Linear Models for Classification

We consider binary classification in dimension p, with a training set $\mathcal{D} = \{(\vec{x}^{\,i},\ y^i)\}_{i=1,\dots,n}$ consisting of n labeled observations $(\vec{x}^{\,i},\ y^i) \in \mathbb{R}^{p+1} \times \{0,1\}$.

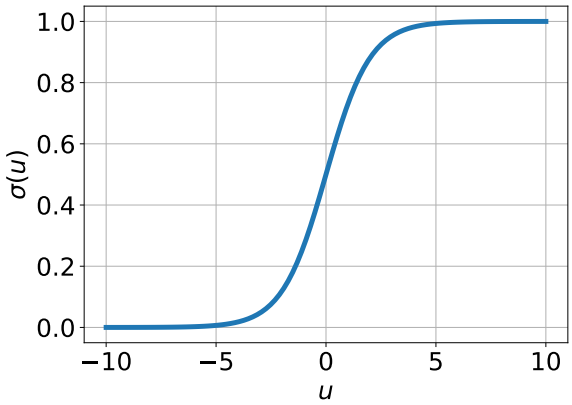
Here we take $\vec{x} \in \mathbb{R}^{p+1}$, after prepending a 1 to a p-dimensional vector, to simplify vector and matrix notation as mentionned in the course: $\beta_0 + \sum_{j=1}^{p} \beta_j x_j$ can then be written as $\langle \vec{\beta},\ \vec{x} \rangle$.

The **logistic function** (not to be confused with the *logistic loss function*) is defined by

$$\begin{aligned} \sigma : \mathbb{R} &\to [0,1] \\ u &\mapsto \frac{1}{1+e^{-u}}. \end{aligned}$$

Its graph is shown alongside.
This function is differentiable, and its derivative satisfies the following identity (which you can verify):

$$\sigma'(u) = \sigma(u)(1-\sigma(u)) \text{ for all } u \in \mathbb{R}. \qquad (1)$$



Graph of the logistic function.

1 Empirical risk minimization

1. Why is a linear parametric model, that is, a model of the form $f : \vec{x} \mapsto \langle \vec{\beta},\ \vec{x} \rangle$, inappropriate for a binary classification problem?

One could use a linear model as a *decision function*: $f(\vec{x}) \geq 0$ leads to predicting a positive label, and $f(\vec{x}) < 0$ leads to predicting a negative label. In **logistic regression**, we instead use the composition of a linear function and the logistic function as the decision function:

$$f(\vec{x}) = \sigma(\langle \vec{\beta},\ \vec{x} \rangle). \qquad (2)$$

2. How can $f(\vec{x})$ then be interpreted? Pay attention to the codomain of σ.

We now consider the hypothesis space $\mathcal{F} = \{f : \vec{x} \mapsto \sigma(\langle \vec{\beta},\ \vec{x} \rangle); \vec{\beta} \in \mathbb{R}^{p+1}\}$.

3. Use this hypothesis space and cross-entropy function

$$\begin{aligned} L_H : \{0,1\} \times (0,1) &\to \mathbb{R} \\ y, f(\vec{x}) &\mapsto -y \ln f(\vec{x}) - (1-y)\ln(1-f(\vec{x})) \end{aligned}$$

as the logistic loss function to formulate the training of a binary classifier as an empirical risk minimization problem.

4. Show or assume that this empirical risk is convex. How can it be minimized? You may show or assume that the gradient of the empirical risk at $\vec{\beta}$ is

$$\nabla_{\vec{\beta}}\, R_n(\vec{\beta}) = -\frac{1}{n}\sum_{i=1}^{n}\left(y^i - \frac{1}{1+e^{-\langle \vec{\beta},\ \vec{x}^{\,i} \rangle}}\right)\vec{x}^{\,i}.$$

Hint: set $\sigma_i = \sigma(\langle \vec{\beta},\ \vec{x}^{\,i} \rangle)$ and start by expressing $\nabla_{\vec{\beta}}\sigma_i$ in terms of $\vec{x}^{\,i}$ and σ_i.