Logistic regression and convex analysis

Alessandro Rudi, Pierre Gaillard

February 26, 2021

In this class, we will see logistic regression, a widely used classification algorithm. Contrary to linear regression, there is no closed-form solution and one needs to solve it thanks to iterative convex optimization algorithms. We will then see the basics of convex analysis.

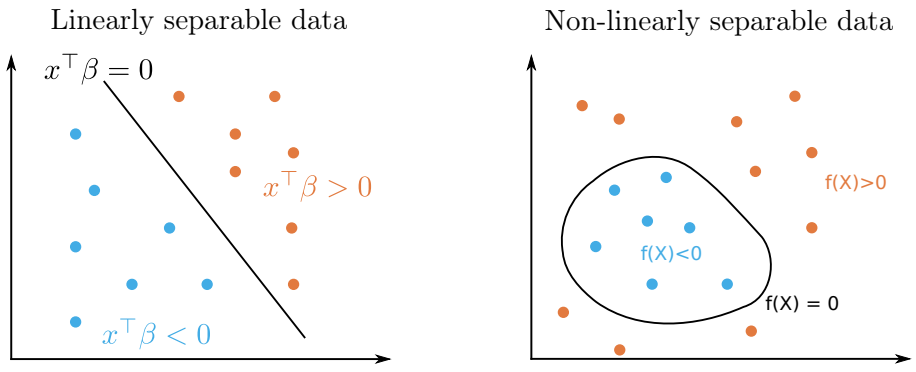
1 Logistic regression

We will consider the binary classification problem in which one wants to predict outputs in $\{0,1\}$ from inputs in $\mathbb{R}^d$. We consider a training set $D_n := \big\{(X_i, Y_i)\big\}_{1\leqslant i\leqslant n}$. The data points (X_i, Y_i) are i.i.d. random variables and follow a distribution $\mathcal{P}$ in $\mathcal{X}\times\mathcal{Y}$. Here, $\mathcal{X}=\mathbb{R}^d$ and $\mathcal{Y}=\{0,1\}$.

Goal We would like to use a similar algorithm to linear regression. However, since the outputs Y_i are binary and belong to $\{0,1\}$ we cannot predict them by linear transformation of the inputs X_i (which belong to $\mathbb{R}^d$). We will thus classify the data thanks to classification rules $f:\mathbb{R}^d\mapsto\mathbb{R}$ such that:

$$f(X_i)\begin{cases}\geqslant 0 & \Rightarrow \quad Y_i=+1\\ <0 & \Rightarrow \quad Y_i=0\end{cases},$$

to separate the data into two groups. In particular, we will consider linear functions f of the form $f_\beta : x\mapsto x^\top\beta$. This assumes that the data are well-explained by a linear separation (see figure below).



Of course, if the data does not seem to be linearly separable, we can use similar tricks that we mentioned for linear regression (polynomial regression, kernel regression, splines,...). We search a feature map $x\mapsto\phi(x)$ into a higher dimensional space in which the data are linearly separable. This will be the topic of the class on Kernel methods.

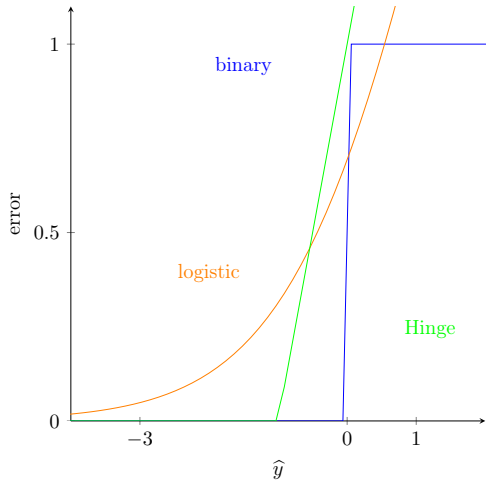


Figure 1: Binary, logistic and Hinge loss incured for a prediction $\widehat{y} := x^\top \beta$ when the true observation is $y = 0$.

Loss function To minimize the empirical risk, it remains to choose a loss function to assess the performance of a prediction. A natural loss is the *binary loss*: 1 if there is a mistake ($f(X_i) \neq Y_i$) and 0 otherwise. The empirical risk is then:

$$\widehat{\mathcal{R}}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \neq \mathbb{1}_{X_i^\top \beta \geqslant 0}} .$$

This loss function is however not convex neither in β. The minimization problem $\min_\beta \widehat{\mathcal{R}}_n(\beta)$ is extremely hard to solve. The idea of logistic regression consists in replacing the binary loss with another similar loss function which is convex in β. This is the case of the *Hinge loss* and of the logistic loss $\ell : \mathbb{R} \times \{0,1\} \to \mathbb{R}_+$. The latter assigns to a linear prediction $\widehat{y} = x^\top \beta$ and an observation $y \in \{0,1\}$ the loss

$$\ell(\widehat{y}, y) := y \log\left(1 + e^{-\widehat{y}}\right) + (1-y)\log\left(1 + e^{\widehat{y}}\right). \tag{1}$$

The binary loss, Hinge loss and logistic loss are plotted in Figure 1.

Definition 1.1 (Logistic regression estimator)**.** *The logistic regression estimator is the solution of the following minimization problem:*

$$\widehat{\beta}_{(logit)} = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell(X_i^\top \beta, Y_i) ,$$

where ℓ is the logistic loss defined in Equation (1)*.*

The advantage of the logistic loss with respect to the Hinge loss is that it has a probabilistic interpretation by modeling $\mathbb{P}(Y = 1|X)$, where (X, Y) is a couple of random variables following the law of (X_i, Y_i). We will see more on this in the lecture on Maximum Likelihood.

Computation of $\widehat{\beta}_{(\text{logit})}$ Similarly to OLS, we may try to analytically solve the minimization problem by canceling the gradient of the empirical risk. Since $\frac{\partial \ell(\widehat{y},y)}{\partial \widehat{y}} = \sigma(\widehat{y}) - y$, where $\sigma : z \mapsto$

$\frac{1}{1+e^{-z}}$ is the logistic function, we have:

$$\nabla \widehat{\mathcal{R}}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} X_i\big(\sigma(X_i^\top \beta) - Y_i\big) = \frac{1}{n} X\big(Y - \sigma(X\beta)\big)$$

where $X := (X_1, \dots, X_n)^\top$, $Y := (Y_1, \dots, Y_n)$, and $\sigma(X\beta)_i := \sigma(X_i^\top \beta)$ for $1 \leqslant i \leqslant n$. Bad news: the equation $\nabla \widehat{\mathcal{R}}_n(\beta) = 0$ has no closed-form solution. It needs to be solved through iterative algorithm (gradient descent, Newton's method,...). Fortunately, this is possible because the logistic loss is convex in its first argument. Indeed,

$$\frac{\partial^2 \ell(\widehat{y}, y)}{\partial \widehat{y}} = \sigma(\widehat{y})\sigma(-\widehat{y}) > 0\,.$$

The loss is strictly convex, the solution is thus unique. In this class and the next one, we will see tools and methods to solve convex optimization problems.

Regularization Similarly to linear regression, logistic regression may over-fit the data (especially when $p > n$). One needs then to add a regularization such as $\lambda\|\beta\|_2^2$ to the logistic loss.

2 Convex analysis

We will now see notions of convex analysis to solve convex optimization problems such as the one of logistic regression. For more details on this topic, we refer to the monograph Boyd and Vandenberghe, 2004. This class and the next one, we will see two aspects:

- convex analysis: properties of convex functions and convex optimization problems
- convex optimization: algorithms (gradient descent, Newton's method, stochastic gradient descent, ...)

Convexity is a crucial notion in many fields of mathematics and computer sciences. In machine learning, convexity allows to get well-defined problems with efficient solutions. A typical example is the problem of *empirical risk minimization*:

$$\widehat{f}_n \in \operatorname*{arg\,min}_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \ell\big(f(X_i), Y_i\big) + \lambda\Omega(f)\,, \qquad (*)$$

where $D_n = \big\{(X_i, Y_i)\big\}_{1 \leqslant i \leqslant n}$ is the data set, $\mathcal{F}$ is a convex set of predictors $f : \mathcal{X} \mapsto \mathbb{R}$, $\widehat{y} \mapsto \ell(\widehat{y}, y)$ are convex loss functions for all $y \in \mathcal{Y}$ and Ω is a convex penalty ($\|\cdot\|_2, \|\cdot\|_1, \dots$).

Convexity will be useful to analyze

- the statistical properties of the solution $\widehat{f}_n$ and its generalization error (i.e., its risk):

$$\mathcal{R}(\widehat{f}_n) := \mathbb{E}\big[\ell(f(X), Y)\big|D_n\big]$$

- get efficient algorithms to solve the minimization problem $(*)$ and find $\widehat{f}_n$.

2.1 Convex sets

In this class, we will only consider finite dimensional Euclidean space (typically $\mathbb{R}^d$).

Definition 2.1 (Convex set)**.** *A set $K \subseteq \mathbb{R}^d$ is convex if and only if for all $x, y \in K$ $[x, y] \subset K$ (or equivalently for all $\alpha \in (0, 1)$, $\alpha x + (1-\alpha)y \in K$).*

Example 2.1. *Here are a few examples of convex sets*

- *Hyperplans:* $K = \{x \in \mathbb{R}^d : a^\top x = b, a \neq 0, b \in \mathbb{R}\}$
- *Half spaces:* $K = \{x \in \mathbb{R}^d : a^\top x \geqslant b, a \neq 0, b \in \mathbb{R}\}$
- *Affine subspaces:* $K = \left\{x \in \mathbb{R}^d : Ax = b, A \in M_d(\mathbb{R}), b \in \mathbb{R}\right\}$
- *Balls:* $\|x\| \leqslant R$, $R > 0$
- *Cones:* $K = \{(x, r) \in \mathbb{R}^{d+1}, \|x\| \leqslant r\}$
- *Convex polytopes: intersections of half spaces.*

Properties of Convex sets :

- stability by intersection (not necessarily countable)
- stability by affine transformation
- convex separation: if C, D are disjoints convex sets ($C \cap D = \emptyset$), then there exists a hyperplane which separates C and D:

$$\exists a \neq 0, b \in \mathbb{R} \text{ such that } C \subset \left\{a^\top x \geqslant b\right\} \text{ and } D \subset \left\{a^\top x \leqslant b\right\}.$$

  The inequalities are strict if C and D are compact. Exercise: show this property when C and D are compact (clue: define $(x, y) \in \arg\min_{x \in C, y \in D} \|x - y\|$).

Definition 2.2 (Convex Hull)**.** *Let $A \in \mathbb{R}^d$. The convex hull, denoted* $\mathrm{Conv}(A)$*, of A is the smallest convex set that contains A. In other words:*

$$\mathrm{Conv}(A) = \bigcap \left\{B \subseteq \mathbb{R}^d : A \subseteq B \text{ and } B \text{convex}\right\}$$

$$\left\{x \in \mathbb{R}^d : \exists p \geqslant 1, \alpha \in \mathbb{R}_+^p, \sum_{i=1}^p \alpha_i = 1 \text{ and } z_1, \ldots, z_p \in A \text{such that } x = \sum_{i=1}^p \alpha_i z_i\right\}$$

2.2 Convex functions

Definition 2.3 (Convex function)**.** *A function $f : D \subset \mathbb{R}^d \to \mathbb{R}$ with D convex is*

- *convex iff for all $x, y \in D$ and $0 \leqslant \alpha \leqslant 1$,*

$$f(\alpha x + (1-\alpha)y) \leqslant \alpha f(x) + (1-\alpha) f(y)$$

- *strictly convex iff for all $x, y \in D$ and $0 \leqslant \alpha \leqslant 1$,*

$$f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha) f(y)$$

- *μ-strongly convex if there exists $\mu > 0$ such that*

$$f(\alpha x + (1-\alpha)y) \leqslant \alpha f(x) + (1-\alpha) f(y) - \frac{\mu}{2}\alpha(1-\alpha)\|x - y\|^2$$

Proposition 2.1. *f is μ-strongly convex if and only if $x \mapsto f(x) - \frac{\mu}{2}\|x\|^2$ is convex.*

A few examples of useful convex functions:

- dimension $d = 1$: x, x^2, $-\log x$, e^x, $\log(1 + e^{-x})$, $|x|^p$ for $p \geqslant 1$, $-x^p$ for $p < 1$ and $x \geqslant 0$
- higher dimension $d \geqslant 1$: linear functions $x \mapsto a^\top x$, quadratic functions $x \mapsto x^\top Q x$ for Q semidefinite (symmetric) positive matrix (i.e., all eigenvalues are nonnegative, or for all x $x^\top Q x \geqslant 0$), norms, $\max\{x_1, \ldots, x_d\}$, $\log\left(\sum_{i=1}^d e_i^x\right)$

Characterization of convex functions

- if f is C^1: f convex $\Leftrightarrow \forall x, y \in D$ $f(x) \geqslant f(y) + f'(y)(x - y)$
- if f is twice differentiable: f convex $\Leftrightarrow \forall x \in D$ its Hessian is semi-definite positive ($f''(x) \geqslant 0$)

Operations which preserve convexity

- supremum of a family $x \mapsto \sup_{i \in I} f_i(x)$
- linear combination with non-negative coefficients
- partial minimization: f convex on $C \times D \Rightarrow \quad y \mapsto \inf_{x \in C} f(x, y)$ is convex on D

Proposition 2.2. *If f is convex on D, then f is continuous on the interior of D. Furthermore, the epigraph of f $\big\{(x, t) \in D \times \mathbb{R}, f(x) \leqslant t\big\}$ is convex.*

Proposition 2.3 (Jensen's inequality)**.** *If f is convex. For all $x_1, \ldots, x_n \in D$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{R}_+$ such that $\sum_{i=1}^n \alpha_i = 1$ then*

$$f\left(\sum_{i=1}^n \alpha_i x_i\right) \leqslant \sum_{i=1}^n \alpha_i f(x_i)\,.$$

Jensen's inequality can be extended to infinite sums, integrals and expected values: if f is convex

- integral formulation: if $p(x) \geqslant 0$ on $S \subset D$ such that $\int_S p(x)dx = 1$ then

$$f\left(\int_S p(x) x dx\right) \leqslant \int_S p(x) f(x) dx\,.$$

- expected value formulation: if X is a random variable such that $X \in D$ almost surely and $\mathbb{E}[X]$ exists then if

$$f\big(\mathbb{E}\big[X\big]\big) \leqslant \mathbb{E}\big[f(X)\big]\,.$$

2.3 Unconstrained optimization problems

Let $f : \mathbb{R}^d \to \mathbb{R}$ convex and finite on $\mathbb{R}^d$. We consider the problem

$$\inf_{x \in \mathbb{R}^d} f(x)\,.$$

First, remark that we use the notation $\min_x f(x)$ only when the minimum is reached. If no point achieves the minimum, we use the notation $\inf_x f(x)$.

There are three possible cases

- $\inf_{x \in \mathbb{R}^d} f(x) = -\infty$: there is no minimum. For instance, $x \mapsto x$.
- $\inf_{x \in \mathbb{R}^d} f(x) > -\infty$ and the infimum is not reached. This is the case for instance for $x \mapsto \log(1 + e^{-x})$ or for $x \mapsto e^{-x}$.

– $\inf_{x \in \mathbb{R}^d} f(x) > -\infty$ and the minimum is reached and equals $\min_{x \in \mathbb{R}^d} f(x)$. This is the case for instance for coercive functions $\lim_{\|x\| \to \infty} f(x) = +\infty$.

Definition 2.4 (Local minimum)**.** *Let $f : D \to \mathbb{R}$ and $x \in D$. x is a local minimum if and only if there exists an open set $V \subset D$ such that $x \in V$ and $f(x) = \min_{x' \in V} f(x')$.*

Properties:

– f convex $\Rightarrow$ any local minimum is a global minimum.
– f strictly convex $\Rightarrow$ at most one minimum.
– f convex and C^1 then x is a minimum of f on $\mathbb{R}^d$ if and only if $f'(x) = 0$.

As we saw for linear regression and we will use in the next class, canceling the gradient provides an efficient solution to solve the minimization problem in closed form.

References

Boyd, S. and L. Vandenberghe (2004). *Convex optimization*. Cambridge university press.