Lecture Notes : Basics of ML

Scott Pesme
INRIA Grenoble

September 30, 2025

Contents

1 Introduction

Slides of the general introduction can be found here.

2 Supervised learning framework

2.1 Regression setting

2.2 Classification setting

A Useful Formulas and Good Practices

A.1 Definitions and notations

Keep in mind that these are the notations I like to use, but these are obviously personal and others will use different ones!!

- the notation x will always be used for input data. E.g. $x_1, \dots, x_n \in \mathbb{R}^d$ could be some dataset. n is then the number of training samples, and d the dimension of each data point. Similarly, y will used for output data, e.g. $y_1, \dots, y_n \in \mathbb{R}$ (or $\in \{0, 1\}$) are output data (or labels). When $d > n$, we often say that we are in the *over-parametrised setting* (also called *under-determined setting*, which can be confusing!), if $d < n$ we say that we are in the *under-parametrised setting* (also called, *under-determined setting*).
- $X = \begin{pmatrix} - & x_1 & - \\ & \vdots & \\ - & x_n & - \end{pmatrix} \in \mathbb{R}^{n \times d}$ corresponds to data / feature / design /observation matrix. It has $n =$ *"number of samples"* rows and $d =$ *"dimension of datapoints"* columns.
- $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ corresponds to the output vector.
- for a vector $u \in \mathbb{R}^d$, we let $\|u\|_2 \coloneqq \sqrt{\sum_{i=1}^d u_i^2}$ denote the Euclidean norm of u (also called ℓ_2-norm).
- for vectors $u, v \in \mathbb{R}^d$, we denote $\langle u, v\rangle \coloneqq \sum_{i=1}^d u_i, v_i$ to be the inner product of u and v. Notice that $\|u\|^2 = \langle u, u\rangle$. Two vectors are said to be orthogonal if their inner product is null.
- for vectors $x_1, \dots, x_n$, their span corresponds to the linear space which they generate: $\mathrm{span}(x_1, \dots, x_n) \coloneqq \{\lambda_1 x_1 + \dots, \lambda_n x_n, \text{where } \lambda_1, \dots, \lambda_n \in \mathbb{R}\}$.
- the rank of a matrix corresponds to the dimension of the span of its columns, which is equal to the dimension of the span of its rows. Therefore for $X \in \mathbb{R}^{n \times d}$, it holds that $\mathrm{rank}(X) \leq \min(n, d)$. A matrix is said to be full rank if $\mathrm{rank}(X) = \min(n, d)$.
- the null space (sometimes called kernel) of a matrix $A \in \mathbb{R}^{n \times d}$ is defined as $\mathrm{Ker}(A) = \{w \in \mathbb{R}^d, Aw = 0\}$.
- a matrix $A \in \mathbb{R}^{d \times d}$ is said to be symmetric if $A^\top = A$. It is said to be positive semi-definite (we write this as $A \succeq 0$) if for all vector $u \in \mathbb{R}^d$, $u^\top A u \geq 0$. This is equivalent to saying that all the eigenvalues of A are positive.

A.2 Gradients

If $f : \mathbb{R}^d \to \mathbb{R}$ is twice differentiable, then its gradient $\nabla f : \mathbb{R}^d \to \mathbb{R}^d$ and hessian $\nabla^2 f : \mathbb{R}^d \to \mathbb{R}^{d \times d}$ are defined as:

$$\nabla f(w) = \Big(\frac{\partial f}{\partial w_i}(w)\Big)_{1 \leq i \leq d} \quad \nabla^2 f(w) = \Big(\frac{\partial^2 f}{\partial w_i \partial w_i}(w)\Big)_{1 \leq i,j \leq d}$$

- For a vector $b \in \mathbb{R}^d$, the gradient of the linear function $f : \mathbb{R}^d \to \mathbb{R}$, $f(w) = \langle a, w\rangle$ is equal to $\nabla f(w) = b$.

- For a (not necessarily symmetric) matrix $A \in \mathbb{R}^{d\times d}$, $f(w) = \frac{1}{2} w^\top A w$ is a quadratic function. Its gradient is $\nabla f(w) = \frac{1}{2}(A + A^\top)w$, which is equal to Aw if and only if A is a symmetric matrix. The hessian of f is equal to $\nabla^2 f(w) = \frac{1}{2}(A + A^\top)$.

A.3 Linear Algebra

- For a matrix $A \in \mathbb{R}^{d\times d}$, it holds that $w^\top A w = \sum_{i,j=1}^d w_i w_j A_{i,j}$.
- It holds that $X^\top X = \sum_{i=1}^n x_i x_i^\top \in \mathbb{R}^{d\times d}$ and $XX^\top = \left(\langle x_i, x_j\rangle\right)_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$.
- Let $L(w) = \frac{1}{2}\sum_{i=1}^n (y_i - \langle w, x_i\rangle)^2$, then $L(w) = \frac{1}{2}\|y - Xw\|^2$ and $\nabla L(w) = X^\top(Xw - y)$.
- if $n < d$ ("underparametrised setting), then the matrix $X^\top X$ cannot be invertible, because $\mathrm{span}(x_1, \dots, x_n)$ cannot be equal to $\mathbb{R}^d$. However, if $n \geq d$ and $\mathrm{span}(x_1, \dots, x_n) = \mathbb{R}^d$, then $X^\top X$ is invertible.
- *Rank–nullity theorem:* for $A \in \mathbb{R}^{n\times d}$, it holds that $\mathrm{rank}(A) + \dim \mathrm{Ker}(A) = d$.

A.4 Convexity

A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be convex if for all $w_1, w_2 \in \mathbb{R}^d$ and $\lambda \in [0, 1]$, $f(\lambda w_1 + (1-\lambda)w_2) \leq \lambda f(w_1) + (1-\lambda) f(w_2)$ (DO A DRAWING TO VISUALISE THIS!).

- if f is convex and differentiable, then for all $w_1, w_2 \in \mathbb{R}^d$, it holds that $f(w_2) \geq f(w_1) + \langle \nabla f(w_1), w_2 - w_1\rangle$ (DO A DRAWING TO VISUALISE THIS!).
- if f is convex, then all local minima are global. Therefore, if $\nabla f(w^\star) = 0$, then $w^\star$ is a global minimum. However, keep in mind that there exist convex functions which do not have any minima! (the exponential function for example)
- if a function f is convex, then its sublevel sets $\{w \in \mathbb{R}^d, f(w) \leq c\}$ are convex sets for all $c \in \mathbb{R}$. The converse is false (for example think of $x \mapsto \sqrt{|x|}$ in 1d).
- let $\lambda_1, \lambda_2 \geq 0$, if f_1 and f_2 are convex functions, then so is $\lambda_1 f_1 + \lambda_2 f_2$.
- $f : \mathbb{R}^n \to \mathbb{R}$ is convex and $A \in \mathbb{R}^{n\times d}$ and $b \in \mathbb{R}^n$, then $w \in \mathbb{R}^d \mapsto f(Aw + b)$ is convex too.

A.5 Good practices and sanity checks

Math sanity checks Always check that the math makes sense!

- If $f : \mathbb{R}^d \to \mathbb{R}$, then for a vector $w \in \mathbb{R}^d$, $\nabla f(w)$ must belong to $\mathbb{R}^d$! So for example if $f(w) = \frac{1}{2}\|w\|^2$, then writing that "$\nabla f(w) = \|w\|$" doesn't make sense.
- Check that the matrix operations are allowed: if $L(w) = \|y - Xw\|^2$, writing that "$\nabla L(w) = X(Xw - y)$" doesn't make sense because the operations XX and Xy don't make sense for $n \neq d$ (and also because of the remark right above).

Dimensional sanity check. Even if an expression is mathematically well-formed, it might be meaningless from the point of view of "units" or "dimensions." A quick check is to make sure your formulas are *homogeneous*: every term you add or compare should have the same "type."
Example. Suppose our data are temperature observations y_i measured in degrees Celsius. A feature vector $x \in \mathbb{R}^d$ could represent input quantities such as:

$$x = \big([\text{altitude in meters}], [\text{pressure in pascals}], [\text{wind speed in meters per second}]\big).$$

A parameter vector $w \in \mathbb{R}^d$ scales each feature so that the inner product $\langle w, x \rangle$ has the same unit as y (degrees Celsius). In our small example the units of w are

$$([{}^\circ\text{C}] \cdot [\text{meters}]^{-1}, [{}^\circ\text{C}] \cdot [\text{pascals}]^{-1}, [{}^\circ\text{C}] \cdot [\text{meters}]^{-1} \cdot [\text{seconds}])$$

Now notice that:

- $\langle w, x \rangle$ makes sense: each coordinate of w carries the reciprocal unit of the corresponding coordinate of x, so the sum yields something in degrees Celsius.
- $w + x$ does *not* make sense: w and x do not have the same units (adding "degrees per meter" to "meters" is meaningless).