The A”allon koneoppimisen sanakirja

Alexander Jung[1], Konstantina Olioumtsevits[1], Ekkehard Schnoor[1], Tommi Flores Ryynänen[1], Juliette Gronier[2], and Salvatore Rastelli[1]
[1]Aalto University [2]ENS Lyon

October 10, 2025

please cite as: A. Jung, K. Olioumtsevits, E. Schnoor, T. Ryynänen, J. Gronier ja S. Rastelli, *Aallon koneoppimisen sanakirja*. Käännös Janita Thusberg, Helli Manninen ja Mikko Seesto. Espoo, Finland: Aalto University, 2025.

Acknowledgment

This dictionary of machine learning evolved through the development and teaching of several courses, including 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, and CS-E407507 Human-Centered Machine Learning. These courses were offered at Aalto University `https://www.aalto.fi/en`, to adult learners via The Finnish Institute of Technology (FITech) `https://fitech.io/en/`, and to international students through the European University Alliance Unite! `https://www.aalto.fi/en/unite`.

We are grateful to the students who provided valuable feedback that helped shape this dictionary. Special thanks to Mikko Seesto for his meticulous proofreading.

This work was supported by

- the Research Council of Finland (grants 331197, 363624, 349966);
- the European Union (grant 952410);
- the Jane and Aatos Erkko Foundation (grant A835);
- Business Finland, as part of the project Forward-Looking AI Governance in Banking and Insurance (FLAIG).

Contents

Lists of Symbols

Sets and Functions

$a \in \mathcal{A}$	The object a is an element of the set $\mathcal{A}$.
$a := b$	We use a as a shorthand for b.
$\lvert\mathcal{A}\rvert$	The cardinality (i.e., number of elements) of a finite set $\mathcal{A}$.
$\mathcal{A} \subseteq \mathcal{B}$	$\mathcal{A}$ is a subset of $\mathcal{B}$.
$\mathcal{A} \subset \mathcal{B}$	$\mathcal{A}$ is a strict subset of $\mathcal{B}$.
$\mathcal{A} \times \mathcal{B}$	The Cartesian product of the sets $\mathcal{A}$ and $\mathcal{B}$.
$\mathbb{N}$	The natural numbers 1, 2, ...
$\mathbb{R}$	The real numbers x [?].
$\mathbb{R}_+$	The nonnegative real numbers $x \geq 0$.
$\mathbb{R}_{++}$	The positive real numbers $x > 0$.
$\{0, 1\}$	The set consisting of the two real numbers 0 and 1.
$[0, 1]$	The closed interval of real numbers x with $0 \leq x \leq 1$.

$\arg\min_{\mathbf{w}} f(\mathbf{w})$	The set of minimizers for a real-valued function $f(\mathbf{w})$. See also: function.
$\mathbb{S}^{(n)}$	The set of unit-norm vectors in $\mathbb{R}^{n+1}$. See also: norm, vector.
$\exp(a)$	The exponential function evaluated at the real number $a \in \mathbb{R}$. See also: function.
$\log a$	The logarithm of the positive number $a \in \mathbb{R}_{++}$.
$f(\cdot): \mathcal{A} \rightarrow \mathcal{B} : a \mapsto f(a)$	A function (or map) from a set $\mathcal{A}$ to a set $\mathcal{B}$, which assigns to each input $a \in \mathcal{A}$ a well-defined output $f(a) \in \mathcal{B}$. The set $\mathcal{A}$ is the domain of the function f and the set $\mathcal{B}$ is the co-domain of f. Machine learning (ML) aims to learn a function h that maps features $\mathbf{x}$ of a data point to a prediction $h(\mathbf{x})$ for its label y. See also: function, map, ML, feature, data point, prediction, label.
$\operatorname{epi}(f)$	The epigraph of a real-valued function $f : \mathbb{R}^d \rightarrow \mathbb{R}$. See also: epigraph, function.
$\frac{\partial f(w_1, \ldots, w_d)}{\partial w_j}$	The partial derivative (if it exists) of a real-valued function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ with respect to w_j [**?**, Ch. 9]. See also: function.

$\nabla f(\mathbf{w})$ The gradient of a differentiable real-valued function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is the vector $\nabla f(\mathbf{w}) = \big(\partial f/\partial w_1, \ldots, \partial f/\partial w_d\big)^T \in \mathbb{R}^d$ [**?**, Ch. 9].
See also: gradient, differentiable, function, vector.

Matrices and Vectors

$\mathbf{x} = \big(x_1, \ldots, x_d\big)^T$	A vector of length d, with its jth entry being x_j. See also: vector.
$\mathbb{R}^d$	The set of vectors $\mathbf{x} = \big(x_1, \ldots, x_d\big)^T$ consisting of d real-valued entries $x_1, \ldots, x_d \in \mathbb{R}$. See also: vector.
$\mathbf{I}_{l \times d}$	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 otherwise equal to 0. See also: matrix.
$\mathbf{I}_d, \mathbf{I}$	A square identity matrix of size $d \times d$. If the size is clear from context, we drop the subscript. See also: matrix.
$\lVert\mathbf{x}\rVert_2$	The Euclidean (or ℓ_2) norm of the vector $\mathbf{x} = \big(x_1, \ldots, x_d\big)^T \in \mathbb{R}^d$ defined as $\lVert\mathbf{x}\rVert_2 := \sqrt{\sum_{j=1}^{d} x_j^2}$. See also: norm, vector.
$\lVert\mathbf{x}\rVert$	Some norm of the vector $\mathbf{x} \in \mathbb{R}^d$ [?]. Unless otherwise specified, we mean the Euclidean norm $\lVert\mathbf{x}\rVert_2$. See also: norm, vector.
$\mathbf{x}^T$	The transpose of a matrix that has the vector $\mathbf{x} \in \mathbb{R}^d$ as its single column. See also: matrix, vector.

$\mathbf{X}^T$	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 $\mathbf{X} = \mathbf{X}^T$. See also: matrix.
$\mathbf{X}^{-1}$	The inverse matrix of a matrix $\mathbf{X} \in \mathbb{R}^{d \times d}$. See also: inverse matrix, matrix.
$\mathbf{0} = \big(0, \dots, 0\big)^T$	The vector in $\mathbb{R}^d$ with each entry equal to zero. See also: vector.
$\mathbf{1} = \big(1, \dots, 1\big)^T$	The vector in $\mathbb{R}^d$ with each entry equal to one. See also: vector.
$\big(\mathbf{v}^T, \mathbf{w}^T\big)^T$	The vector of length $d + d'$ obtained by concatenating the entries of vector $\mathbf{v} \in \mathbb{R}^d$ with the entries of $\mathbf{w} \in \mathbb{R}^{d'}$. See also: vector.
$\mathrm{span}\{\mathbf{B}\}$	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 $\mathbf{B}$, such that $\mathrm{span}\{\mathbf{B}\} = \big\{\mathbf{Ba} : \mathbf{a} \in \mathbb{R}^b\big\} \subseteq \mathbb{R}^a$. See also: matrix.
$\mathrm{null}(\mathbf{A})$	The nullspace of a matrix $\mathbf{A} \in \mathbb{R}^{a \times b}$, which is the subspace of vectors $\mathbf{a} \in \mathbb{R}^b$ such that $\mathbf{Aa} = \mathbf{0}$. See also: nullspace, matrix, vector.

$\det(\mathbf{C})$	The determinant of the matrix $\mathbf{C}$. See also: determinant, matrix.
$\mathbf{A} \otimes \mathbf{B}$	The Kronecker product of $\mathbf{A}$ and $\mathbf{B}$ [**?**]. See also: Kronecker product.

Probability Theory

$\mathbf{x} \sim p(\mathbf{z})$	The random variable (RV) $\mathbf{x}$ is distributed according to the probability distribution $p(\mathbf{z})$ [**?**], [**?**]. See also: RV, probability distribution.
$\mathbb{E}_p\{f(\mathbf{z})\}$	The expectation of an RV $f(\mathbf{z})$ that is obtained by applying a deterministic function f to an RV $\mathbf{z}$ whose probability distribution is $\mathbb{P}(\mathbf{z})$. If the probability distribution is clear from context, we just write $\mathbb{E}\{f(\mathbf{z})\}$. See also: expectation, RV, function, probability distribution.
$\mathrm{cov}\,(x, y)$	The covariance between two real-valued RVs defined over a common probability space. See also: covariance, RV, probability distribution.
$\mathbb{P}(\mathbf{x}, y)$	A (joint) probability distribution of an RV whose realizations are data points with features $\mathbf{x}$ and label y. See also: probability distribution, RV, realization, data point, feature, label.
$\mathbb{P}(\mathbf{x}\|y)$	A conditional probability distribution of an RV $\mathbf{x}$ given the value of another RV y [**?**, Sec. 3.5]. See also: probability distribution, RV.
$\mathbb{P}(\mathcal{A})$	The probability of the measurable event $\mathcal{A}$. See also: probability, measurable, event.

$\mathbb{P}(\mathbf{x};\mathbf{w})$	A parameterized probability distribution of an RV $\mathbf{x}$. The probability distribution depends on a parameter vector $\mathbf{w}$. For example, $\mathbb{P}(\mathbf{x};\mathbf{w})$ could be a multivariate normal distribution with the parameter vector $\mathbf{w}$ given by the entries of the mean vector $\mathbb{E}\{\mathbf{x}\}$ and the covariance matrix $\mathbb{E}\Big\{\big(\mathbf{x}-\mathbb{E}\{\mathbf{x}\}\big)\big(\mathbf{x}-\mathbb{E}\{\mathbf{x}\}\big)^{T}\Big\}$. See also: probability distribution, parameter, probabilistic model.
$\mathcal{N}(\mu,\sigma^2)$	The probability distribution of a Gaussian random variable (Gaussian RV) $x \in \mathbb{R}$ with mean (or expectation) $\mu = \mathbb{E}\{x\}$ and variance $\sigma^2 = \mathbb{E}\big\{(x-\mu)^2\big\}$. See also: probability distribution, Gaussian RV.
$\mathcal{N}(\boldsymbol{\mu},\mathbf{C})$	The multivariate normal distribution of a vector-valued Gaussian RV $\mathbf{x} \in \mathbb{R}^d$ with mean (or expectation) $\boldsymbol{\mu} = \mathbb{E}\{\mathbf{x}\}$ and covariance matrix $\mathbf{C} = \mathbb{E}\big\{\big(\mathbf{x}-\boldsymbol{\mu}\big)\big(\mathbf{x}-\boldsymbol{\mu}\big)^{T}\big\}$. See also: multivariate normal distribution, Gaussian RV.
Ω	A sample space of all possible outcomes of a random experiment. See also: event.
$\mathcal{F}$	A collection of measurable subsets of a sample space Ω. See also: sample space, event.
$\mathcal{P}$	A probability space that consists of a sample space Ω, a σ-algebra $\mathcal{F}$ of measurable subsets of Ω, and a probability distribution $\mathbb{P}(\cdot)$. See also: sample space, measurable, probability distribution.

Machine Learning

r	An index $r = 1, 2, \ldots$ that enumerates data points. See also: data point.
m	The number of data points in (i.e., the size of) a dataset. See also: data point, dataset.
$\mathcal{D}$	A dataset $\mathcal{D} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}\}$ is a list of individual data points $\mathbf{z}^{(r)}$, for $r = 1, \ldots, m$. See also: dataset, data point.
d	The number of features that characterize a data point. See also: feature, data point.
x_j	The jth feature of a data point. The first feature is denoted by x_1, the second feature x_2, and so on. See also: data point, feature.
$\mathbf{x}$	The feature vector $\mathbf{x} = \big(x_1, \ldots, x_d\big)^T$ of a data point. The vector's entries are the individual features of a data point. See also: feature vector, data point, vector, feature.
$\mathcal{X}$	The feature space $\mathcal{X}$ is the set of all possible values that the features $\mathbf{x}$ of a data point can take on. See also: feature space, feature, data point.

$\mathbf{z}$	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 [**?**, Ch. 9]. See also: vector, feature, data point.
$\mathbf{x}^{(r)}$	The feature vector of the rth data point within a dataset. See also: feature vector, data point, dataset.
$x_j^{(r)}$	The jth feature of the rth data point within a dataset. See also: feature, data point, dataset.
$\mathcal{B}$	A mini-batch (or subset) of randomly chosen data points. See also: batch, data point.
B	The size of (i.e., the number of data points in) a mini-batch. See also: data point, batch.
y	The label (or quantity of interest) of a data point. See also: label, data point.
$y^{(r)}$	The label of the rth data point. See also: label, data point.
$\left(\mathbf{x}^{(r)}, y^{(r)}\right)$	The features and label of the rth data point. See also: feature, label, data point.

$\mathcal{Y}$	The label space $\mathcal{Y}$ of an ML method consists of all potential label values that a data point can carry. The nominal label space might be larger than the set of different label values arising in a given dataset (e.g., a training set). 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 classification problems (or methods). See also: label space, ML, label, data point, dataset, training set, regression, classification.
η	Learning rate (or step size) used by gradient-based methods. See also: learning rate, step size, gradient-based methods.
$h(\cdot)$	A hypothesis map that maps the features of a data point to a prediction $\hat{y} = h(\mathbf{x})$ for its label y. See also: hypothesis, map, feature, data point, prediction, label.
$\mathcal{Y}^{\mathcal{X}}$	Given two sets $\mathcal{X}$ and $\mathcal{Y}$, we denote by $\mathcal{Y}^{\mathcal{X}}$ the set of all possible hypothesis maps $h : \mathcal{X} \to \mathcal{Y}$. See also: hypothesis, map.
$\mathcal{H}$	A hypothesis space or model used by an ML method. The hypothesis space consists of different hypothesis maps $h : \mathcal{X} \to \mathcal{Y}$, between which the ML method must choose. See also: hypothesis space, model, ML, hypothesis, map.

$d_{\text{eff}}(\mathcal{H})$	The effective dimension of a hypothesis space $\mathcal{H}$. See also: effective dimension, hypothesis space.
B^2	The squared bias of a learned hypothesis $\hat{h}$, or its parameters. Note that $\hat{h}$ becomes an RV if it is learned from data points being RVs themselves. See also: bias, hypothesis, parameter, RV, data point.
V	The variance of a learned hypothesis $\hat{h}$, or its parameters. Note that $\hat{h}$ becomes an RV if it is learned from data points being RVs themselves. See also: variance, hypothesis, parameter, RV, data point.
$L((\mathbf{x}, y), h)$	The loss incurred by predicting the label y of a data point using the prediction $\hat{y} = h(\mathbf{x})$. The prediction $\hat{y}$ is obtained by evaluating the hypothesis $h \in \mathcal{H}$ for the feature vector $\mathbf{x}$ of the data point. See also: loss, label, data point, prediction, hypothesis, feature vector.
E_v	The validation error of a hypothesis h, which is its average loss incurred over a validation set. See also: validation error, hypothesis, loss, validation set.
$\widehat{L}(h\|\mathcal{D})$	The empirical risk, or average loss, incurred by the hypothesis h on a dataset $\mathcal{D}$. See also: empirical risk, loss, hypothesis, dataset.

E_t	The training error of a hypothesis h, which is its average loss incurred over a training set. See also: training error, hypothesis, loss, training set.
t	A discrete-time index $t = 0, 1, \ldots$ used to enumerate sequential events (or time instants). See also: event.
t	An index that enumerates learning tasks within a multitask learning problem. See also: learning task, multitask learning.
α	A regularization parameter that controls the amount of regularization. See also: regularization, parameter.
$\lambda_j(\mathbf{Q})$	The jth eigenvalue (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. See also: eigenvalue, psd, matrix.
$\sigma(\cdot)$	The activation function used by an artificial neuron within an artificial neural network (ANN). See also: activation function, ANN.
$\mathcal{R}_{\hat{y}}$	A decision region within a feature space. See also: decision region, feature space.

$\mathbf{w}$	A parameter vector $\mathbf{w} = \big(w_1, \ldots, w_d\big)^T$ of a model, e.g., the weights of a linear model or an ANN. See also: parameter, vector, model, weights, linear model, ANN.
$h^{(\mathbf{w})}(\cdot)$	A hypothesis map that involves tunable model parameters $w_1, \ldots, w_d$ stacked into the vector $\mathbf{w} = \big(w_1, \ldots, w_d\big)^T$. See also: hypothesis, map, model parameters, vector.
$\phi(\cdot)$	A feature map $\phi : \mathcal{X} \rightarrow \mathcal{X}' : \mathbf{x} \mapsto \phi\big(\mathbf{x}\big)$ that transforms the feature vector $\mathbf{x}$ of a data point into a new feature vector $\mathbf{x}' = \phi\big(\mathbf{x}\big) \in \mathcal{X}'$. See also: feature map.
$K\big(\cdot, \cdot\big)$	Given some feature space $\mathcal{X}$, a kernel is a map $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{C}$ that is psd. See also: feature space, kernel, map, psd.
$\mathrm{VCdim}\left(\mathcal{H}\right)$	The Vapnik–Chervonenkis dimension (VC dimension) of the hypothesis space $\mathcal{H}$. See also: VC dimension, hypothesis space.

Federated Learning

$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	An undirected graph whose nodes $i \in \mathcal{V}$ represent devices within a federated learning network (FL network). The undirected weighted edges $\mathcal{E}$ represent connectivity between devices and statistical similarities between their datasets and learning tasks. See also: graph, device, FL network, dataset, learning task.
$i \in \mathcal{V}$	A node that represents some device within an FL network. The device can access a local dataset and train a local model. See also: device, FL network, local dataset, 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 $\mathcal{G}$. See also: Laplacian matrix, graph.
$\mathbf{L}^{(\mathcal{C})}$	The Laplacian matrix of the induced graph $\mathcal{G}^{(\mathcal{C})}$. See also: Laplacian matrix, graph.
$\mathcal{N}^{(i)}$	The neighborhood of the node i in a graph $\mathcal{G}$. See also: neighborhood, graph.
$d^{(i)}$	The weighted node degree $d^{(i)} := \sum_{i' \in \mathcal{N}^{(i)}} A_{i,i'}$ of node i. See also: node degree.
$d_{\max}^{(\mathcal{G})}$	The maximum weighted node degree of a graph $\mathcal{G}$. See also: maximum, node degree, graph.

$\mathcal{D}^{(i)}$	The local dataset $\mathcal{D}^{(i)}$ carried by node $i \in \mathcal{V}$ of an FL network. See also: local dataset, FL network.
m_i	The number of data points (i.e., sample size) contained in the local dataset $\mathcal{D}^{(i)}$ at node $i \in \mathcal{V}$. See also: data point, sample size, local dataset.
$\mathbf{x}^{(i,r)}$	The features of the rth data point in the local dataset $\mathcal{D}^{(i)}$. See also: feature, data point, local dataset.
$y^{(i,r)}$	The label of the rth data point in the local dataset $\mathcal{D}^{(i)}$. See also: label, data point, local dataset.
$\mathbf{w}^{(i)}$	The local model parameters of device i within an FL network. See also: model parameters, device, FL network.
$L_i(\mathbf{w})$	The local loss function used by device i to measure the usefulness of some choice $\mathbf{w}$ for the local model parameters. See also: loss function, device, model parameters.
$L^{(\mathrm{d})}\left(\mathbf{x}, h\big(\mathbf{x}\big), h'\big(\mathbf{x}\big)\right)$	The loss incurred by a hypothesis h' on a data point with features $\mathbf{x}$ and label $h\big(\mathbf{x}\big)$ that is obtained from another hypothesis. See also: loss, hypothesis, data point, feature, label.

$\mathrm{stack}\big\{\mathbf{w}^{(i)}\big\}_{i=1}^{n}$

The vector $\left(\left(\mathbf{w}^{(1)}\right)^{T}, \ldots, \left(\mathbf{w}^{(n)}\right)^{T}\right)^{T} \in \mathbb{R}^{dn}$ that is obtained by vertically stacking the local model parameters $\mathbf{w}^{(i)} \in \mathbb{R}^{d}$, for $i = 1, \ldots, n$.

See also: vector, model parameters.

Tools

This document is incomplete. The external file associated with the glossary 'math' (which should be called `ADictML_Finnish.math-gls`) hasn't been created.

Check the contents of the file `ADictML_Finnish.math-glo`. If it's empty, that means you haven't indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can't be generated. If the file isn't empty, the document build process hasn't been completed.

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`. For example:

  ```
  \usepackage[automake]{glossaries-extra}
  ```

- Run the external (Lua) application:

  ```
  makeglossaries-lite.lua "ADictML_Finnish"
  ```

- Run the external (Perl) application:

  ```
  makeglossaries "ADictML_Finnish"
  ```

Then rerun LaTeX on this document.

This message will be removed once the problem has been fixed.

Machine Learning Concepts

This document is incomplete. The external file associated with the glossary 'main' (which should be called `ADictML_Finnish.gls`) hasn't been created.

Check the contents of the file `ADictML_Finnish.glo`. If it's empty, that means you haven't indexed any of your entries in this glossary (using commands like `\gls` or `\glsadd`) so this list can't be generated. If the file isn't empty, the document build process hasn't been completed.

If you don't want this glossary, add `nomain` to your package option list when you load `glossaries-extra.sty`. For example:

```
\usepackage[nomain]{glossaries-extra}
```

Try one of the following:

- Add `automake` to your package option list when you load `glossaries-extra.sty`. For example:

  ```
  \usepackage[automake]{glossaries-extra}
  ```

- Run the external (Lua) application:

  ```
  makeglossaries-lite.lua "ADictML_Finnish"
  ```

- Run the external (Perl) application:

  ```
  makeglossaries "ADictML_Finnish"
  ```

Then rerun LaTeX on this document.

This message will be removed once the problem has been fixed.

Machine Learning Systems

This document is incomplete. The external file associated with the glossary 'systems' (which should be called `ADictML_Finnish.systems-gls`) hasn't been created.

This has probably happened because there are no entries defined in this glossary. Did you forget to use `type=systems` when you defined your entries? If you tried to load entries into this glossary with `\loadglsentries` did you remember to use `[systems]` as the optional argument? If you did, check that the definitions in the file you loaded all had the type set to `\glsdefaulttype`.

This message will be removed once the problem has been fixed.