FL network.

τυχαίο δάσος A random forest is a set of different decision trees. Each of these decision trees is obtained by fitting a perturbed copy of the original σύνολο δεδομένων.

Βλέπε επίσης: decision tree, σύνολο δεδομένων.

υπερπροσαρμογή Consider an ml method that uses εμπειρική ελαχιστοποίηση διακινδύνευσης to learn a υπόθεση with the ολικό ελάχιστο empirical risk on a given σύνολο εκπαίδευσης. Such a method is overfitting the σύνολο εκπαίδευσης if it learns a υπόθεση with a small empirical risk on the σύνολο εκπαίδευσης but a significantly larger loss outside the σύνολο εκπαίδευσης.

Βλέπε επίσης: ml, εμπειρική ελαχιστοποίηση διακινδύνευσης, υπόθεση, ολικό ελάχιστο, empirical risk, σύνολο εκπαίδευσης, loss.

υπόθεση A hypothesis refers to a map (or function) $h : \mathcal{X} \to \mathcal{Y}$ from the feature space $\mathcal{X}$ to the label space $\mathcal{Y}$. Given a data point with features $\mathbf{x}$, we use a hypothesis map h to estimate (or approximate) the ετικέτα y using the πρόβλεψη $\hat{y} = h(\mathbf{x})$. Ml is all about learning (or finding)

a hypothesis map h such that $y \approx h(\mathbf{x})$ for any data point (having features $\mathbf{x}$ and ετικέτα y).

Βλέπε επίσης: feature space, label space, data point, feature, ετικέτα, πρόβλεψη, ml.

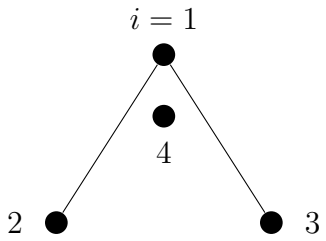
υπολογιστικές διαστάσεις By computational aspects of an ml method, we mainly refer to the computational resources required for its implementation. For example, if an ml method uses iterative optimization techniques to solve εμπειρική ελαχιστοποίηση διακινδύνευσης, then its computational aspects include: 1) how many arithmetic operations are needed to implement a single iteration (βήμα κλίσης); and 2) how many iterations are needed to obtain useful παράμετροι μοντέλου. One important example of an iterative optimization technique is κάθοδος κλίσης.

Βλέπε επίσης: ml, εμπειρική ελαχιστοποίηση διακινδύνευσης, βήμα κλίσης, παράμετροι μοντέλου, κάθοδος κλίσης.

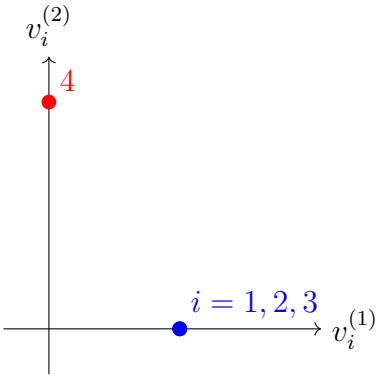
υποπροσαρμογή Consider an ml method that uses εμπειρική ελαχιστοποίηση διακινδύνευσης to learn a υπόθεση with the ολικό ελάχιστο empirical risk on a given σύνολο εκπαίδευσης. Such a method is underfitting the σύνολο εκπαίδευσης if it is not able to learn a υπόθεση with a sufficiently small empirical risk on the σύνολο εκπαίδευσης. If a method is underfitting, it will typically also not be able to learn a υπόθεση with a small διακινδύνευση.

Βλέπε επίσης: ml, εμπειρική ελαχιστοποίηση διακινδύνευσης, υπόθεση, ολικό ελάχιστο, empirical risk, σύνολο εκπαίδευσης, διακινδύνευση.

φασματική συσταδοποίηση Spectral συσταδοποίηση is a particular instance of συσταδοποίηση γράφου, i.e., it clusters data points represented as the nodes $i = 1, \ldots, n$ of a graph $\mathcal{G}$. Spectral συσταδοποίηση uses the ιδιοδιάνυσμαs of the Laplacian matrix $\mathbf{L}^{(\mathcal{G})}$ to construct feature vectors $\mathbf{x}^{(i)} \in \mathbb{R}^d$ for each node (i.e., for each data point) $i = 1, \ldots, n$. We can feed these feature vectors into Euclidean space-based συσταδοποίηση methods, such as αλγόριθμος k-μέσων or soft clustering via GMM. Roughly speaking, the feature vectors of nodes belonging to a well-connected subset (or συστάδα) of nodes in $\mathcal{G}$ are located nearby in the Euclidean space $\mathbb{R}^d$ (see Fig. 15).



$$\mathbf{L}^{(\mathcal{G})} = \begin{pmatrix} 2 & -1 & -1 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$$



$$\mathbf{V} = \left(\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}, \mathbf{v}^{(4)}\right)$$

$$\mathbf{v}^{(1)} = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \\ 0 \end{pmatrix}, \; \mathbf{v}^{(2)} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

Fig. 15. **Top.** Left: An undirected graph $\mathcal{G}$ with four nodes $i = 1, 2, 3, 4$, each representing a data point. Right: The Laplacian matrix $\mathbf{L}^{(\mathcal{G})} \in \mathbb{R}^{4\times 4}$ and its ανάλυση ιδιοτιμών. **Bottom.** Left: A scatterplot of data points using the feature vectors $\mathbf{x}^{(i)} = \left(v_i^{(1)}, v_i^{(2)}\right)^T$. Right: Two ιδιοδιάνυσμας $\mathbf{v}^{(1)}, \mathbf{v}^{(2)} \in \mathbb{R}^d$ corresponding to the ιδιοτιμή $\lambda = 0$ of the Laplacian matrix $\mathbf{L}^{(\mathcal{G})}$.

Βλέπε επίσης: συσταδοποίηση, συσταδοποίηση γράφου, data point, graph, ιδιοδιάνυσμα, Laplacian matrix, feature vector, Euclidean space, αλγόριθμος k-μέσων, soft clustering, GMM, συστάδα, ανάλυση ιδιοτιμών, scatterplot, ιδιοτιμή.

Φινλανδικό Μετεωρολογικό Ινστιτούτο The FMI (Finnish Meteorological Institute; FMI) is a government agency responsible for gathering and reporting weather data in Finland.
Βλέπε επίσης: data.

χαρακτηριστικό A feature of a data point is one of its properties that can be measured or computed easily without the need for human supervision. For example, if a data point is a digital image (e.g., stored as a `.jpeg` file), then we could use the red-green-blue intensities of its pixels as features. Domain-specific synonyms for the term feature are "covariate," "explanatory variable," "independent variable," "input (variable)," "predictor (variable)," or "regressor" [46], [47], [48].
Βλέπε επίσης: data point.

χώρος παραμέτρων The parameter space $\mathcal{W}$ of an ml model $\mathcal{H}$ is the set of all feasible choices for the παράμετροι μοντέλου (see Fig. 16). Many important ml methods use a model that is parametrized by vectors of the Euclidean space $\mathbb{R}^d$. Two widely used examples of parametrized models are γραμμικό μοντέλος and βαθύ δίκτυος. The parameter space is then often a subset $\mathcal{W} \subseteq \mathbb{R}^d$, e.g., all vectors $\mathbf{w} \in \mathbb{R}^d$ with a νόρμα smaller than one.

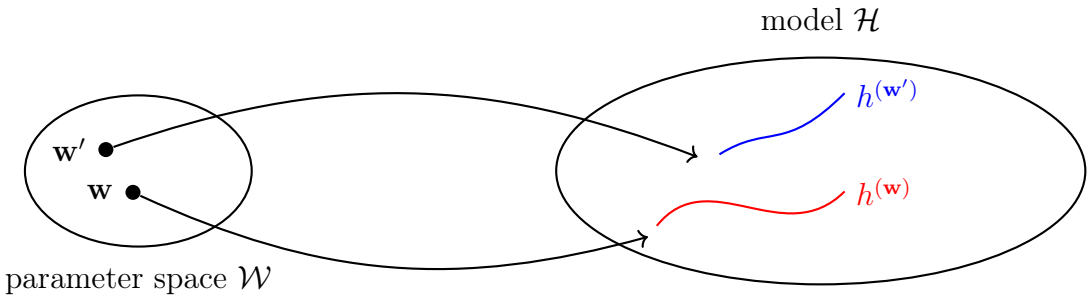


Fig. 16. The parameter space $\mathcal{W}$ of an ml model $\mathcal{H}$ consists of all feasible choices for the παράμετροι μοντέλου. Each choice $\mathbf{w}$ for the παράμετροι μοντέλου selects a υπόθεση map $h^{(\mathbf{w})} \in \mathcal{H}$.

Βλέπε επίσης: ml, model, παράμετροι μοντέλου, Euclidean space, γραμμικό μοντέλο, βαθύ δίκτυο, νόρμα, υπόθεση.

χώρος υποθέσεων Every practical ml method uses a υπόθεση space (or model) $\mathcal{H}$. The υπόθεση space of an ml method is a subset of all possible maps from the feature space to the label space. The design choice of the υπόθεση space should take into account available computational resources and στατιστικές διαστάσεις. If the computational infrastructure allows for efficient matrix operations, and there is an (approximately) linear relation between a set of features and a ετικέτα, a useful choice for the υπόθεση space might be the γραμμικό μοντέλο.

Βλέπε επίσης: ml, υπόθεση, model, feature space, label space, στατιστικές διαστάσεις, feature, ετικέτα, γραμμικό μοντέλο.

χώρος Hilbert A Hilbert space is a complete inner product space [78]. That is, it is a linear vector space equipped with an inner product

between pairs of vectors, and it satisfies the additional requirement of completeness, i.e., every Cauchy sequence of vectors converges to a limit within the space. A canonical example of a Hilbert space is the Euclidean space $\mathbb{R}^d$, for some dimension d, consisting of vectors $\mathbf{u} = \big(u_1, \ldots, u_d\big)^T$ and the standard inner product $\mathbf{u}^T\mathbf{v}$.

Βλέπε επίσης: Euclidean space.

0/1 απώλεια The 0/1 loss $L^{(0/1)}\left((\mathbf{x}, y), h\right)$ measures the quality of a ταξινομητής $h(\mathbf{x})$ that delivers a πρόβλεψη $\hat{y}$ (e.g., via thresholding (8)) for the ετικέτα y of a data point with features $\mathbf{x}$. It is equal to 0 if the πρόβλεψη is correct, i.e., $L^{(0/1)}\left((\mathbf{x}, y), h\right) = 0$ when $\hat{y} = y$. It is equal to 1 if the πρόβλεψη is wrong, i.e., $L^{(0/1)}\left((\mathbf{x}, y), h\right) = 1$ when $\hat{y} \neq y$.

Βλέπε επίσης: loss, ταξινομητής, πρόβλεψη, ετικέτα, data point, feature.

supremum (or least upper bound) The supremum of a set of real numbers is the smallest number that is greater than or equal to every element in the set. More formally, a real number a is the supremum of a set $\mathcal{A} \subseteq \mathbb{R}$ if: 1) a is an upper bound of $\mathcal{A}$; and 2) no number smaller than a is an upper bound of $\mathcal{A}$. Every non-empty set of real numbers that is bounded above has a supremum, even if it does not contain its supremum as an element [2, Sec. 1.4].

vertical federated learning (VFL) VFL uses τοπικό σύνολο δεδομένωνs that are constituted by the same data points but characterizing them with different features [79]. For example, different healthcare providers might all contain information about the same population of patients.

However, different healthcare providers collect different measurements (e.g., blood values, electrocardiography, lung X-ray) for the same patients.

Βλέπε επίσης: FL, τοπικό σύνολο δεδομένων, data point, feature.

local interpretable model-agnostic explanations (LIME) Consider a trained model (or learned υπόθεση) $\widehat{h} \in \mathcal{H}$, which maps the feature vector of a data point to the πρόβλεψη $\widehat{y} = \widehat{h}$. Local interpretable model-agnostic επεξήγησηs is a technique for explaining the behavior of $\widehat{h}$, locally around a data point with feature vector $\mathbf{x}^{(0)}$ [25]. The explanation is given in the form of a local approximation $g \in \mathcal{H}'$ of $\widehat{h}$ (see Fig. 17). This approximation can be obtained by an instance of εμπειρική ελαχιστοποίηση διακινδύνευσης with carefully designed σύνολο εκπαίδευσης. In particular, the σύνολο εκπαίδευσης consists of data points with feature vector $\mathbf{x}$ close to $\mathbf{x}^{(0)}$ and the (pseudo-)label $\widehat{h}(\mathbf{x})$. Note that we can use a different model $\mathcal{H}'$ for the approximation from the original model $\mathcal{H}$. For example, we can use a decision tree to approximate (locally) a βαθύ δίκτυο. Another widely-used choice for $\mathcal{H}'$ is the γραμμικό μοντέλο.

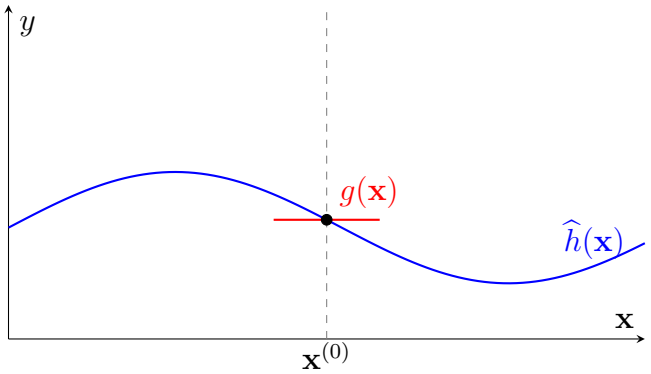


Fig. 17. To explain a trained model $\widehat{h} \in \mathcal{H}$, around a given feature vector $\mathbf{x}^{(0)}$, we can use a local approximation $g \in \mathcal{H}'$.

Βλέπε επίσης: model, υπόθεση, feature vector, data point, πρόβλεψη, επεξήγηση, εμπειρική ελαχιστοποίηση διακινδύνευσης, σύνολο εκπαίδευσης, decision tree, βαθύ δίκτυο, γραμμικό μοντέλο

Gaussian random variable (Gaussian RV) A standard Gaussian RV is a real-valued RV x with συνάρτηση πυκνότητας πιθανότητας [5], [22], [62]

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp^{-x^2/2}.$$

Given a standard Gaussian RV x, we can construct a general Gaussian RV x' with μέση τιμή μ and διακύμανση σ^2 via $x' := \sigma(x + \mu)$. The κατανομή πιθανότητας of a Gaussian RV is referred to as normal distribution, denoted $\mathcal{N}(\mu, \sigma)$.

A Gaussian random vector $\mathbf{x} \in \mathbb{R}^d$ with πίνακας συνδιακύμανσης $\mathbf{C}$ and μέση τιμή $\boldsymbol{\mu}$ can be constructed via $\mathbf{x} := \mathbf{A}\big(\mathbf{z} + \boldsymbol{\mu}\big)$. Here, $\mathbf{A}$ is any matrix that satisfies $\mathbf{A}\mathbf{A}^T = \mathbf{C}$ and $\mathbf{z} := \big(z_1, \ldots, z_d\big)^T$ is a vector whose

entries are i.i.d. standard Gaussian RVs $z_1, \ldots, z_d$. Gaussian random processes generalize Gaussian random vectors by applying linear transformations to infinite sequences of standard Gaussian RVs [80].

Gaussian RVs are widely used πιθανοτικό μοντέλος for the statistical analysis of ml methods. Their significance arises partly from the central limit theorem, which states that the average of an increasing number of independent RVs (not necessarily Gaussian themselves) converges to a Gaussian RV [81].

Βλέπε επίσης: RV, συνάρτηση πυκνότητας πιθανότητας, μέση τιμή, διακύμανση, κατανομή πιθανότητας, πίνακας συνδιακύμανσης, i.i.d., πιθανοτικό μοντέλο.

trustworthy artificial intelligence (trustworthy AI) Besides the υπολογιστικές διαστάσεις and στατιστικές διαστάσεις, a third main design aspect of ml methods is their trustworthiness [82]. The EU has put forward seven key requirements (KRs) for trustworthy τεχνητή νοημοσύνη (that typically build on ml methods) [83]:

1) KR1 - Human agency and oversight;
2) KR2 - Technical robustness and safety;
3) KR3 - Privacy and data governance;
4) KR4 - Transparency;
5) KR5 - Diversity, non-discrimination and fairness;
6) KR6 - Societal and environmental well-being;
7) KR7 - Accountability.

Βλέπε επίσης: υπολογιστικές διαστάσεις, στατιστικές διαστάσεις, ml, τεχνητή νοημοσύνη.

stability Stability is a desirable property of an ml method $\mathcal{A}$ that maps a σύνολο δεδομένων $\mathcal{D}$ (e.g., a σύνολο εκπαίδευσης) to an output $\mathcal{A}(\mathcal{D})$. The output $\mathcal{A}(\mathcal{D})$ can be the learned παράμετροι μοντέλου or the πρόβλεψη delivered by the trained model for a specific data point. Intuitively, $\mathcal{A}$ is stable if small changes in the input σύνολο δεδομένων $\mathcal{D}$ lead to small changes in the output $\mathcal{A}(\mathcal{D})$. Several formal notions of stability exist that enable bounds on the generalization error or διακινδύνευση of the method (see [7, Ch. 13]). To build intuition, consider the three σύνολο δεδομένωνs depicted in Fig. 18, each of which is equally likely under the same data-generating κατανομή πιθανότητας. Since the optimal παράμετροι μοντέλου are determined by this underlying κατανομή πιθανότητας, an accurate ml method $\mathcal{A}$ should return the same (or very similar) output $\mathcal{A}(\mathcal{D})$ for all three σύνολο δεδομένωνs. In other words, any useful $\mathcal{A}$ must be robust to variability in δείγμα realizations from the same κατανομή πιθανότητας, i.e., it must be stable.

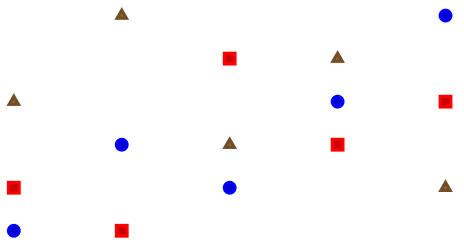


Fig. 18. Three σύνολο δεδομένωνs $\mathcal{D}^{(*)}$, $\mathcal{D}^{(\square)}$, and $\mathcal{D}^{(\triangle)}$, each sampled independently from the same data-generating κατανομή πιθανότητας. A stable ml method should return similar outputs when trained on any of these σύνολο δεδομένωνs.

Βλέπε επίσης: ml, σύνολο δεδομένων, σύνολο εκπαίδευσης, παράμετροι μοντέλου, πρόβλεψη, model, data point, generalization, διακινδύνευση, data, κατανομή πιθανότητας, δείγμα, realization.

multi-armed bandit (MAB) A MAB problem models a repeated decision-making scenario in which, at each time step k, a learner must choose one out of several possible actions, often referred to as arms, from a finite set $\mathcal{A}$. Each arm $a \in \mathcal{A}$ yields a stochastic ανταμοιβή $r^{(a)}$ drawn from an unknown κατανομή πιθανότητας with μέση τιμή $\mu^{(a)}$. The learner's goal is to maximize the cumulative ανταμοιβή over time by strategically balancing exploration (gathering information about uncertain arms) and exploitation (selecting arms known to perform well). This balance

is quantified by the notion of regret, which measures the performance gap between the learner's strategy and the optimal strategy that always selects the best arm. MAB problems form a foundational model in online learning, reinforcement learning, and sequential experimental design [8].

Βλέπε επίσης: ανταμοιβή, κατανομή πιθανότητας, μέση τιμή, regret, model.

federated learning network (FL network) An FL network is an undirected weighted graph whose nodes represent data generators that aim to train a local (or personalized) model. Each node in an FL network represents some συσκευή capable of collecting a τοπικό σύνολο δεδομένων and, in turn, train a local model. FL methods learn a local υπόθεση $h^{(i)}$, for each node $i \in \mathcal{V}$, such that it incurs small loss on the τοπικό σύνολο δεδομένωνs.

Βλέπε επίσης: FL, graph, data, model, συσκευή, τοπικό σύνολο δεδομένων, local model, υπόθεση, loss.

dual norm Every νόρμα $\|\cdot\|$ defined on an Euclidean space $\mathbb{R}^d$ has an associated dual νόρμα, denoted $\|\cdot\|_*$, defined as $\|\mathbf{y}\|_* := \sup_{\|\mathbf{x}\|\leq 1} \mathbf{y}^T\mathbf{x}$. The dual νόρμα measures the largest possible inner product between $\mathbf{y}$ and any vector in the unit ball of the original νόρμα. For further details, see [52, Sec. A.1.6].

Βλέπε επίσης: νόρμα, Euclidean space.

distributed algorithm A distributed αλγόριθμος is an αλγόριθμος designed

for a special type of computer: a collection of interconnected computing devices (or nodes). These devices communicate and coordinate their local computations by exchanging messages over a network [84], [85]. Unlike a classical αλγόριθμος, which is implemented on a single συσκευή, a distributed αλγόριθμος is executed concurrently on multiple συσκευήs with computational capabilities. Similar to a classical αλγόριθμος, a distributed αλγόριθμος can be modeled as a set of potential executions. However, each execution in the distributed setting involves both local computations and message-passing events. A generic execution might look as follows:

$$\begin{aligned}
&\text{Node 1: } \text{input}_1, s_1^{(1)}, s_2^{(1)}, \ldots, s_{T_1}^{(1)}, \text{output}_1; \\
&\text{Node 2: } \text{input}_2, s_1^{(2)}, s_2^{(2)}, \ldots, s_{T_2}^{(2)}, \text{output}_2; \\
&\quad \vdots \\
&\text{Node N: } \text{input}_N, s_1^{(N)}, s_2^{(N)}, \ldots, s_{T_N}^{(N)}, \text{output}_N.
\end{aligned}$$

Each συσκευή i starts from its own local input and performs a sequence of intermediate computations $s_k^{(i)}$ at discrete time instants $k = 1, \ldots, T_i$. These computations may depend on both: the previous local computations at the συσκευή and messages received from other συσκευήs. One important application of distributed αλγόριθμοςs is in FL where a network of συσκευήs collaboratively train a personal model for each συσκευή.

Βλέπε επίσης: αλγόριθμος, συσκευή, FL, model.

online algorithm An online αλγόριθμος processes input data incrementally, receiving data items sequentially and making decisions or producing

outputs (or decisions) immediately without having access to the entire input in advance [86], [87]. Unlike an offline αλγόριθμος, which has the entire input available from the start, an online αλγόριθμος must handle αβεβαιότητα about future inputs and cannot revise past decisions. Similar to an offline αλγόριθμος, an online αλγόριθμος can be modeled formally as a collection of possible executions. However, the execution sequence for an online αλγόριθμος has a distinct structure:

$$\text{init}, s_1, \text{out}_1, \text{in}_2, s_2, \text{out}_2, \ldots, \text{in}_T, s_T, \text{out}_T.$$

Each execution begins from an initial state (init) and proceeds through alternating computational steps, outputs (or decisions), and inputs. Specifically, at step k, the αλγόριθμος performs a computational step s_k, generates an output out_k, and then subsequently receives the next input in_{k+1}. A notable example of an online αλγόριθμος in ml is online gradient descent (online GD), which incrementally updates παράμετροι μοντέλου as new data points arrive.

Βλέπε επίσης: αλγόριθμος, data, αβεβαιότητα, ml, online GD, παράμετροι μοντέλου, data point.

spectrogram A spectrogram represents the time-frequency distribution of the energy of a time signal $x(t)$. Intuitively, it quantifies the amount of signal energy present within a specific time segment $[t_1, t_2] \subseteq \mathbb{R}$ and frequency interval $[f_1, f_2] \subseteq \mathbb{R}$. Formally, the spectrogram of a signal is defined as the squared magnitude of its short-time Fourier transform (STFT) [88]. Fig. 19 depicts a time signal along with its spectrogram.

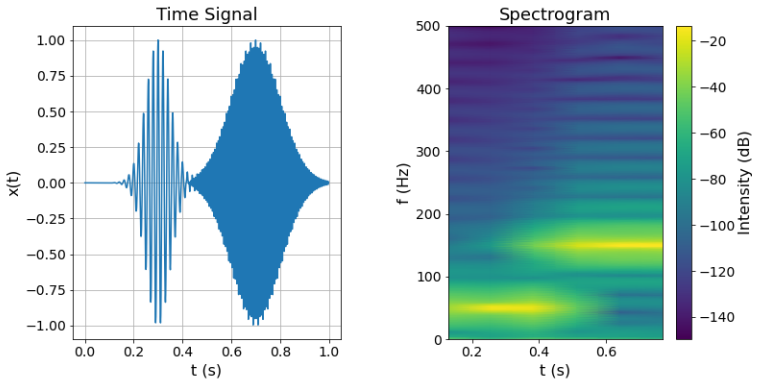


Fig. 19. Left: A time signal consisting of two modulated Gaussian pulses. Right: An intensity plot of the spectrogram.

The intensity plot of its spectrogram can serve as an image of a signal. A simple recipe for audio signal ταξινόμηση is to feed this signal image into βαθύ δίκτυοs originally developed for image ταξινόμηση and object detection [89]. It is worth noting that, beyond the spectrogram, several alternative representations exist for the time-frequency distribution of signal energy [90], [91].

Βλέπε επίσης: ταξινόμηση, βαθύ δίκτυο.

generalized total variation minimization (GTVMin) GTV minimization is an instance of regularized empirical risk minimization (RERM) using the GTV of local παράμετροι μοντέλου as a regularizer [92].

Βλέπε επίσης: GTV, RERM, παράμετροι μοντέλου, regularizer.

model inversion TBD.

bagging (or bootstrap aggregation) Bagging (or bootstrap aggregation) is a generic technique to improve (the robustness of) a given ml method. The idea is to use the εκκίνηση to generate perturbed copies of a given σύνολο δεδομένων and then to learn a separate υπόθεση for each copy. We then predict the ετικέτα of a data point by combining or aggregating the individual πρόβλεψηs of each separate υπόθεση. For υπόθεση maps delivering numeric ετικέτα values, this aggregation could be implemented by computing the average of individual πρόβλεψηs.

Βλέπε επίσης: ml, εκκίνηση, σύνολο δεδομένων, υπόθεση, ετικέτα, data point, πρόβλεψη.

online gradient descent (online GD) Consider an ml method that learns παράμετροι μοντέλου $\mathbf{w}$ from some χώρος παραμέτρων $\mathcal{W} \subseteq \mathbb{R}^d$. The learning process uses data points $\mathbf{z}^{(t)}$ that arrive at consecutive time-instants $t = 1, 2, \ldots$. Let us interpret the data points $\mathbf{z}^{(t)}$ as i.i.d. copies of an RV $\mathbf{z}$. The διακινδύνευση $\mathbb{E}\{L(\mathbf{z}, \mathbf{w})\}$ of a υπόθεση $h^{(\mathbf{w})}$ can then (under mild conditions) be obtained as the limit $\lim_{T\to\infty}(1/T)\sum_{t=1}^{T} L\left(\mathbf{z}^{(t)}, \mathbf{w}\right)$. We might use this limit as the objective function for learning the παράμετροι μοντέλου $\mathbf{w}$. Unfortunately, this limit can only be evaluated if we wait infinitely long in order to collect all data points. Some ml applications require methods that learn online: as soon as a new data point $\mathbf{z}^{(t)}$ arrives at time t, we update the current παράμετροι μοντέλου $\mathbf{w}^{(t)}$. Note that the new data point $\mathbf{z}^{(t)}$ contributes the component $L\left(\mathbf{z}^{(t)}, \mathbf{w}\right)$ to the διακινδύνευση. As its name suggests, online κάθοδος

κλίσης updates $\mathbf{w}^{(t)}$ via a (projected) βήμα κλίσης

$$\mathbf{w}^{(t+1)} := P_{\mathcal{W}}\big(\mathbf{w}^{(t)} - \eta_t \nabla_{\mathbf{w}} L\left(\mathbf{z}^{(t)}, \mathbf{w}\right)\big). \tag{9}$$

Note that (9) is a βήμα κλίσης for the current component $L\left(\mathbf{z}^{(t)}, \cdot\right)$ of the διακινδύνευση. The update (9) ignores all the previous components $L\left(\mathbf{z}^{(t')}, \cdot\right)$, for $t' < t$. It might therefore happen that, compared to $\mathbf{w}^{(t)}$, the updated παράμετροι μοντέλου $\mathbf{w}^{(t+1)}$ increase the retrospective average loss $\sum_{t'=1}^{t-1} L\left(\mathbf{z}^{(t')}, \cdot\right)$. However, for a suitably chosen learning rate η_t, online κάθοδος κλίσης can be shown to be optimal in practically relevant settings. By optimal, we mean that the παράμετροι μοντέλου $\mathbf{w}^{(T+1)}$ delivered by online κάθοδος κλίσης after observing T data points $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(T)}$ are at least as good as those delivered by any other learning method [87], [93].

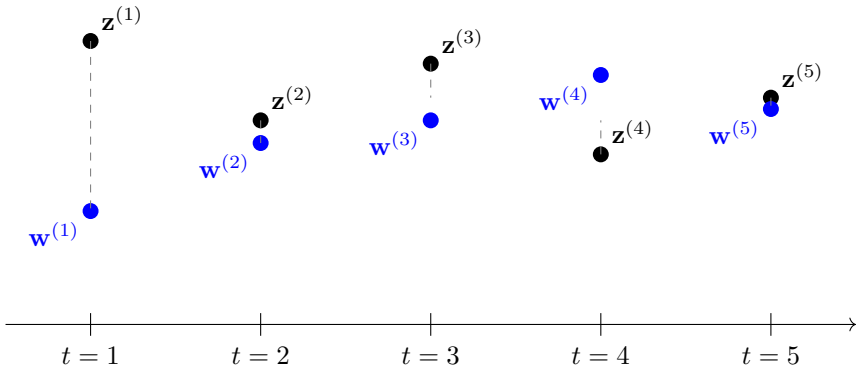


Fig. 20. An instance of online κάθοδος κλίσης that updates the παράμετροι μοντέλου $\mathbf{w}^{(t)}$ using the data point $\mathbf{z}^{(t)} = x^{(t)}$ arriving at time t. This instance uses the τετραγωνική απώλεια σφάλματος $L\left(\mathbf{z}^{(t)}, w\right) = (x^{(t)} - w)^2$.

Βλέπε επίσης: ml, παράμετροι μοντέλου, χώρος παραμέτρων, data point,

i.i.d., RV, διακινδύνευση, υπόθεση, objective function, κάθοδος κλίσης, βήμα κλίσης, loss, learning rate, τετραγωνική απώλεια σφάλματος.

gradient-based methods Gradient-based methods are iterative techniques for finding the ολικό ελάχιστο (or maximum) of a παραγωγίσιμη objective function of the παράμετροι μοντέλου. These methods construct a sequence of approximations to an optimal choice for παράμετροι μοντέλου that results in a ολικό ελάχιστο (or maximum) value of the objective function. As their name indicates, gradient-based methods use the gradients of the objective function evaluated during previous iterations to construct new, (hopefully) improved παράμετροι μοντέλου. One important example of a gradient-based method is κάθοδος κλίσης.
Βλέπε επίσης: gradient, ολικό ελάχιστο, maximum, παραγωγίσιμη, objective function, παράμετροι μοντέλου, κάθοδος κλίσης.

independent and identically distributed assumption (i.i.d. assumption) The i.i.d. assumption interprets data points of a σύνολο δεδομένων as the realizations of i.i.d. RVs.
Βλέπε επίσης: i.i.d., data point, σύνολο δεδομένων, realization, RV.

probabilistic principal component analysis (PPCA) Probabilistic principal component analysis extends basic principal component analysis by using a πιθανοτικό μοντέλο for data points. The πιθανοτικό μοντέλο of probabilistic principal component analysis reduces the task of dimensionality reduction to an estimation problem that can be solved using EM methods.

Βλέπε επίσης: principal component analysis, πιθανοτικό μοντέλο, data point, EM.

Gaussian mixture model (GMM) A GMM is a particular type of πιθανοτικό μοντέλο for a numeric vector $\mathbf{x}$ (e.g., the features of a data point). Within a GMM, the vector $\mathbf{x}$ is drawn from a randomly selected πολυμεταβλητή κανονική κατανομή $p^{(c)} = \mathcal{N}\left(\boldsymbol{\mu}^{(c)}, \mathbf{C}^{(c)}\right)$ with $c = I$. The index $I \in \{1, \ldots, k\}$ is an RV with probabilities $p(I = c) = p_c$. Note that a GMM is parametrized by the probability p_c, the μέση τιμή vector $\boldsymbol{\mu}^{(c)}$, and the πίνακας συνδιακύμανσης $\mathbf{\Sigma}^{(\mathbf{c})}$ for each $c = 1, \ldots, k$. GMMs are widely used for συσταδοποίηση, density estimation, and as a generative model.

Βλέπε επίσης: πιθανοτικό μοντέλο, feature, data point, πολυμεταβλητή κανονική κατανομή, RV, μέση τιμή, πίνακας συνδιακύμανσης, συσταδοποίηση, model.

expectation-maximization (EM) Consider a πιθανοτικό μοντέλο $p(\mathbf{z}; \mathbf{w})$ for the data points $\mathcal{D}$ generated in some ml application. The μέγιστη πιθανοφάνεια estimator for the παράμετροι μοντέλου $\mathbf{w}$ is obtained by maximizing $p(\mathcal{D}; \mathbf{w})$. However, the resulting optimization problem might be computationally challenging. Expectation-maximization approximates the μέγιστη πιθανοφάνεια estimator by introducing a latent RV $\mathbf{z}$ such that maximizing $p(\mathcal{D}, \mathbf{z}; \mathbf{w})$ would be easier [77], [58], [94]. Since we do not observe $\mathbf{z}$, we need to estimate it from the observed σύνολο δεδομένων $\mathcal{D}$ using a conditional expectation. The resulting estimate $\widehat{\mathbf{z}}$ is then used to compute a new estimate $\widehat{\mathbf{w}}$ by solving $\max_{\mathbf{w}} p(\mathcal{D}, \widehat{\mathbf{z}}; \mathbf{w})$.

The crux is that the conditional expectation $\widehat{\mathbf{z}}$ depends on the παράμετροι μοντέλου $\widehat{\mathbf{w}}$, which we have updated based on $\widehat{\mathbf{z}}$. Thus, we have to re-calculate $\widehat{\mathbf{z}}$, which, in turn, results in a new choice $\widehat{\mathbf{w}}$ for the παράμετροι μοντέλου. In practice, we repeat the computation of the conditional expectation (i.e., the E-step) and the update of the παράμετροι μοντέλου (i.e., the M-step) until some κριτήριο τερματισμού is met.

Βλέπε επίσης: πιθανοτικό μοντέλο, data point, ml, μέγιστη πιθανοφάνεια, παράμετροι μοντέλου, expectation, RV, σύνολο δεδομένων, κριτήριο τερματισμού.

high-dimensional regime The high-dimensional regime of εμπειρική ελαχιστοποίηση διακινδύνευσης is characterized by the effective dimension of the model being larger than the μέγεθος δείγματος, i.e., the number of (labeled) data points in the σύνολο εκπαίδευσης. For example, γραμμική παλινδρόμηση methods operate in the high-dimensional regime whenever the number d of features used to characterize data points exceeds the number of data points in the σύνολο εκπαίδευσης. Another example of ml methods that operate in the high-dimensional regime is large τεχνητό νευρωνικό δίκτυος, which have far more tunable βάρη (and bias terms) than the total number of data points in the σύνολο εκπαίδευσης. High-dimensional statistics is a recent main thread of probability theory that studies the behavior of ml methods in the high-dimensional regime [95], [96].

Βλέπε επίσης: εμπειρική ελαχιστοποίηση διακινδύνευσης, effective dimen-

sion, model, μέγεθος δείγματος, data point, σύνολο εκπαίδευσης, γραμμική παλινδρόμηση, feature, ml, τεχνητό νευρωνικό δίκτυο, βάρη, probability.

federated learning (FL) FL is an umbrella term for ml methods that train models in a collaborative fashion using decentralized data and computation.

Βλέπε επίσης: ml, model, data.

clustered federated learning (CFL) CFL trains local models for the συσκευήs in a FL application by using a παραδοχή συσταδοποίησης, i.e., the συσκευήs of an FL network form συστάδαs. Two συσκευήs in the same συστάδα generate τοπικό σύνολο δεδομένωνs with similar statistical properties. CFL pools the τοπικό σύνολο δεδομένωνs of συσκευήs in the same συστάδα to obtain a σύνολο εκπαίδευσης for a συστάδα-specific model. Generalized total variation minimization (GTVMin) clusters συσκευήs implicitly by enforcing approximate similarity of παράμετροι μοντέλου across well-connected nodes of the FL network.

Βλέπε επίσης: local model, συσκευή, FL, παραδοχή συσταδοποίησης, FL network, συστάδα, τοπικό σύνολο δεδομένων, σύνολο εκπαίδευσης, model, GTVMin, παράμετροι μοντέλου.

explainable machine learning (explainable ML) Explainable ml methods aim at complementing each πρόβλεψη with an επεξήγηση of how the πρόβλεψη has been obtained. The construction of an explicit επεξήγηση might not be necessary if the ml method uses a sufficiently simple (or

interpretable) model [24].

Βλέπε επίσης: ml, πρόβλεψη, επεξήγηση, model.

algebraic connectivity The algebraic connectivity of an undirected graph is the second-smallest ιδιοτιμή λ_2 of its Laplacian matrix. A graph is connected if and only if $\lambda_2 > 0$.

Βλέπε επίσης: graph, ιδιοτιμή, Laplacian matrix.

Courant–Fischer–Weyl min-max characterization Consider a psd matrix $\mathbf{Q} \in \mathbb{R}^{d \times d}$ with ανάλυση ιδιοτιμών (or spectral decomposition),

$$\mathbf{Q} = \sum_{j=1}^{d} \lambda_j \mathbf{u}^{(j)} \big(\mathbf{u}^{(j)}\big)^T.$$

Here, we use the ordered (in increasing fashion) ιδιοτιμής

$$\lambda_1 \leq \ldots \leq \lambda_n.$$

The Courant–Fischer–Weyl min-max characterization [3, Th. 8.1.2] represents the ιδιοτιμής of $\mathbf{Q}$ as the solutions to certain optimization problems.

Βλέπε επίσης: psd, ανάλυση ιδιοτιμών, ιδιοτιμή.

networked exponential families (nExpFam) A collection of exponential families, each of them assigned to a node of an FL network. The παράμετροι μοντέλου are coupled via the network structure by requiring them to have a small GTV [97].

Βλέπε επίσης: FL network, παράμετροι μοντέλου, GTV.

regularized loss minimization (RLM) See RERM.

data poisoning Data poisoning refers to the intentional manipulation (or fabrication) of data points to steer the training of an ml model [98], [99]. The protection against data poisoning is particularly important in distributed ml applications where σύνολο δεδομένωνs are decentralized. Βλέπε επίσης: data, data point, ml, model, σύνολο δεδομένων.

geometric median (GM) The GM of a set of input vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$ in $\mathbb{R}^d$ is a point $\mathbf{z} \in \mathbb{R}^d$ that minimizes the sum of distances to the vectors [52],

$$\mathbf{z} \in \operatorname*{argmin}_{\mathbf{y} \in \mathbb{R}^d} \sum_{r=1}^{m} \left\| \mathbf{y} - \mathbf{x}^{(r)} \right\|_2 . \tag{10}$$

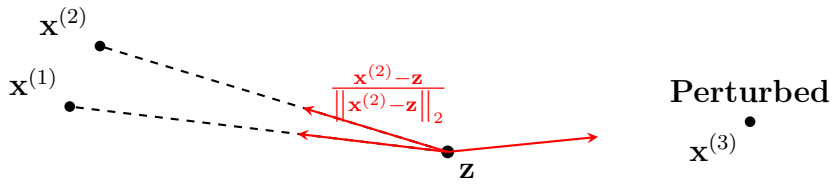


Fig. 21. Consider a solution $\mathbf{z}$ of (10) that does not coincide with any of the input vectors. The optimality condition for (10) requires that the unit vectors from $\mathbf{z}$ to the input vectors sum to zero.

Figure 21 illustrates a fundamental property of the GM: If $\mathbf{z}$ does not coincide with any of the input vectors, then the unit vectors pointing from $\mathbf{z}$ to each $\mathbf{x}^{(r)}$ must sum to zero - this is the zero-subgradient (optimality) condition of (10). It turns out that the solution to (10)

cannot be arbitrarily pulled away from trustworthy input vectors as long as they are the majority [100, Th. 2.2].

FedRelax A distributed FL αλγόριθμος.

See also: FL, αλγόριθμος.

FedAvg A FL αλγόριθμος using a server-client setting.

See also: FL, αλγόριθμος.

FedGD A distributed FL αλγόριθμος that can be implemented as message passing across a FL network.

See also: FL, αλγόριθμος, βήμα κλίσης, gradient-based methods.

FedSGD A distributed FL αλγόριθμος that can be implemented as message passing across a FL network.

See also: FL, αλγόριθμος, βήμα κλίσης, gradient-based methods, στοχαστική κάθοδος κλίσης.

regression Regression problems revolve around the prediction of a numeric ετικέτα solely from the features of a data point [6, Ch. 2].

Βλέπε επίσης: ετικέτα, feature, data point.

expert ml aims to learn a υπόθεση h that accurately predicts the ετικέτα of a data point based on its features. We measure the πρόβλεψη error using some συνάρτηση απώλειας. Ideally, we want to find a υπόθεση that incurs minimal loss on any data point. We can make this informal goal precise via the i.i.d. assumption and by using the διακινδύνευση

Bayes as the βάση αναφοράς for the (average) loss of a υπόθεση. An alternative approach to obtaining a βάση αναφοράς is to use the υπόθεση h' learned by an existing ml method. We refer to this υπόθεση h' as an expert [86]. Regret minimization methods learn a υπόθεση that incurs a loss comparable to the best expert [86], [87].

Βλέπε επίσης: ml, υπόθεση, ετικέτα, data point, feature, πρόβλεψη, συνάρτηση απώλειας, loss, i.i.d. assumption, διακινδύνευση Bayes, βάση αναφοράς, regret.

networked federated learning (NFL) Networked FL refers to methods that learn personalized models in a distributed fashion. These methods learn from τοπικό σύνολο δεδομένωνs that are related by an intrinsic network structure.

Βλέπε επίσης: FL, model, τοπικό σύνολο δεδομένων.

regret The regret of a υπόθεση h relative to another υπόθεση h', which serves as a βάση αναφοράς, is the difference between the loss incurred by h and the loss incurred by h' [86]. The βάση αναφοράς υπόθεση h' is also referred to as an expert.

Βλέπε επίσης: υπόθεση, βάση αναφοράς, loss, expert.

strongly convex A continuously παραγωγίσιμη real-valued function $f(\mathbf{x})$ is strongly convex with coefficient σ if $f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + (\sigma/2)\left\|\mathbf{y}-\mathbf{x}\right\|_2^2$ [53], [55, Sec. B.1.1].

Βλέπε επίσης: παραγωγίσιμη, convex.

subgradient For a real-valued function $f : \mathbb{R}^d \to \mathbb{R} : \mathbf{w} \mapsto f(\mathbf{w})$, a vector $\mathbf{a}$ such that $f(\mathbf{w}) \geq f(\mathbf{w}') + \left(\mathbf{w} - \mathbf{w}'\right)^T \mathbf{a}$ is referred to as a subgradient of f at $\mathbf{w}'$ [101], [102].

federated averaging (FedAvg) FedAvg refers to an iterative FL αλγόριθμος that alternates between separately training local models and combining the updated local παράμετροι μοντέλου. The training of local models is implemented via several στοχαστική κάθοδος κλίσης steps [103].

Βλέπε επίσης: FL, αλγόριθμος, local model, παράμετροι μοντέλου, στοχαστική κάθοδος κλίσης.

FedProx FedProx refers to an iterative FL αλγόριθμος that alternates between separately training local models and combining the updated local παράμετροι μοντέλου. In contrast to FedAvg, which uses στοχαστική κάθοδος κλίσης to train local models, FedProx uses a εγγύς τελεστής for the training [104].

Βλέπε επίσης: FL, αλγόριθμος, local model, παράμετροι μοντέλου, FedAvg, στοχαστική κάθοδος κλίσης, εγγύς τελεστής.

rectified linear unit (ReLU) The ReLU is a popular choice for the συνάρτηση ενεργοποίησης of a neuron within an τεχνητό νευρωνικό δίκτυο. It is defined as $\sigma(z) = \max\{0, z\}$, with z being the weighted input of the artificial neuron.

Βλέπε επίσης: συνάρτηση ενεργοποίησης, τεχνητό νευρωνικό δίκτυο.

Vapnik–Chervonenkis dimension (VC dimension) The VC dimension of an infinite χώρος υποθέσεων is a widely-used measure for its size. We

refer to the literature (see [7]) for a precise definition of VC dimension as well as a discussion of its basic properties and use in ml.

Βλέπε επίσης: χώρος υποθέσεων, ml.

effective dimension The effective dimension $d_{\text{eff}}(\mathcal{H})$ of an infinite χώρος υποθέσεων $\mathcal{H}$ is a measure of its size. Loosely speaking, the effective dimension is equal to the effective number of independent tunable παράμετροι μοντέλου. These παράμετροι might be the coefficients used in a linear map or the βάρη and bias terms of an τεχνητό νευρωνικό δίκτυο.

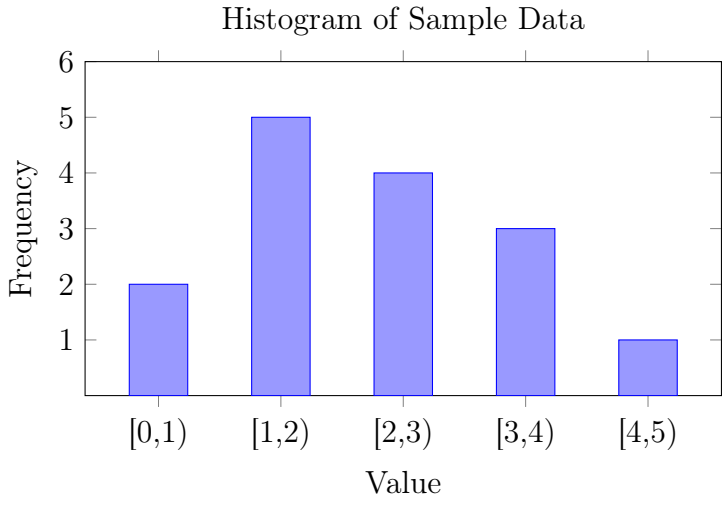
Βλέπε επίσης: χώρος υποθέσεων, παράμετροι μοντέλου, παράμετροι, βάρη, τεχνητό νευρωνικό δίκτυο.

label space Consider an ml application that involves data points characterized by features and ετικέτας. The ετικέτα space is constituted by all potential values that the ετικέτα of a data point can take on. Regression methods, aiming at predicting numeric ετικέτας, often use the ετικέτα space $\mathcal{Y} = \mathbb{R}$. Binary ταξινόμηση methods use a ετικέτα space that consists of two different elements, e.g., $\mathcal{Y} = \{-1, 1\}$, $\mathcal{Y} = \{0, 1\}$, or $\mathcal{Y} = \{$"cat image","no cat image"$\}$.

Βλέπε επίσης: ml, data point, feature, ετικέτα, regression, ταξινόμηση.

histogram Consider a σύνολο δεδομένων $\mathcal{D}$ that consists of m data points $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}$, each of them belonging to some cell $[-U, U] \times \ldots \times [-U, U] \subseteq \mathbb{R}^d$ with side length U. We partition this cell evenly into smaller elementary cells with side length Δ. The histogram of $\mathcal{D}$ assigns each elementary cell to the corresponding fraction of data points in $\mathcal{D}$

that fall into this elementary cell.



Βλέπε επίσης: σύνολο δεδομένων, data point.

feature space The feature space of a given ml application or method is constituted by all potential values that the feature vector of a data point can take on. A widely used choice for the feature space is the Euclidean space $\mathbb{R}^d$, with the dimension d being the number of individual features of a data point.

Βλέπε επίσης: feature, ml, feature vector, data point, feature, Euclidean space.

missing data Consider a σύνολο δεδομένων constituted by data points collected via some physical συσκευή. Due to imperfections and failures, some of the feature or ετικέτα values of data points might be corrupted or simply missing. Data imputation aims at estimating these missing values [105]. We can interpret data imputation as an ml problem where the ετικέτα of a data point is the value of the corrupted feature.

Βλέπε επίσης: σύνολο δεδομένων, data point, συσκευή, feature, ετικέτα, data, ml.

positive semi-definite (psd) A (real-valued) symmetric matrix $\mathbf{Q} = \mathbf{Q}^T \in \mathbb{R}^{d \times d}$ is referred to as psd if $\mathbf{x}^T \mathbf{Q} \mathbf{x} \geq 0$ for every vector $\mathbf{x} \in \mathbb{R}^d$. The property of being psd can be extended from matrices to (real-valued) symmetric kernel maps $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (with $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x}', \mathbf{x})$) as follows: For any finite set of feature vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$, the resulting matrix $\mathbf{Q} \in \mathbb{R}^{m \times m}$ with entries $Q_{r,r'} = K\big(\mathbf{x}^{(r)}, \mathbf{x}^{(r')}\big)$ is psd [106].
Βλέπε επίσης: kernel, feature vector.

feature vector Feature vector refers to a vector $\mathbf{x} = \big(x_1, \ldots, x_d\big)^T$ whose entries are individual features $x_1, \ldots, x_d$. Many ml methods use feature vectors that belong to some finite-dimensional Euclidean space $\mathbb{R}^d$. For some ml methods, however, it can be more convenient to work with feature vectors that belong to an infinite-dimensional vector space (e.g., see kernel method).
Βλέπε επίσης: feature, ml, Euclidean space, kernel method.

predictor A predictor is a real-valued υπόθεση map. Given a data point with features $\mathbf{x}$, the value $h(\mathbf{x}) \in \mathbb{R}$ is used as a πρόβλεψη for the true numeric ετικέτα $y \in \mathbb{R}$ of the data point.
Βλέπε επίσης: υπόθεση, data point, feature, πρόβλεψη, ετικέτα.

labeled datapoint A data point whose ετικέτα is known or has been determined by some means which might require human labor.

Βλέπε επίσης: data point, ετικέτα.

random variable (RV) An RV is a function that maps from a probability space $\mathcal{P}$ to a value space [22], [49]. The probability space consists of elementary events and is equipped with a probability measure that assigns probabilities to subsets of $\mathcal{P}$. Different types of RVs include

- binary RVs, which map elementary events to a set of two distinct values, such as $\{-1, 1\}$ or $\{\text{cat}, \text{no cat}\}$;
- real-valued RVs, which take values in the real numbers $\mathbb{R}$;
- vector-valued RVs, which map elementary events to the Euclidean space $\mathbb{R}^d$.

Probability theory uses the concept of measurable spaces to rigorously define and study the properties of (large) collections of RVs [49].
Βλέπε επίσης: probability space, probability, Euclidean space.

probability space A probability space is a mathematical model of a physical process (a random experiment) with an uncertain outcome. Formally, a probability space $\mathcal{P}$ is a triplet $(\Omega, \mathcal{F}, P)$ where

- Ω is a sample space containing all possible elementary outcomes of a random experiment;
- $\mathcal{F}$ is a sigma-algebra, a collection of subsets of Ω (called events) that satisfies certain closure properties under set operations;
- P is a probability measure, a function that assigns a probability $P(\mathcal{A}) \in [0, 1]$ to each event $\mathcal{A} \in \mathcal{F}$. The function must satisfy

$P(\Omega) = 1$ and $P\left(\bigcup_{i=1}^{\infty} \mathcal{A}_i\right) = \sum_{i=1}^{\infty} P(\mathcal{A}_i)$ for any countable sequence of pairwise disjoint events $\mathcal{A}_1, \mathcal{A}_2, \ldots$ in $\mathcal{F}$.

Probability spaces provide the foundation for defining RVs and to reason about αβεβαιότητα in ml applications [22], [49], [81].

Βλέπε επίσης: probability, model, RV, αβεβαιότητα, ml.

realization Consider an RV x which maps each element (i.e., outcome or elementary event) $\omega \in \mathcal{P}$ of a probability space $\mathcal{P}$ to an element a of a measurable space $\mathcal{N}$ [2], [66], [49]. A realization of x is any element $a' \in \mathcal{N}$ such that there is an element $\omega' \in \mathcal{P}$ with $x(\omega') = a'$.

Βλέπε επίσης: RV, probability space.

networked model A networked model over an FL network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ assigns a local model (i.e., a χώρος υποθέσεων) to each node $i \in \mathcal{V}$ of the FL network $\mathcal{G}$.

Βλέπε επίσης: model, FL network, local model, χώρος υποθέσεων.

networked data Networked data consists of τοπικό σύνολο δεδομένωνs that are related by some notion of pairwise similarity. We can represent networked data using a graph whose nodes carry τοπικό σύνολο δεδομένωνs and edges encode pairwise similarities. One example of networked data arises in FL applications where τοπικό σύνολο δεδομένωνs are generated by spatially distributed συσκευήs.

Βλέπε επίσης: data, τοπικό σύνολο δεδομένων, graph, FL, συσκευή.

training error The average loss of a υπόθεση when predicting the ετικέτας of the data points in a σύνολο εκπαίδευσης. We sometimes refer by training error also to minimal average loss which is achieved by a solution of εμπειρική ελαχιστοποίηση διακινδύνευσης.

Βλέπε επίσης: loss, υπόθεση, ετικέτα, data point, σύνολο εκπαίδευσης, εμπειρική ελαχιστοποίηση διακινδύνευσης.

quadratic function A function $f : \mathbb{R}^d \to \mathbb{R}$ of the form

$$f(\mathbf{w}) = \mathbf{w}^T \mathbf{Q} \mathbf{w} + \mathbf{q}^T \mathbf{w} + a,$$

with some matrix $\mathbf{Q} \in \mathbb{R}^{d \times d}$, vector $\mathbf{q} \in \mathbb{R}^d$, and scalar $a \in \mathbb{R}$.

test set A set of data points that have been used neither to train a model (e.g., via εμπειρική ελαχιστοποίηση διακινδύνευσης) nor in a σύνολο επικύρωσης to choose between different models.

Βλέπε επίσης: data point, model, εμπειρική ελαχιστοποίηση διακινδύνευσης, σύνολο επικύρωσης.

model selection In ml, model selection refers to the process of choosing between different candidate models. In its most basic form, model selection amounts to: 1) training each candidate model; 2) computing the σφάλμα επικύρωσης for each trained model; and 3) choosing the model with the smallest σφάλμα επικύρωσης [6, Ch. 6].

Βλέπε επίσης: glsml, model, σφάλμα επικύρωσης.

linear classifier Consider data points characterized by numeric features $\mathbf{x} \in \mathbb{R}^d$ and a ετικέτα $y \in \mathcal{Y}$ from some finite label space $\mathcal{Y}$. A linear ταξινομητής is characterized by having περιοχή αποφάσεωνs that are separated by hyperplanes in $\mathbb{R}^d$ [6, Ch. 2].

Βλέπε επίσης: data point, feature, ετικέτα, label space, ταξινομητής, περιοχή αποφάσεων.

multi-label classification Multi-ετικέτα ταξινόμηση problems and methods use data points that are characterized by several ετικέτας. As an example, consider a data point representing a picture with two ετικέτας. One ετικέτα indicates the presence of a human in this picture and another ετικέτα indicates the presence of a car.

Βλέπε επίσης: ετικέτα, ταξινόμηση, data point.

semi-supervised learning (SSL) SSL methods use unlabeled data points to support the learning of a υπόθεση from labeled datapoints [65]. This approach is particularly useful for ml applications that offer a large amount of unlabeled data points, but only a limited number of labeled datapoints.

Βλέπε επίσης: data point, υπόθεση, labeled datapoint, ml.

objective function An objective function is a map that assigns each value of an optimization variable, such as the παράμετροι μοντέλου $\mathbf{w}$ of a υπόθεση $h^{(\mathbf{w})}$, to an objective value $f(\mathbf{w})$. The objective value $f(\mathbf{w})$ could be the διακινδύνευση or the empirical risk of a υπόθεση $h^{(\mathbf{w})}$.

Βλέπε επίσης: παράμετροι μοντέλου, υπόθεση, διακινδύνευση, empirical risk.

regularizer A regularizer assigns each υπόθεση h from a χώρος υποθέσεων $\mathcal{H}$ a quantitative measure $\mathcal{R}\big\{h\big\}$ for how much its πρόβλεψη error on a σύνολο εκπαίδευσης might differ from its πρόβλεψη errors on data points outside the σύνολο εκπαίδευσης. Ridge regression uses the regularizer $\mathcal{R}\big\{h\big\} := \|\mathbf{w}\|_2^2$ for linear υπόθεση maps $h^{(\mathbf{w})}(\mathbf{x}) := \mathbf{w}^T\mathbf{x}$ [6, Ch. 3]. Lasso uses the regularizer $\mathcal{R}\big\{h\big\} := \|\mathbf{w}\|_1$ for linear υπόθεση maps $h^{(\mathbf{w})}(\mathbf{x}) := \mathbf{w}^T\mathbf{x}$ [6, Ch. 3].

Βλέπε επίσης: υπόθεση, χώρος υποθέσεων, πρόβλεψη, σύνολο εκπαίδευσης, data point, ridge regression, Lasso.

regularization A key challenge of modern ml applications is that they often use large models, which have an effective dimension in the order of billions. Training a high-dimensional model using basic εμπειρική ελαχιστοποίηση διακινδύνευσης-based methods is prone to υπερπροσαρμογή: the learned υπόθεση performs well on the σύνολο εκπαίδευσης but poorly outside the σύνολο εκπαίδευσης. Regularization refers to modifications of a given instance of εμπειρική ελαχιστοποίηση διακινδύνευσης in order to avoid υπερπροσαρμογή, i.e., to ensure that the learned υπόθεση performs not much worse outside the σύνολο εκπαίδευσης. There are three routes for implementing regularization:

1) Model pruning: We prune the original model $\mathcal{H}$ to obtain a smaller model $\mathcal{H}'$. For a parametric model, the pruning can be implemented

via constraints on the παράμετροι μοντέλου (such as $w_1 \in [0.4, 0.6]$ for the weight of feature x_1 in γραμμική παλινδρόμηση).

2) Loss penalization: We modify the objective function of εμπειρική ελαχιστοποίηση διακινδύνευσης by adding a penalty term to the training error. The penalty term estimates how much larger the expected loss (or διακινδύνευση) is compared to the average loss on the σύνολο εκπαίδευσης.

3) Data augmentation: We can enlarge the σύνολο εκπαίδευσης $\mathcal{D}$ by adding perturbed copies of the original data points in $\mathcal{D}$. One example for such a perturbation is to add the realization of an RV to the feature vector of a data point.

Fig. 22 illustrates the above three routes to regularization. These routes are closely related and sometimes fully equivalent: data augmentation using Gaussian RVs to perturb the feature vectors in the σύνολο εκπαίδευσης of γραμμική παλινδρόμηση has the same effect as adding the penalty $\lambda \|\mathbf{w}\|_2^2$ to the training error (which is nothing but ridge regression). The decision on which route to use for regularization can be based on the available computational infrastructure. For example, it might be much easier to implement data augmentation than model pruning.

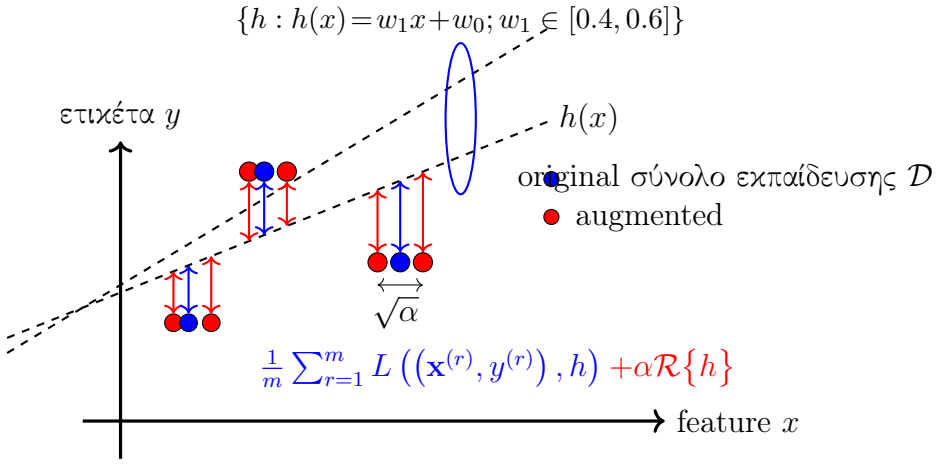


Fig. 22. Three approaches to regularization: 1) data augmentation; 2) loss penalization; and 3) model pruning (via constraints on παράμετροι μοντέλου).

Βλέπε επίσης: ml, model, effective dimension, εμπειρική ελαχιστοποίηση διακινδύνευσης, υπερπροσαρμογή, υπόθεση, σύνολο εκπαίδευσης, παράμετροι μοντέλου, feature, γραμμική παλινδρόμηση, loss, objective function, training error, διακινδύνευση, data augmentation, data point, realization, RV, feature vector, Gaussian RV, ridge regression, ετικέτα.

regularized empirical risk minimization (RERM) Basic εμπειρική ελαχιστοποίηση διακινδύνευσης learns a υπόθεση (or trains a model) $h \in \mathcal{H}$ based solely on the empirical risk $\widehat{L}\big(h|\mathcal{D}\big)$ incurred on a σύνολο εκπαίδευσης $\mathcal{D}$. To make εμπειρική ελαχιστοποίηση διακινδύνευσης less prone to υπερπροσαρμογή, we can implement regularization by including a (scaled) regularizer $\mathcal{R}\{h\}$ in the learning objective. This leads to

regularized εμπειρική ελαχιστοποίηση διακινδύνευσης,

$$\hat{h} \in \operatorname*{argmin}_{h \in \mathcal{H}} \widehat{L}\big(h|\mathcal{D}\big) + \alpha \mathcal{R}\big\{h\big\}. \tag{11}$$

The parameter $\alpha \geq 0$ controls the regularization strength. For $\alpha = 0$, we recover standard εμπειρική ελαχιστοποίηση διακινδύνευσης without regularization. As α increases, the learned υπόθεση is increasingly biased toward small values of $\mathcal{R}\big\{h\big\}$. The component $\alpha \mathcal{R}\big\{h\big\}$ in the objective function of (11) can be intuitively understood as a surrogate for the increased average loss that may occur when predicting ετικέτας for data points outside the σύνολο εκπαίδευσης. This intuition can be made precise in various ways. For example, consider a γραμμικό μοντέλο trained using τετραγωνική απώλεια σφάλματος and the regularizer $\mathcal{R}\big\{h\big\} = \|\mathbf{w}\|_2^2$. In this setting, $\alpha \mathcal{R}\big\{h\big\}$ corresponds to the expected increase in loss caused by adding Gaussian RVs to the feature vectors in the σύνολο εκπαίδευσης [6, Ch. 3]. A principled construction for the regularizer $\mathcal{R}\big\{h\big\}$ arises from approximate upper bounds on the generalization error. The resulting regularized εμπειρική ελαχιστοποίηση διακινδύνευσης instance is known as SRM [107, Sec. 7.2].

Βλέπε επίσης: εμπειρική ελαχιστοποίηση διακινδύνευσης, υπόθεση, model, empirical risk, σύνολο εκπαίδευσης, υπερπροσαρμογή, regularization, regularizer, objective function, loss, ετικέτα, data point, γραμμικό μοντέλο, τετραγωνική απώλεια σφάλματος, Gaussian RV, feature vector, generalization, SRM.

generalization Many current ml (and τεχνητή νοημοσύνη) systems are based on εμπειρική ελαχιστοποίηση διακινδύνευσης: At their core, they train a

model (i.e., learn a υπόθεση $\hat{h} \in \mathcal{H}$) by minimizing the average loss (or empirical risk) on some data points $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}$, which serve as a σύνολο εκπαίδευσης $\mathcal{D}^{(\text{train})}$. Generalization refers to an ml method's ability to perform well outside the σύνολο εκπαίδευσης. Any mathematical theory of generalization needs some mathematical concept for the "outside the σύνολο εκπαίδευσης." For example, statistical learning theory uses a πιθανοτικό μοντέλο such as the i.i.d. assumption for data generation: the data points in the σύνολο εκπαίδευσης are i.i.d. realizations of some underlying κατανομή πιθανότητας $p(\mathbf{z})$. A πιθανοτικό μοντέλο allows us to explore the outside of the σύνολο εκπαίδευσης by drawing additional i.i.d. realizations from $p(\mathbf{z})$. Moreover, using the i.i.d. assumption allows us to define the διακινδύνευση of a trained model $\hat{h} \in \mathcal{H}$ as the expected loss $\bar{L}\big(\hat{h}\big)$. What is more, we can use concentration bounds or convergence results for sequences of i.i.d. RVs to bound the deviation between the empirical risk $\widehat{L}\big(\hat{h}|\mathcal{D}^{(\text{train})}\big)$ of a trained model and its διακινδύνευση [7]. It is possible to study generalization also without using πιθανοτικό μοντέλος. For example, we could use (deterministic) perturbations of the data points in the σύνολο εκπαίδευσης to study its outside. In general, we would like the trained model to be robust, i.e., its πρόβλεψης should not change too much for small perturbations of a data point. Consider a trained model for detecting an object in a smartphone snapshot. The detection result should not change if we mask a small number of randomly chosen pixels in the image [108].

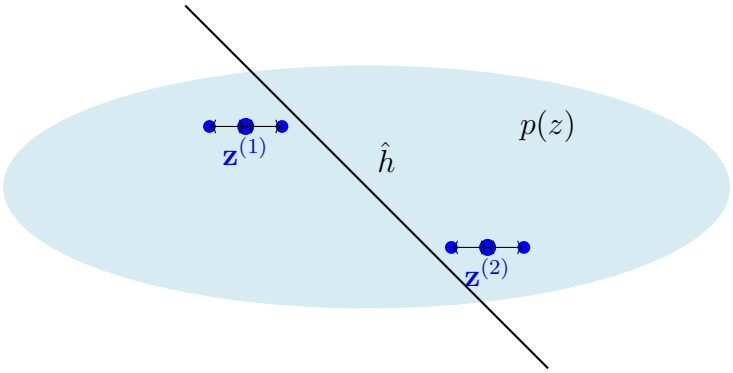


Fig. 23. Two data points $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}$ that are used as a σύνολο εκπαίδευσης to learn a υπόθεση $\hat{h}$ via εμπειρική ελαχιστοποίηση διακινδύνευσης. We can evaluate $\hat{h}$ outside $\mathcal{D}^{(\text{train})}$ either by an i.i.d. assumption with some underlying κατανομή πιθανότητας $p(\mathbf{z})$ or by perturbing the data points.

Βλέπε επίσης: ml, τεχνητή νοημοσύνη, εμπειρική ελαχιστοποίηση διακινδύνευσης, model, υπόθεση, loss, empirical risk, data point, σύνολο εκπαίδευσης, πιθανοτικό μοντέλο, i.i.d. assumption, data, i.i.d., realization, διακινδύνευση, RV, πρόβλεψη, κατανομή πιθανότητας.

generalized total variation (GTV) GTV is a measure of the variation of trained local models $h^{(i)}$ (or their παράμετροι μοντέλου $\mathbf{w}^{(i)}$) assigned to the nodes $i = 1, \ldots, n$ of an undirected weighted graph $\mathcal{G}$ with edges $\mathcal{E}$. Given a measure $d^{(h,h')}$ for the απόκλιση between υπόθεση maps h, h', the GTV is

$$\sum_{\{i,i'\} \in \mathcal{E}} A_{i,i'} d^{(h^{(i)}, h^{(i')})}.$$

Here, $A_{i,i'} > 0$ denotes the weight of the undirected edge $\{i, i'\} \in \mathcal{E}$.

Βλέπε επίσης: local model, παράμετροι μοντέλου, graph, απόκλιση, υ-

πόθεση.

structural risk minimization (SRM) Structural διακινδύνευση minimization is an instance of RERM, which which the model $\mathcal{H}$ can be expressed as a countable union of submodels: $\mathcal{H} = \bigcup_{n=1}^{\infty} \mathcal{H}^{(n)}$. Each submodel $\mathcal{H}^{(n)}$ permits the derivation of an approximate upper bound on the generalization error incurred when applying εμπειρική ελαχιστοποίηση διακινδύνευσης to train $\mathcal{H}^{(n)}$. These individual bounds—one for each submodel—are then combined to form a regularizer used in the RERM objective. These approximate upper bounds (one for each $\mathcal{H}^{(n)}$) are then combined to construct a regularizer for RERM [7, Sec. 7.2].

Βλέπε επίσης: διακινδύνευση, RERM, model, generalization, εμπειρική ελαχιστοποίηση διακινδύνευσης, regularizer.

denial-of-service attack A denial-of-service attack aims (e.g., via data poisoning) to steer the training of a model such that it performs poorly for typical data points.

Βλέπε επίσης: data poisoning, model, data point.

scatterplot A visualization technique that depicts data points by markers in a two-dimensional plane. Fig. 24 depicts an example of a scatterplot.

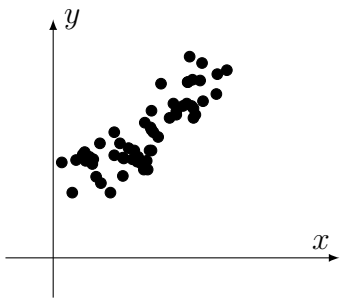


Fig. 24. A scatterplot of some data points representing daily weather conditions in Finland. Each data point is characterized by its ολικό ελάχιστο daytime temperature x as the feature and its maximum daytime temperature y as the ετικέτα. The temperatures have been measured at the Φινλανδικό Μετεωρολογικό Ινστιτούτο weather station Helsinki Kaisaniemi during 1.9.2024 - 28.10.2024.

Βλέπε επίσης: data point, ολικό ελάχιστο, feature, maximum, ετικέτα, Φινλανδικό Μετεωρολογικό Ινστιτούτο.

step size See learning rate.

learning rate Consider an iterative ml method for finding or learning a useful υπόθεση $h \in \mathcal{H}$. Such an iterative method repeats similar computational (update) steps that adjust or modify the current υπόθεση to obtain an improved υπόθεση. One well-known example of such an iterative learning method is κάθοδος κλίσης and its variants, στοχαστική κάθοδος κλίσης and προβεβλημένη κάθοδος κλίσης. A key parameter of an iterative method is the learning rate. The learning rate controls the extent to which the current υπόθεση can be modified during a single iteration. A well-known example of such a parameter is the step size

used in κάθοδος κλίσης [6, Ch. 5].

Βλέπε επίσης: ml, υπόθεση, κάθοδος κλίσης, στοχαστική κάθοδος κλίσης, προβεβλημένη κάθοδος κλίσης, step size.

feature map Feature map refers to a map that transforms the original features of a data point into new features. The so-obtained new features might be preferable over the original features for several reasons. For example, the arrangement of data points might become simpler (or more linear) in the new feature space, allowing the use of γραμμικό μοντέλος in the new features. This idea is a main driver for the development of kernel methods [106]. Moreover, the hidden layers of a βαθύ δίκτυο can be interpreted as a trainable feature map followed by a γραμμικό μοντέλο in the form of the output layer. Another reason for learning a feature map could be that learning a small number of new features helps to avoid υπερπροσαρμογή and ensures ερμηνευσιμότητα [25]. The special case of a feature map delivering two numeric features is particularly useful for data visualization. Indeed, we can depict data points in a scatterplot by using two features as the coordinates of a data point.

Βλέπε επίσης: feature, data point, feature space, γραμμικό μοντέλο, kernel method, βαθύ δίκτυο, υπερπροσαρμογή, ερμηνευσιμότητα, data, scatterplot.

least absolute shrinkage and selection operator (Lasso) The Lasso is an instance of SRM. It learns the βάρη $\mathbf{w}$ of a linear map $h(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ based on a σύνολο εκπαίδευσης. Lasso is obtained from γραμμική παλινδρόμηση by adding the scaled ℓ_1-νόρμα $\alpha \|\mathbf{w}\|_1$ to the average

τετραγωνική απώλεια σφάλματος incurred on the σύνολο εκπαίδευσης.
Βλέπε επίσης: SRM, βάρη, σύνολο εκπαίδευσης, γραμμική παλινδρόμηση, νόρμα, τετραγωνική απώλεια σφάλματος.

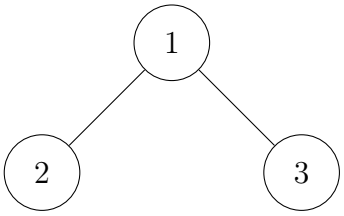
similarity graph Some ml applications generate data points that are related by a domain-specific notion of similarity. These similarities can be represented conveniently using a similarity graph $\mathcal{G} = \big(\mathcal{V} := \{1, \ldots, m\}, \mathcal{E}\big)$. The node $r \in \mathcal{V}$ represents the r-th data point. Two nodes are connected by an undirected edge if the corresponding data points are similar.
Βλέπε επίσης: ml, data point, graph.

Kullback-Leibler divergence (KL divergence) The KL divergence is a quantitative measure of how much one κατανομή πιθανότητας is different from another κατανομή πιθανότητας [13].
Βλέπε επίσης: κατανομή πιθανότητας.

Laplacian matrix The structure of a graph $\mathcal{G}$, with nodes $i = 1, \ldots, n$, can be analyzed using the properties of special matrices that are associated with $\mathcal{G}$. One such matrix is the graph Laplacian matrix $\mathbf{L}^{(\mathcal{G})} \in \mathbb{R}^{n \times n}$, which is defined for an undirected and weighted graph [76], [109]. It is defined element-wise as (see Fig. 25)

$$L_{i,i'}^{(\mathcal{G})} := \begin{cases} -A_{i,i'} & \text{for } i \neq i', \{i,i'\} \in \mathcal{E}, \\ \sum_{i'' \neq i} A_{i,i''} & \text{for } i = i', \\ 0 & \text{else.} \end{cases} \tag{12}$$

Here, $A_{i,i'}$ denotes the βάρος ακμής of an edge $\{i, i'\} \in \mathcal{E}$.



$$\mathbf{L}^{(\mathcal{G})} = \begin{pmatrix} 2 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$$

Fig. 25. Left: Some undirected graph $\mathcal{G}$ with three nodes $i = 1, 2, 3$. Right: The Laplacian matrix $\mathbf{L}^{(\mathcal{G})} \in \mathbb{R}^{3\times 3}$ of $\mathcal{G}$.

Βλέπε επίσης: graph, βάρος ακμής.

kernel Consider data points characterized by a feature vector $\mathbf{x} \in \mathcal{X}$ with a generic feature space $\mathcal{X}$. A (real-valued) kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ assigns each pair of feature vectors $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ a real number $K\big(\mathbf{x}, \mathbf{x}'\big)$. The value $K\big(\mathbf{x}, \mathbf{x}'\big)$ is often interpreted as a measure for the similarity between $\mathbf{x}$ and $\mathbf{x}'$. Kernel methods use a kernel to transform the feature vector $\mathbf{x}$ to a new feature vector $\mathbf{z} = K\big(\mathbf{x}, \cdot\big)$. This new feature vector belongs to a linear feature space $\mathcal{X}'$ which is (in general) different from the original feature space $\mathcal{X}$. The feature space $\mathcal{X}'$ has a specific mathematical structure, i.e., it is a reproducing kernel χώρος Hilbert [106], [15].

Βλέπε επίσης: data point, feature vector, feature space, kernel method, χώρος Hilbert.

kernel method A kernel method is an ml method that uses a kernel K to map the original (raw) feature vector $\mathbf{x}$ of a data point to a new (transformed) feature vector $\mathbf{z} = K\big(\mathbf{x}, \cdot\big)$ [106], [15]. The motivation for transforming the feature vectors is that, by using a suitable kernel, the

data points have a "more pleasant" geometry in the transformed feature space. For example, in a binary ταξινόμηση problem, using transformed feature vectors $\mathbf{z}$ might allow us to use γραμμικό μοντέλος, even if the data points are not linearly separable in the original feature space (see Fig. 26).

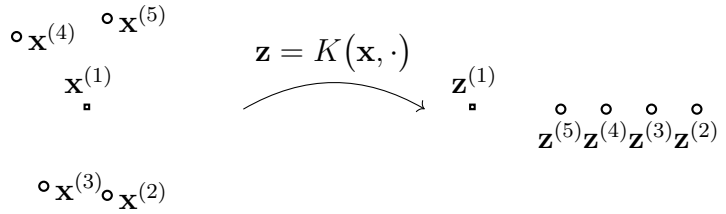


Fig. 26. Five data points characterized by feature vectors $\mathbf{x}^{(r)}$ and ετικέτας $y^{(r)} \in \{\circ, \square\}$, for $r = 1, \ldots, 5$. With these feature vectors, there is no way to separate the two classes by a straight line (representing the όριο απόφασης of a linear classifier). In contrast, the transformed feature vectors $\mathbf{z}^{(r)} = K(\mathbf{x}^{(r)}, \cdot)$ allow us to separate the data points using a linear classifier.

Βλέπε επίσης: kernel, ml, feature vector, data point, feature space, ταξινόμηση, γραμμικό μοντέλο, ετικέτα, όριο απόφασης, linear classifier.

feature matrix Consider a σύνολο δεδομένων $\mathcal{D}$ with m data points with feature vectors $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \in \mathbb{R}^d$. It is convenient to collect the individual feature vectors into a feature matrix $\mathbf{X} := \big(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\big)^T$ of size $m \times d$.

Βλέπε επίσης: σύνολο δεδομένων, data point, feature vector, feature.

density-based spatial clustering of applications with noise (DBSCAN)

DBSCAN refers to a συσταδοποίηση αλγόριθμος for data points that are characterized by numeric feature vectors. Like αλγόριθμος k-μέσων and soft clustering via GMM, also DBSCAN uses the Euclidean distances between feature vectors to determine the συστάδαs. However, in contrast to αλγόριθμος k-μέσων and GMM, DBSCAN uses a different notion of similarity between data points. DBSCAN considers two data points as similar if they are connected via a sequence (path) of close-by intermediate data points. Thus, DBSCAN might consider two data points as similar (and therefore belonging to the same cluster) even if their feature vectors have a large Euclidean distance.

Βλέπε επίσης: συσταδοποίηση, αλγόριθμος, data point, feature vector, αλγόριθμος k-μέσων, soft clustering, GMM, συστάδα.

independent and identically distributed (i.i.d.) It can be useful to interpret data points $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}$ as realizations of i.i.d. RVs with a common κατανομή πιθανότητας. If these RVs are continuous-valued, their joint συνάρτηση πυκνότητας πιθανότητας is $p\big(\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}\big) = \prod_{r=1}^{m} p\big(\mathbf{z}^{(r)}\big)$, with $p(\mathbf{z})$ being the common marginal συνάρτηση πυκνότητας πιθανότητας of the underlying RVs.

Βλέπε επίσης: data point, realization, RV, κατανομή πιθανότητας, συνάρτηση πυκνότητας πιθανότητας.

outlier Many ml methods are motivated by the i.i.d. assumption, which interprets data points as realizations of i.i.d. RVs with a common κατανομή πιθανότητας. The i.i.d. assumption is useful for applications

where the statistical properties of the data generation process are stationary (or time-invariant) [110]. However, in some applications the data consists of a majority of regular data points that conform with an i.i.d. assumption as well as a small number of data points that have fundamentally different statistical properties compared to the regular data points. We refer to a data point that substantially deviates from the statistical properties of most data points as an outlier. Different methods for outlier detection use different measures for this deviation. Stastistical learning theory studies fundamental limits on the ability to mitigate outliers reliably [111], [112].

Βλέπε επίσης: ml, i.i.d. assumption, data point, realization, i.i.d., RV, κατανομή πιθανότητας, data.

Euclidean space The Euclidean space $\mathbb{R}^d$ of dimension $d \in \mathbb{N}$ consists of vectors $\mathbf{x} = \big(x_1, \ldots, x_d\big)$, with d real-valued entries $x_1, \ldots, x_d \in \mathbb{R}$. Such an Euclidean space is equipped with a geometric structure defined by the inner product $\mathbf{x}^T\mathbf{x}' = \sum_{j=1}^{d} x_j x'_j$ between any two vectors $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$ [2].

explainable empirical risk minimization (EERM) Explainable εμπειρική ελαχιστοποίηση διακινδύνευσης is an instance of SRM that adds a regularization term to the average loss in the objective function of εμπειρική ελαχιστοποίηση διακινδύνευσης. The regularization term is chosen to favor υπόθεση maps that are intrinsically explainable for a specific user. This user is characterized by their πρόβλεψηs provided for the data points in a σύνολο εκπαίδευσης [41].

Βλέπε επίσης: εμπειρική ελαχιστοποίηση διακινδύνευσης, SRM, regular-

ization, loss, objective function, υπόθεση, πρόβλεψη, data point, σύνολο εκπαίδευσης.

μέγιστη πιθανοφάνεια Consider data points $\mathcal{D} = \left\{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}\right\}$ that are interpreted as the realizations of i.i.d. RVs with a common κατανομή πιθανότητας $p(\mathbf{z}; \mathbf{w})$ which depends on the παράμετροι μοντέλου $\mathbf{w} \in \mathcal{W} \subseteq \mathbb{R}^n$. Maximum likelihood methods learn παράμετροι μοντέλου $\mathbf{w}$ by maximizing the probability (density) $p(\mathcal{D}; \mathbf{w}) = \prod_{r=1}^{m} p(\mathbf{z}^{(r)}; \mathbf{w})$ of the observed σύνολο δεδομένων. Thus, the maximum likelihood estimator is a solution to the optimization problem $\max_{\mathbf{w} \in \mathcal{W}} p(\mathcal{D}; \mathbf{w})$.

Βλέπε επίσης: data point, realization, i.i.d., RV, κατανομή πιθανότητας, παράμετροι μοντέλου, maximum, σύνολο δεδομένων.

Index

References

[1] W. Rudin, *Real and Complex Analysis*, 3rd ed. New York, NY, USA: McGraw-Hill, 1987.

[2] W. Rudin, *Principles of Mathematical Analysis*, 3rd ed. New York, NY, USA: McGraw-Hill, 1976.

[3] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 4th ed. Baltimore, MD, USA: The Johns Hopkins Univ. Press, 2013.

[4] G. H. Golub and C. F. Van Loan, "An analysis of the total least squares problem," *SIAM J. Numer. Anal.*, vol. 17, no. 6, pp. 883–893, Dec. 1980, doi: 10.1137/0717073.

[5] D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to Probability*, 2nd ed. Belmont, MA, USA: Athena Scientific, 2008.

[6] A. Jung, *Machine Learning: The Basics*. Singapore, Singapore: Springer Nature, 2022.

[7] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. New York, NY, USA: Cambridge Univ. Press, 2014.

[8] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *Found. Trends Mach. Learn.*, vol. 5, no. 1, pp. 1–122, Dec. 2012, doi: 10.1561/2200000024.

[9] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. Cambridge, MA, USA: MIT Press, 2022. [Online].

Available: http://ebookcentral.proquest.com/lib/aalto-ebooks/detail.action?docID=6925615

[10] M. Sipser, *Introduction to the Theory of Computation*, 3rd ed. Andover, U.K.: Cengage Learning, 2013.

[11] R. Motwani and P. Raghavan, *Randomized Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 1995.

[12] R. G. Gallager, *Stochastic Processes: Theory for Applications*. New York, NY, USA: Cambridge Univ. Press, 2013.

[13] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[14] I. Csiszar, "Generalized cutoff rates and Renyi's information measures," *IEEE Trans. Inf. Theory*, vol. 41, no. 1, pp. 26–34, Jan. 1995, doi: 10.1109/18.370121.

[15] C. H. Lampert, "Kernel methods in computer vision," *Found. Trends Comput. Graph. Vis.*, vol. 4, no. 3, pp. 193–285, Sep. 2009, doi: 10.1561/0600000027.

[16] European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance)," L 119/1, May 4, 2016. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj

[17] European Union, "Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018 on the protection of natural persons with regard to the processing of personal data by the Union institutions, bodies, offices and agencies and on the free movement of such data, and repealing Regulation (EC) No 45/2001 and Decision No 1247/2002/EC (Text with EEA relevance)," L 295/39, Nov. 21, 2018. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2018/1725/oj

[18] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[19] M. P. Salinas et al., "A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis," *npj Digit. Med.*, vol. 7, no. 1, May 2024, Art. no. 125, doi: 10.1038/s41746-024-01103-x.

[20] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, 2nd ed. New York, NY, USA: Springer-Verlag, 1998.

[21] G. F. Cooper, "The computational complexity of probabilistic inference using bayesian belief networks," *Artif. Intell.*, vol. 42, no. 2–3, pp. 393–405, Mar. 1990, doi: 10.1016/0004-3702(90)90060-D.

[22] R. M. Gray, *Probability, Random Processes, and Ergodic Properties*, 2nd ed. New York, NY, USA: Springer Science+Business Media, 2009.

[23] N. Parikh and S. Boyd, "Proximal algorithms," *Found. Trends Optim.*, vol. 1, no. 3, pp. 127–239, Jan. 2014, doi: 10.1561/2400000003.

[24] C. Rudin, "Stop explaining black box machine learning models for high-stakes decisions and use interpretable models instead," *Nature Mach.*

Intell., vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[25] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[26] R. T. Rockafellar, *Network Flows and Monotropic Optimization.* Belmont, MA, USA: Athena Scientific, 1998.

[27] E. F. Codd, "A relational model of data for large shared data banks," *Commun. ACM*, vol. 13, no. 6, pp. 377–387, Jun. 1970, doi: 10.1145/362384.362685.

[28] A. Ünsal and M. Önen, "Information-theoretic approaches to differential privacy," *ACM Comput. Surv.*, vol. 56, no. 3, Oct. 2023, Art. no. 76, doi: 10.1145/3604904.

[29] A. Makhdoumi, S. Salamatian, N. Fawaz, and M. Médard, "From the information bottleneck to the privacy funnel," in *2014 IEEE Inf. Theory Workshop*, pp. 501–505, doi: 10.1109/ITW.2014.6970882.

[30] High-Level Expert Group on Artificial Intelligence, "Ethics guidelines for trustworthy AI," European Commission, Apr. 8, 2019. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

[31] A. Jung and P. H. J. Nardelli, "An information-theoretic approach to

personalized explainable machine learning," *IEEE Signal Process. Lett.*, vol. 27, pp. 825–829, 2020, doi: 10.1109/LSP.2020.2993176.

[32] C. Gallese, "The AI act proposal: A new right to technical interpretability?," *SSRN Electron. J.*, Feb. 2023. [Online]. Available: https://ssrn.com/abstract=4398206

[33] T. Gebru et al., "Datasheets for datasets," *Commun. ACM*, vol. 64, no. 12, pp. 86–92, Nov. 2021, doi: 10.1145/3458723.

[34] M. Mitchell et al., "Model cards for model reporting," in *Proc. Conf. Fairness, Accountability, Transparency*, 2019, pp. 220–229, doi: 10.1145/3287560.3287596.

[35] K. Shahriari and M. Shahriari, "IEEE standard review — Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems," in *2017 IEEE Canada Int. Humanitarian Technol. Conf.*, pp. 197–201, doi: 10.1109/IHTC.2017.8058187.

[36] L. Richardson and M. Amundsen, *RESTful Web APIs*. Sebastopol, CA, USA: O'Reilly Media, 2013.

[37] H. H. Bauschke and P. L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, 2nd ed. New York, NY, USA: Springer Science+Business Media, 2017.

[38] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 3rd ed., 2025. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[39] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *2017 IEEE Int. Conf. Comput. Vis.*, pp. 618–626, doi: 10.1109/ICCV.2017.74.

[40] J. Colin, T. Fel, R. Cadène, and T. Serre, "What I cannot predict, I do not understand: A human-centered evaluation framework for explainability methods," in *Adv. Neural Inf. Process. Syst.*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds. vol. 35, 2022, pp. 2832–2845. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/hash/13113e938f2957891c0c5e8df811dd01-Abstract-Conference.html

[41] L. Zhang, G. Karakasidis, A. Odnoblyudova, L. Dogruel, Y. Tian, and A. Jung, "Explainable empirical risk minimization," *Neural Comput. Appl.*, vol. 36, no. 8, pp. 3983–3996, Mar. 2024, doi: 10.1007/s00521-023-09269-3.

[42] J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation," in *Proc. 35th Int. Conf. Mach. Learn.*, J. Dy and A. Krause, Eds. vol. 80, 2018, pp. 883–892. [Online]. Available: https://proceedings.mlr.press/v80/chen18j.html

[43] R. Caruana, "Multitask learning," *Mach. Learn.*, vol. 28, pp. 41–75, Jul. 1997, doi: 10.1023/A:1007379606734.

[44] A. Jung, G. Hannak, and N. Goertz, "Graphical lasso based model

selection for time series," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1781–1785, Oct. 2015, doi: 10.1109/LSP.2015.2425434.

[45] A. Jung, "Learning the conditional independence structure of stationary time series: A multitask learning approach," *IEEE Trans. Signal Process.*, vol. 63, no. 21, Nov. 2015, doi: 10.1109/TSP.2015.2460219.

[46] D. N. Gujarati and D. C. Porter, *Basic Econometrics*, 5th ed. New York, NY, USA: McGraw-Hill/Irwin, 2009.

[47] Y. Dodge, Ed. *The Oxford Dictionary of Statistical Terms*. New York, NY, USA: Oxford Univ. Press, 2003.

[48] B. S. Everitt, *The Cambridge Dictionary of Statistics*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[49] P. Billingsley, *Probability and Measure*, 3rd ed. New York, NY, USA: Wiley, 1995.

[50] D. Sun, K.-C. Toh, and Y. Yuan, "Convex clustering: Model, theoretical guarantee and efficient algorithm," *J. Mach. Learn. Res.*, vol. 22, no. 9, pp. 1–32, Jan. 2021. [Online]. Available: http://jmlr.org/papers/v22/18-694.html

[51] K. Pelckmans, J. De Brabanter, J. A. K. Suykens, and B. De Moor, "Convex clustering shrinkage," presented at the PASCAL Workshop Statist. Optim. Clustering Workshop, 2005.

[52] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[53] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Boston, MA, USA: Kluwer Academic Publishers, 2004.

[54] S. Bubeck, "Convex optimization: Algorithms and complexity," *Found. Trends Mach. Learn.*, vol. 8, no. 3–4, pp. 231–357, Nov. 2015, 10.1561/2200000050.

[55] D. P. Bertsekas, *Convex Optimization Algorithms*. Belmont, MA, USA: Athena Scientific, 2015.

[56] A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, I. Guyon, U. von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. vol. 30, 2017, pp. 5998–6008. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[57] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1993.

[58] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer Science+Business Media, 2006.

[59] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. New York, NY, USA: Cambridge Univ. Press, 2000.

[60] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton, FL, USA: CRC Press, 2015.

[61] E. Abbe, "Community detection and stochastic block models: Recent developments," *J. Mach. Learn. Res.*, vol. 18, no. 177, pp. 1–86, Apr. 2018. [Online]. Available: http://jmlr.org/papers/v18/16-480.html

[62] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed. New York, NY, USA: McGraw-Hill Higher Education, 2002.

[63] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed. New York, NY, USA: Cambridge Univ. Press, 2013.

[64] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Horizontal federated learning," in *Federated Learning*. Cham, Switzerland: Springer Nature, 2020, ch. 4, pp. 49–67.

[65] O. Chapelle, B. Schölkopf, and A. Zien, Eds. *Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, 2006.

[66] P. R. Halmos, *Measure Theory*. New York, NY, USA: Springer-Verlag, 1974.

[67] O. Kallenberg, *Foundations of Modern Probability*. New York, NY, USA: Springer-Verlag, 1997.

[68] A. Lapidoth, *A Foundation in Digital Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[69] L. Condat, "A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms," *J. Optim.*

Theory Appl., vol. 158, no. 2, pp. 460–479, Aug. 2013, doi: 10.1007/s10957-012-0245-9.

[70] L. Bottou, "On-line learning and stochastic approximations," in *On-Line Learning in Neural Networks*, D. Saad, Ed. New York, NY, USA: Cambridge Univ. Press, 1999, ch. 2, pp. 9–42.

[71] A. Silberschatz, H. F. Korth, and S. Sudarshan, *Database System Concepts*, 7th ed. New York, NY, USA: McGraw-Hill Education, 2019. [Online]. Available: https://db-book.com/

[72] S. Abiteboul, R. Hull, and V. Vianu, *Foundations of Databases*. Reading, MA, USA: Addison-Wesley Publishing Company, 1995.

[73] S. Hoberman, *Data Modeling Made Simple: A Practical Guide for Business and IT Professionals*, 2nd ed. Basking Ridge, NJ, USA: Technics Publications, 2009.

[74] R. Ramakrishnan and J. Gehrke, *Database Management Systems*, 3rd ed. New York, NY, USA: McGraw-Hill, 2002.

[75] Y. SarcheshmehPour, Y. Tian, L. Zhang, and A. Jung, "Flow-based clustering and spectral clustering: A comparison," in *2021 55th Asilomar Conf. Signals, Syst., Comput.*, M. B. Matthews, Ed. pp. 1292–1296, doi: 10.1109/IEEECONF53345.2021.9723162.

[76] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, Dec. 2007, doi: 10.1007/s11222-007-9033-z.

[77] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York, NY, USA: Springer Science+Business Media, 2001.

[78] N. Young, *An Introduction to Hilbert Space.* New York, NY, USA: Cambridge Univ. Press, 1988.

[79] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, "Vertical federated learning," in *Federated Learning.* Cham, Switzerland: Springer Nature, 2020, ch. 5, pp. 69–81.

[80] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning.* Cambridge, MA, USA: MIT Press, 2006.

[81] S. Ross, *A First Course in Probability*, 9th ed. Boston, MA, USA: Pearson Education, 2014.

[82] D. Pfau and A. Jung, "Engineering trustworthy AI: A developer guide for empirical risk minimization," Nov. 2024. [Online]. Available: https://arxiv.org/abs/2410.19361

[83] High-Level Expert Group on Artificial Intelligence, "The assessment list for trustworthy artificial intelligence (ALTAI): For self assessment," European Commission, Jul. 17, 2020. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment

[84] G. Tel, *Introduction to Distributed Algorithms*, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[85] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*. Belmont, MA, USA: Athena Scientific, 2015.

[86] N. Cesa-Bianchi and G. Lugosi, *Prediction, Learning, and Games*. New York, NY, USA: Cambridge Univ. Press, 2006.

[87] E. Hazan, "Introduction to online convex optimization," *Found. Trends Optim.*, vol. 2, no. 3–4, pp. 157–325, Aug. 2016, doi: 10.1561/2400000013.

[88] L. Cohen, *Time-Frequency Analysis*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1995.

[89] J. Li, L. Han, X. Li, J. Zhu, B. Yuan, and Z. Gou, "An evaluation of deep neural network models for music classification using spectrograms," *Multimedia Tools Appl.*, vol. 81, no. 4, pp. 4621–4647, Feb. 2022, doi: 10.1007/s11042-020-10465-9.

[90] B. Boashash, Ed. *Time Frequency Signal Analysis and Processing: A Comprehensive Reference*. Oxford, U.K.: Elsevier, 2003.

[91] S. Mallat, *A Wavelet Tour of Signal Processing: The Sparse Way*, 3rd ed. Burlington, MA, USA: Academic, 2009.

[92] Y. SarcheshmehPour, Y. Tian, L. Zhang, and A. Jung, "Clustered federated learning via generalized total variation minimization," *IEEE Trans. Signal Process.*, vol. 71, pp. 4240–4256, 2023, doi: 10.1109/TSP.2023.3322848.

[93] A. Rakhlin, O. Shamir, and K. Sridharan, "Making gradient descent optimal for strongly convex stochastic optimization," in *Proc. 29th Int.*

Conf. Mach. Learn., J. Langford and J. Pineau, Eds. 2012, pp. 449–456. [Online]. Available: https://icml.cc/Conferences/2012/papers/261.pdf

[94] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, no. 1–2, pp. 1–305, Nov. 2008, doi: 10.1561/2200000001.

[95] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[96] P. Bühlmann and S. van de Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin, Germany: Springer-Verlag, 2011.

[97] A. Jung, "Networked exponential families for big data over networks," *IEEE Access*, vol. 8, pp. 202 897–202 909, Nov. 2020, doi: 10.1109/ACCESS.2020.3033817.

[98] X. Liu, H. Li, G. Xu, Z. Chen, X. Huang, and R. Lu, "Privacy-enhanced federated learning against poisoning adversaries," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4574–4588, 2021, doi: 10.1109/TIFS.2021.3108434.

[99] J. Zhang, B. Chen, X. Cheng, H. T. T. Binh, and S. Yu, "PoisonGAN: Generative poisoning attacks against federated learning in edge computing systems," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3310–3322, Mar. 2021, doi: 10.1109/JIOT.2020.3023126.

[100] H. P. Lopuhaä and P. J. Rousseeuw, "Breakdown points of affine

equivariant estimators of multivariate location and covariance matrices," *Ann. Statist.*, vol. 19, no. 1, pp. 229–248, Mar. 1991, doi: 10.1214/aos/1176347978.

[101] D. P. Bertsekas, A. Nedic, and A. E. Ozdaglar, *Convex Analysis and Optimization.* Belmont, MA, USA: Athena Scientific, 2003.

[102] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont, MA, USA: Athena Scientific, 1999.

[103] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, A. Singh and J. Zhu, Eds. vol. 54, 2017, pp. 1273–1282. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

[104] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, I. Dhillon, D. Papailiopoulos, and V. Sze, Eds. vol. 2, 2020. [Online]. Available: https://proceedings.mlsys.org/paper_files/paper/2020/hash/1f5fe83998a09396ebe6477d9475ba0c-Abstract.html

[105] K. Abayomi, A. Gelman, and M. Levy, "Diagnostics for multivariate imputations," *J. Roy. Statist. Soc.: Ser. C (Appl. Statist.)*, vol. 57, no. 3, pp. 273–291, Jun. 2008, doi: 10.1111/j.1467-9876.2007.00613.x.

[106] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* Cambridge, MA, USA: MIT Press, 2002.

[107] S. Shalev-Shwartz and A. Tewari, "Stochastic methods for ℓ_1 regularized loss minimization," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, L. Bottou and M. Littman, Eds. Jun. 2009, pp. 929–936.

[108] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019, doi: 10.1109/TEVC.2019.2890858.

[109] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Adv. Neural Inf. Process. Syst.*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds. vol. 14, 2001, pp. 849–856. [Online]. Available: https://papers.nips.cc/paper_files/paper/2001/hash/801272ee79cfde7fa5960571fee36b9b-Abstract.html

[110] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*, 2nd ed. New York, NY, USA: Springer-Verlag, 1991.

[111] M. Kearns and M. Li, "Learning in the presence of malicious errors," *SIAM J. Comput.*, vol. 22, no. 4, pp. 807–837, Aug. 1993, doi: 10.1137/0222052.

[112] G. Lugosi and S. Mendelson, "Robust multivariate mean estimation: The optimality of trimmed mean," *Ann. Statist.*, vol. 49, no. 1, pp. 393–410, Feb. 2021, doi: 10.1214/20-AOS1961.