

# Mathematical Notations for Learning from Data Course

Islem Rekik\*

Istanbul Technical University

**Abstract. Inspiring quotes.** *“You’ll need to set small, specific goals to master a skill, but first you’ll want to be sure of the basics.”* Source: *Learn better* by Ulrich Boser.

*“Learning is an iterative process that requires that you revisit what you have learnt.”* Source: *Make it stick* by Henry L. Roediger III and Mark A. McDaniel.

*“The good news is that we now know of simple and practical strategies that anybody can use; at any point in life, to learn better and remember longer: various forms of retrieval practice, such as low-stakes quizzing and self-testing, spacing out practice, interleaving the practice of different but related topics or skills, trying to solve a problem before being taught the solution, distilling the underlying principles or rules that differentiate types of problems, and so on.”* Source: *Make it stick* by Henry L. Roediger III and Mark A. McDaniel.

**Pre-requisites:** Linear algebra. To refresh your memory, you can check 3blue1brown YouTube playlist<sup>a</sup>.

**Machine Learning Blinks.** Check the YouTube link below to watch the lectures<sup>b</sup>.

---

\* Corresponding author: irekik@itu.edu.tr; <http://basira-lab.com>

<sup>a</sup> [https://www.youtube.com/watch?v=fNk\\_zzaMoSs&list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE\\_ab](https://www.youtube.com/watch?v=fNk_zzaMoSs&list=PLZHQObOWTQDPD3MizzM2xVFitgF8hE_ab)

<sup>b</sup> <https://www.youtube.com/watch?v=HyWmnlahXAA&list=PLug43ldmRSolLDlvQOPzgoJ6wKnfmzimQ>

Table 1: *Major mathematical notations used in lecture 1.*

Mathematical notation	Definition
$\mathcal{D}$	dataset
$n$	number of samples in a dataset $\mathcal{D}$
$d$	number of features
$\mathbf{x} \in \mathbb{R}^{d \times 1}$	feature vector or data point (sample)
$\mathbf{x}_{feature}^{sample}$	—
$\mathbf{x}^i \in \mathbb{R}^{d \times 1}$	$i^{th}$ sample in the population
$\mathbf{x}_j^i \in \mathbb{R}$	$j^{th}$ feature of $i^{th}$ sample in the population
$\mathcal{D} = \{\mathbf{x}^i, y^i\}_{i=1}^n$	training dataset where $\mathbf{x}^i \in \mathbb{R}^d$ denotes the feature vector for the $i^{th}$ sample and $y^i \in \mathbb{R}$ denotes its score
$\mathbf{X} \in \mathbb{R}^{d \times n}$	data matrix stacking all samples vertically
$f$	mapping or transformation function to learn
$f : \mathbb{R} \rightarrow \mathbb{R}$	one-to-one mapping
$f : \mathbb{R} \rightarrow \mathbb{R}^n$	one-to-many mapping
$f : \mathbb{R}^p \rightarrow \mathbb{R}$	many-to-one mapping
$f : \mathbb{R}^p \rightarrow \mathbb{R}^m$	many-to-many mapping

Table 2: Major mathematical notations used in lecture 3.

Mathematical notation	Definition
$\mathcal{D}$	dataset
$n$	number of samples in a dataset $\mathcal{D}$
$d$	number of features
$\mathbf{x} \in \mathbb{R}^{d \times 1}$	feature vector or data point (sample)
$\mathbf{x}_{sample}^{feature}$	–
$\mathbf{x}^i \in \mathbb{R}^{d \times 1}$	$i^{th}$ sample in the population
$\mathbf{x}_j^i \in \mathbb{R}$	$j^{th}$ feature of $i^{th}$ sample in the population
$\Sigma \in \mathbb{R}^{d \times d}$	covariance matrix of data population $\{\mathbf{x}^i\}_{i=1}^n$
$ \mathbf{A}  \in \mathbb{R}$	determinant of matrix $A$
$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2} \frac{\ x-\mu\ _2^2}{\sigma^2})$	probability density function of a variable $x \in \mathbb{R}$
$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}  \Sigma ^{1/2}} \exp[-\frac{1}{2} (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)]$	probability density function of a multidimensional variable $\mathbf{x} \in \mathbb{R}^{d \times 1}$
$\mu \in \mathbb{R}^{d \times 1}$	sample mean $\mu = \frac{1}{n} \sum_{i=1}^n \mathbf{x}^i$
$\ \mathbf{x} - \mu\ _{\Sigma^{-1}} = (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu) \in \mathbb{R}$	Mahalanobis distance between $\mathbf{x}$ and $\mu$
$\mathbf{I}_{d \times d} \in \mathbb{R}^{d \times d}$	identity matrix of size $d \times d$
$\ \mathbf{x} - \mu\ _{\mathbf{I}_{d \times d}} = (\mathbf{x} - \mu)^T (\mathbf{x} - \mu) \in \mathbb{R}$	Euclidean distance between $\mathbf{x}$ and $\mu$
$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + \mathbf{w}_{i0}$	also noted as $L_2$ norm $\ \cdot\ _2$ discriminant Bayes function for class $i$ when $\Sigma_i = \sigma_i^2 \mathbf{I}$ <u>general case</u> : when $\sigma_i^2 \neq \sigma_j^2$ for classes $i$ and $j$ (i.e., different means $\mu_i \neq \mu_j$ but constant variance for all data features in each class) <u>special case</u> : when $\sigma_i^2 = \sigma_j^2$ for classes $i$ and $j$ (i.e., different means $\mu_i \neq \mu_j$ but constant variances across all classes) (i.e., lines connecting means of different classes are perpendicular to decision boundaries) if $\ln(p(c_i)) = \ln(p(c_j))$ , $g_i(\mathbf{x}) = -\ \mathbf{x} - \mu_i\ _2^2$
$g_i(\mathbf{x}) = \mathbf{w}_i^T \mathbf{x} + \mathbf{w}_{i0}$	discriminant Bayes function for class $i$ when $\Sigma_i = \Sigma_j = \Sigma$ (i.e., constant data feature covariance $\Sigma$ across classes) (i.e., lines connecting means of different classes are not perpendicular to decision boundaries) if $\ln(p(c_i)) = \ln(p(c_j))$ , $g_i(\mathbf{x}) = -\frac{1}{2} \ \mathbf{x} - \mu_i\ _{\Sigma^{-1}}^2$
$g_i(\mathbf{x}) = \mathbf{x}^T \mathbf{W} \mathbf{x} + \mathbf{w}_i^T \mathbf{x} + \mathbf{w}_{i0}$	quadratic discriminant function (decision boundaries are nonlinear)

Table 3: Major mathematical notations used in lectures 4.

Mathematical notation	Definition
$\arg \min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \frac{1}{n} \sum_{(\mathbf{x}^i, y^i) \in \mathcal{D}} \textcolor{brown}{E}(f_{\mathbf{w}}(\mathbf{x}^i), y^i) + \textcolor{brown}{R}(\dots)$	supervised learning energy cost (loss function)
$\textcolor{brown}{f}$	the mapping function to learn from $\mathbf{x}^i \rightarrow y^i$
$\textcolor{brown}{E}$	the error function between the predicted target by $\textcolor{brown}{f}$ and the ground truth observation $y^i$
$\textcolor{brown}{R}$	regularization term to avoid overfitting and control model complexity
$\mathbf{w}$	optimization parameters (weight vector)
$h(w) = l(v) + l'(v)(w - v)$	set of parameters that minimize the loss function $\mathcal{L}(\mathbf{w})$
$l'(w)$	first-order Taylor approximation of the loss function $l$ at point $w$ (in 1-dimensional space)
$h(w) = l(v) + l'(v)(w - v) + \frac{1}{2}l''(v)(w - v)^2$	first derivative of function $l$ evaluated at point $w$
$l''(w)$	second-order Taylor approximation of the loss function $l$ at point $w$ (in 1-dimensional space)
$h(\mathbf{w}) = l(\mathbf{v}) + \nabla l(\mathbf{w})(\mathbf{w} - \mathbf{v})$	second derivative of function $l$ evaluated at point $w$
$\mathbf{w} = [w_1 \ w_2 \ \dots \ w_N] \in \mathbb{R}^d$	first-order Taylor approximation of high-dimensional loss function $l$ at vector point $\mathbf{w} \in \mathbb{R}^d$
$\nabla l(\mathbf{w}) \in \mathbb{R}^{d \times 1}$	weight vector to learn
$h(\mathbf{w}) = l(\mathbf{v}) + \nabla l(\mathbf{w})(\mathbf{w} - \mathbf{v}) + \frac{1}{2}(\mathbf{w} - \mathbf{v})^T \nabla^2 l(\mathbf{w})(\mathbf{w} - \mathbf{v})$	gradient vector of the multivariate loss function $l$ at location $\mathbf{w}$
$\nabla^2 l(\mathbf{w}) \in \mathbb{R}^{d \times d}$	note that $\nabla l(\mathbf{w})^T$ is $\in \mathbb{R}^{1 \times d}$ (row vector)
$l'(w) = 0$	$\nabla l(\mathbf{v}) = [\frac{\partial}{\partial w_1} l(\mathbf{v}) \ \frac{\partial}{\partial w_2} l(\mathbf{v}) \ \dots \ \frac{\partial}{\partial w_d} l(\mathbf{v})]^T$
$\nabla l(\mathbf{w}) = \mathbf{0}_{d \times 1}$	second-order Taylor approximation of high-dimensional loss function $l$ at vector point $\mathbf{w} \in \mathbb{R}^d$
$\mathbf{Q} = \mathbf{Q}^T$	Hessian symmetric matrix of second derivatives of $l$ along all its dimensions (variables)
$\frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$	$l$ is many times differentiable at the vector valued input $\mathbf{w}$
$\frac{\partial \mathbf{a}^T \mathbf{X} \mathbf{a}}{\partial \mathbf{a}} = 2\mathbf{X} \mathbf{a}$ (for $\mathbf{X} = \mathbf{X}^T$ )	stationary point (min, max or saddle) for a 1-dimensional function
$l''(w) > 0$	stationary point (min, max or saddle) for an $N$ -dimensional function
$l''(w) < 0$	(all elements of the gradient are zero)
$\frac{\partial \mathbf{A} \mathbf{x}}{\partial \mathbf{x}} = \frac{1}{2}(\mathbf{A}^T + \mathbf{A})$	$\mathbf{Q}$ is symmetric
$\nabla^2 l(\mathbf{w}) = \frac{1}{2}(\mathbf{Q}^T + \mathbf{Q}) \in \mathbb{R}^{d \times d}$	matrix cookbook <sup>a</sup> (also check <sup>b</sup> )
$\nabla l(\mathbf{w}) = \mathbf{Q} \mathbf{w} + \mathbf{r} \in \mathbb{R}^{d \times 1}$	matrix cookbook
$\mathbf{w}^k = \mathbf{w}^{k-1} - \alpha_k \nabla l(\mathbf{w}^{k-1})$	convex function (facing upward)
	concave function (facing downward)
	matrix cookbook
	is the Hessian matrix of $l$ equal to $l(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{Q} \mathbf{w} + \mathbf{r}^T \mathbf{w} + d$
	$d \in \mathbb{R}$ and $\mathbf{r} \in \mathbb{R}^{d \times 1}$
	is the gradient vector of $l$ equal to $l(\mathbf{w}) = \frac{1}{2} \mathbf{w}^T \mathbf{Q} \mathbf{w} + \mathbf{r}^T \mathbf{w} + d$
	gradient descent for finding the optimal $\mathbf{w}$

<sup>a</sup> <https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf>

<sup>b</sup> <https://ninova.itu.edu.tr/en/courses/institute-of-science-and-technology/1580/blg-527e/ekkaynaklar?g179937>