MA2823 : Foundations of Machine Learning
Chapter 5 : Linear Regression

Lecturer : Chloé-Agathe Azencott Scribes : Emilie Leblanc, Antoine Salon, Hippolyte Jacomet

In this chapter we will see how to :

— define parametric methods ;

— define the maximum likelihood estimator and compute it for Bernoulli, multinomial and Gaussian densities ;

— define the Bayes estimator and compute it for normal priors ;

— compute the maximum likelihood estimator / least-square fit solution for linear regression ;

— compute the maximum likelihood estimator for logistic regression.

1 Parametric methods

In this chapter we would like to define different ways of building estimators that are part of a statistical model, and how to assess their performances.

Definitions and notations Let us consider a training set of n p-dimensional vectors denoted by $\mathcal{X} = \{x^i\}_{i=1,\dots,n}$, and let's assume that each of these observation follows a law of probability given a parameter θ.

$$\forall i \in [1, \dots, n],\ \boldsymbol{x}^i \sim p(\boldsymbol{x}|\theta)$$

The *parametric estimation* consists in *assuming* a form for $p(\boldsymbol{x}|\theta)$ and then *estimating* the parameter θ using our training set $\mathcal{X}$.
For example, if we assume that the observations of our training set follow a Gaussian distribution $p(x_j|\theta_j) \sim \mathcal{N}(\mu_j, {\sigma_j}^2)$ our goal would be to estimate the parameter $\theta = \{\mu_1, \sigma_1, \dots, \mu_p, \sigma_p\}$.

To achieve that goal, the observations of the training set $\boldsymbol{x}^i$ are usually assumed to be *independent and identically distributed* (i.i.d.).

In the following sections we sill study two different methods to estimate the parameters : the *maximum likelihood estimation* and the *Bayes estimation*.

1.1 Maximum likelihood estimation

This first method relies on the idea that θ should maximize the likelihood for $\mathcal{X}$ to be drawn from the law of probability we chose in our model.

Likelihood and log-likelihood We call *likelihood of* θ given $\mathcal{X}$ the function : $\ell(\cdot|\mathcal{X}) : \theta \mapsto p(\mathcal{X}|\theta)$.
In the case of an i.i.d. sample $\mathcal{X}$, this becomes :

$$\ell(\theta|\mathcal{X}) = p(\mathcal{X}|\theta) = p(\boldsymbol{x}^1|\theta)p(\boldsymbol{x}^2|\theta)\dots p(\boldsymbol{x}^p|\theta)$$

It is often useful to also consider the *log-likelihood*, simply defined by $\mathcal{L}(\cdot|\mathcal{X}) : \theta \mapsto \log \ell(\theta|\mathcal{X})$, which in our previously mentioned case becomes :

$$\mathcal{L}(\theta|\mathcal{X}) = \log p(\boldsymbol{x}^1|\theta) + \ldots + \log p(\boldsymbol{x}^n|\theta)$$

Given these two functions, we will then try to find the estimator that maximizes them, called the *maximum likelihood estimator* or MLE, and defined by :

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta|\mathcal{X}) = \arg\max_{\theta} \mathcal{L}(\theta|\mathcal{X})$$

Let us now look at these MLE for Bernoulli, Multinomial and Normal laws.

1.1.1 Bernoulli density

Such a law corresponds to observations that reflect two states of either failure or success (think of a coin toss for example), and thus let us consider a data set $\mathcal{X} = \{x^i\}_{i=1,\ldots,n}$ with $x^i \in \{0, 1\}$ $\forall i \in [1, \ldots, n]$.

The Bernoulli density is given by $P(X = x|p_0) = {p_0}^x(1 - p_0)^{(1-x)}$, let us try to find the MLE $\hat{p_0}$ of the parameter p_0.

We first compute the log-likelihood :

$$\mathcal{L}(p_0|\mathcal{X}) = \log P(\mathcal{X}|p_0) = \sum_{i=1}^{n} \left(x^i \log p_0 + (1 - x^i) \log (1 - p_0)\right)$$

This is a concave function which we can easily maximize by setting its gradient to 0 :

$$\frac{\sum_{i=1}^n x^i}{\hat{p_0}} - \frac{n}{(1 - \hat{p_0})} + \frac{\sum_{i=1}^n x^i}{(1 - \hat{p_0})} = 0$$

Which yields the MLE of p_0 :

$$\hat{p_0} = \frac{1}{n} \sum_{i=1}^{n} x^i$$

1.1.2 Multinomial density

Let us consider K mutually exclusive and exhaustive classes. The probability for each class to occur is p_k, with $\sum_{k=1}^K p_k = 1$. We represent these classes with K indicator variables $x_1, x_2, \ldots, x_K$ knowing

$$x_k = \begin{cases} 1 & \text{if the outcome is class } k \\ 0 & \text{otherwise} \end{cases}$$

The joint probability distribution is given by :

$$P(x_1, x_2, \ldots, x_K) = \prod_{k=1}^{K} p_k^{x_k}$$

Let us compute the MLE $\hat{\boldsymbol{p}} = (\hat{p_1}, \dots, \hat{p_K})$. The observations are i.i.d. thus the log-likelihood looks dramatically like :

$$\mathcal{L}(\boldsymbol{p}|\mathcal{X}) = \sum_{i=1}^{n} \sum_{k=1}^{K} x_k^i \log p_k$$

And we now have to maximize it under the constraint $\sum_{k=1}^{K} p_k = 1$.
We must call on to a Lagrange multiplier λ and define the Lagrangian function

$$\mathcal{F} : (\boldsymbol{p}, \lambda) \mapsto \mathcal{L}(\boldsymbol{p}|\mathcal{X}) - \lambda \left(\sum_{k=1}^{K} p_k - 1 \right)$$

Setting its gradient to zero gently yields a system of equations from which we are happy to learn that $\lambda = n$ and consequently that

$$\hat{p}_k = \frac{1}{n} \sum_{i=1}^{n} x_k^i \qquad \forall k \in [1, \dots, K]$$

1.1.3 Gaussian distribution

Let us now assume that our observations x^i are i.i.d. following a Gaussian distribution, i.e. :
$\forall i \in [1, \dots, n]$,

$$x^i \sim \mathcal{N}(\mu, \sigma^2)$$

$$p(x^i|\mu, \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x^i - \mu)^2}{2\sigma^2} \right]$$

Setting the gradient of the log-likelihood to zero yields the estimators :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x^i$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x^i - \hat{\mu})^2$$

1.2 Bias-variance trade-off

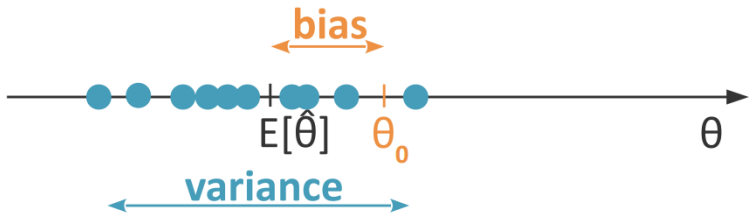
We can assess the performance of the estimator (be it obtain through MLE or another method) by computing the *mean squared error* or MLE, defined by :

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta_0)^2]$$

Bias-variance trade-off Let us try to link this MSE to two previously known notions that help characterize our estimator, the variance and the bias. As a reminder, the *bias* of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated : $\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta_0$.

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= \mathbb{E}[(\hat{\theta} - \theta_0)^2] \\
&= \mathbb{E}\left[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2 + 2\hat{\theta}\mathbb{E}[\hat{\theta}] - \mathbb{E}[\hat{\theta}]^2 - 2\hat{\theta}\theta_0 + \theta_0^2\right] \\
&= \text{Var}(\hat{\theta}) + \mathbb{E}[\hat{\theta}]^2 - 2\mathbb{E}[\hat{\theta}]\theta_0 + \theta_0^2 \\
&= \text{Var}(\hat{\theta}) + \text{Bias}^2(\hat{\theta})
\end{aligned}$$

We see that the MSE is a balance between these two positive expressions, hence the idea of a *bias-variance trade-off*. We understand that a biased estimator may achieve better MSE than an unbiased one, as illustrated below :



1.3 Bayes estimator

From now on let us study another way of building an estimator, using Bayes rule which in its general form is :

$$P(C|\boldsymbol{x}) = \frac{P(C)p(\boldsymbol{x}|C)}{p(\boldsymbol{x})}$$

Here we choose to treat θ as a random variable with a given prior distribution $p(\theta)$. Considering an observation-set $\mathcal{X}$, the Bayes rule yields :

$$p(\theta|\mathcal{X}) = \frac{p(x|\theta)p(\theta)}{p(\mathcal{X})}$$

We define the *Bayes estimate* as the conditional expected value of θ given our data set $\mathcal{X}$:

$$\hat{\theta}_{\text{Bayes}} = \mathbb{E}[\theta|\mathcal{X}] = \int \theta p(\theta|\mathcal{X}) d\theta$$

Here's a reminder of the estimates we have previously seen :

Maximum a posteriori estimate :

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{X})$$

Maximum likelihood estimate :

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} p(\mathcal{X}|\theta)$$

1.3.1 Bayes estimator for normal priors

Let us consider n i.i.d. data points x^i following a normal law, $x^i \sim \mathcal{N}(\theta, \sigma^2)$ and let's assume that θ is of Gaussian prior distribution : $\theta \sim \mathcal{N}(\mu, \sigma_0^2)$.

We know that the MLE of θ is its sample mean : $\hat{\theta}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n} x^i$ and we would like to compare it to its Bayes estimator.

After a long and troublesome computation, a quick look at $p(\theta|\mathcal{X})$ shows us that it follows a normal distribution with mean m and variance s^2 :

$$m = \frac{n\hat{\theta}_{\text{MLE}}\sigma^2 + \mu\sigma_0^2}{n\sigma^2 + \sigma_0^2}$$

$$s^2 = \frac{\sigma^2\sigma_0^2}{n\sigma^2 + \sigma_0^2}$$

Knowing that :

$$\mathbb{E}[\theta|\mathcal{X}] = m$$

We can conclude with :

$$\hat{\theta}_{\text{Bayes}} = \frac{\frac{n}{\sigma_0^2}}{\frac{n}{\sigma_0^2} + \frac{1}{\sigma^2}}\hat{\theta}_{\text{MLE}} + \frac{\frac{1}{\sigma^2}}{\frac{n}{\sigma_0^2} + \frac{1}{\sigma^2}}\mu$$

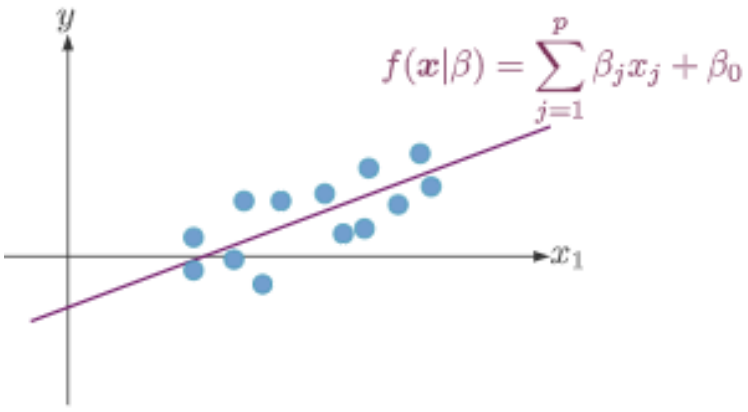
It is interesting to notice that when n increases, $\hat{\theta}_{\text{Bayes}}$ gets closer to the sample average (uses information from the sample) and When σ is small, $\hat{\theta}_{\text{Bayes}}$ gets closer to μ (little uncertainty about the prior).

2 Linear regression

The goal of a *linear regression* is to approximate the law followed by our data y by a linear combination of observed variables $\boldsymbol{x}$ such as :

$$f(\boldsymbol{x}|\beta) = \sum_{j=1}^{p} \beta_j * x_j + \beta_0$$

with $\boldsymbol{x}^i \in \mathbb{R}^p$ and $y^i \in \mathbb{R}$ and noting $\mathcal{D} = \{\boldsymbol{x}^i, y^i\}_{i=1,..,n}$.

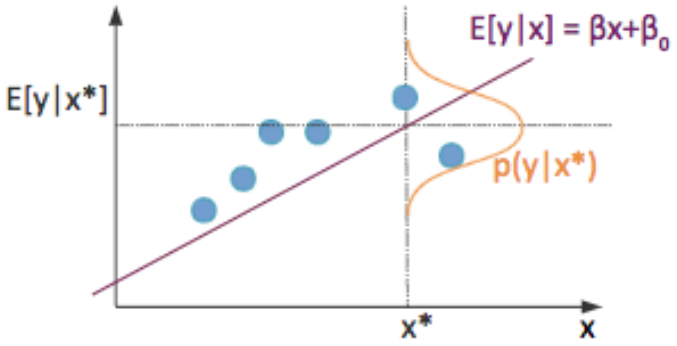


We assume that the error ϵ we have while doing a regression is **Gaussian distributed**. Therefore, with $y = g(\boldsymbol{x}) + \epsilon$, $f(\boldsymbol{x}|\beta)$ being g's estimator, we have :

$$\epsilon = y - g(\boldsymbol{x}) \text{ and } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

We therefore have our $\mathcal{D} = \{\boldsymbol{x}^i, \boldsymbol{y}^i\}_{i=1,..,n}$ observations (the blue dots), the linear regression obtained $\mathbb{E}(y|x) = \beta * x + \beta_0$ and, for each new x^* point, we can compute the hypothetical $y^* = \mathbb{E}(y|x^*)$ value and the corresponding error ϵ^* and confidence interval given by $p(y|x^*)$ since :

$$p(y|\boldsymbol{x}) \sim \mathcal{N}(f(\boldsymbol{x}|\beta), \sigma^2)$$



2.1 Maximum likelihood estimation under Gaussian noise

We are now going to use the *maximum likelihood estimation* in order to find the best linear regression possible to fit the data and predict new points.

Under Gaussian noise, considering i.i.d. observations and with $p(y|\boldsymbol{x}) \sim \mathcal{N}(f(\boldsymbol{x}|\beta), \sigma^2)$, the log-likelihood defined earlier becomes :

$$\begin{aligned}
\mathcal{L}(\beta|\mathcal{D}) &= \log \prod_{i=1}^{n} p(y^i|\boldsymbol{x}^i) + \log \prod_{i=1}^{n} p(\boldsymbol{x}^i) \\
&= \log \left(\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} * exp\left[-\frac{(y^i - f(\boldsymbol{x}^i|\beta))^2}{2\sigma^2} \right] \right) + Cte \\
&= Cte - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y^i - f(\boldsymbol{x}^i|\beta))^2
\end{aligned}$$

Since $\log \prod_{i=1}^{n} p(\boldsymbol{x}^i)$ is independent of β, its value is added to a term Cte along with others.
Assuming Gaussian error, maximizing the likelihood $\mathcal{L}(\beta|\mathcal{D})$ is therefore equivalent to minimizing the sum of squared residuals $\sum_{i=1}^{n} (y^i - f(\boldsymbol{x}^i|\beta))^2$. This expression of the residual sum of squares can also be written in matrix form :

$$\begin{aligned}
RSS(\beta) &= \sum_{i=1}^{n} (y^i - f(\boldsymbol{x}^i))^2 \\
&= \sum_{i=1}^{n} \left(y^i - \beta_0 - \sum_{j=1}^{p} x_j^i \beta_j \right)^2 \\
&= (y - \mathbf{X}\beta)^T * (y - \mathbf{X}\beta)
\end{aligned}$$

with $\boldsymbol{X} = \begin{pmatrix} 1 & x_1^1 & x_2^1 & ... & x_p^1 \\ 1 & x_1^2 & x_2^2 & ... & x_p^2 \\ . & . & . & ... & . \\ . & . & . & ... & . \\ . & . & . & ... & . \\ 1 & x_1^n & x_2^n & ... & x_p^n \end{pmatrix}$. Here, we added a vector $\begin{pmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{pmatrix}$ to $\{\boldsymbol{x}^i\}_{i=1,..,n}$ in $\boldsymbol{X}$'s expression in order to compute the scalar β_0.

Historically, the use of the residual sum of squares minimization can be attributed to Carl Friedrich Gauss (to predict the location of Ceres) and Adrien Marie Legendre.

In this method, under which conditions is β's estimator unique ?

— In order to minimize $RSS(\beta)$, $\boldsymbol{X}$ must have a full column rank, hence $\boldsymbol{X}^T\boldsymbol{X}$ *invertible*. In this case, the β minimizing $RSS(\beta)$ is :

$$\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y$$

— If $\boldsymbol{X}^T\boldsymbol{X}$ is not invertible (or *rank-deficient*), we can use its pseudo-inverse. However, when doing so, we must keep in mind that the solution is *not* unique.
A pseudo-inverse of **A** is a matrix **G** such as **AGA = A**.

2.2 Gauss-Markov Theorem

Gauss-Markov Theorem : Under the assumption that $\epsilon \sim \mathcal{N}(0, \sigma^2)$, the least-squares estimator of β is its best linear unbiased estimator. This Best Linear Unbiased Estimator is unique if $\boldsymbol{X}^T\boldsymbol{X}$ is invertible. We call it the "*BLUE*".

Indeed, with $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y$, we can show that (demonstration of the Gauss-Markov Theorem) :

$\forall \beta^*$ unbiased estimator of β, $\mathbb{V}ar(\beta^*) \geq \mathbb{V}ar(\hat{\beta})$ and when $\mathbb{V}ar(\beta^*) = \mathbb{V}ar(\hat{\beta}) => \beta^* = \hat{\beta}$

Proof of the Gauss-Markov theorem :
We have $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y$ *BLUE* estimator and let $\beta^* = Ay$ be another linear estimator of β. As we're restricting to unbiased estimators, minimum mean squared error implies minimum variance. The goal is therefore to show that such an estimator has a variance no smaller than that of $\hat{\beta}$. We have $A = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + D$ where D is a non-zero matrix.
In order to have β^* unbiased, we compute:

$$\begin{aligned} \mathbb{E}[\beta^*] &= \mathbb{E}[Ay] \\ &= \mathbb{E}[((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + D)(\boldsymbol{X}\beta + \epsilon)] \\ &= ((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + D)\boldsymbol{X}\beta + ((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + D)\mathbb{E}[\epsilon] \\ &= ((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + D)\boldsymbol{X}\beta \\ &= (\boldsymbol{I} + D\boldsymbol{X})\beta \end{aligned}$$

Since $\mathbb{E}(\epsilon) = 0$. Therefore, in order to have an unbiased β^* estimator, we must have $DX = 0$.
Now, we calculate and compare the variances :

$$\begin{aligned} \mathbb{V}ar(\hat{\beta}) &= \mathbb{E}[((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\epsilon\epsilon^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T * \sigma\boldsymbol{I} * \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \end{aligned}$$

$$
\begin{aligned}
\mathbb{V}ar(\beta^*) &= \mathbb{V}ar(Ay) \\
&= A\mathbb{V}ar(y)A^T \\
&= \sigma^2 * AA^T \\
&= \sigma^2 * ((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + D)(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + D)^T \\
&= \sigma^2[(\boldsymbol{X}^T\boldsymbol{X})^{-1} + (\boldsymbol{X}^T\boldsymbol{X})^{-1}(D\boldsymbol{X})^T + D(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} + DD^T] \\
&= \sigma^2\boldsymbol{X}^T\boldsymbol{X})^{-1} + \sigma^2 DD^T \\
&= \mathbb{V}ar(\hat{\beta}) + \sigma^2 DD^T
\end{aligned}
$$

since $DX = 0$.
Therefore $\mathbb{V}ar(\beta^*) = \sigma^2 DD^T + \mathbb{V}ar(\hat{\beta})$ and since $\sigma^2 DD^T$ is positive semi-defined and minimal for $D = 0$, $\mathbb{V}ar(\beta^*)$ exceeds $\mathbb{V}ar(\hat{\beta})$.

2.3 Interpretation with correlated or uncorrelated variables

The interpretation of the coefficients of a linear regression can be tricky. Indeed, once we have found $\beta_0, ..., \beta_p$ such as :

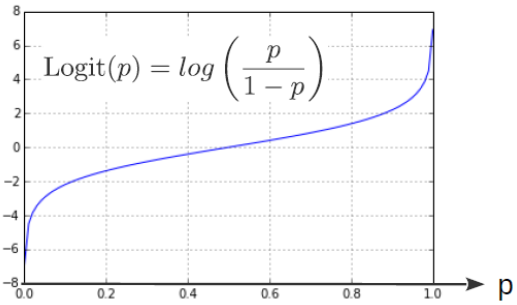
$$f(\boldsymbol{X}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

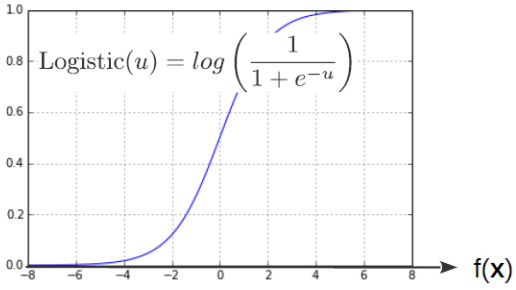
we first have to study the correlation of the variables.

- If the variables are *decorrelated*, each coefficient can be estimated separately and the interpretation is thus easy : "A change in 1 in x_j is associated with a change of β_j in Y, while everything else stays the same."
- This interpretation is no longer true if the variables are *correlated*. The correlations between variables cause problems : the variance of all coefficients tend to increase and the interpretation is much harder (when x_j changes, so does everything else).

3 Logistic regression

What about classification ? There are different conceptual issues with linear regression that make it unsuitable for classification. The main one is that we can't model the probability $P(y = 1|\boldsymbol{x})$ as a linear function of $\boldsymbol{x}$ because it must be between 0 and 1. Moreover, this cannot model "diminishing returns" where the impact of a change in $\boldsymbol{x}$ is not constant along the probability range. Indeed, if $P(y = 1|\boldsymbol{x})$ is close to 0 or +1, $\boldsymbol{x}$ must change a lot for y to change and this is not the case when $P(y = 1|\boldsymbol{x})$ is close to 0.5.
Hence we use a **logit transformation** through a *logistic regression*.





$$\log \frac{P(y = 1|\boldsymbol{x})}{1 - P(y = 1|\boldsymbol{x})} = \beta^T\boldsymbol{x} + \beta_0 \text{ with } \beta = (\beta_0, ..., \beta_p)$$

Maximum likelihood estimation of logistic regression coefficients As usual, let's compute the log likelihood for this logistic regression, knowing our data $\mathcal{D} = \{\boldsymbol{x}^i, y^i\}_{i=1,..,n}$:

$$\mathcal{L}(\beta|\mathcal{D}) = \sum_{i=1}^{n} logP(y^i|\boldsymbol{x}^i) + Cte = \sum_{i=1}^{n} (y^i \log g^i + (1-y^i)\log(1-g^i))$$

$$\text{with } g = P(y=1|\boldsymbol{x}) = \frac{1}{1+e^{-\beta^T \boldsymbol{x}}}$$

We find the maximum of the log-likelihood by calculating its gradient $\nabla_\beta \mathcal{L}$:

$$\nabla_\beta g^i = \boldsymbol{x}^i g^i (1-g^i)$$

$$\nabla_\beta \mathcal{L} = \sum_{i=1}^{n} (y^i - g^i)\boldsymbol{x}^i$$

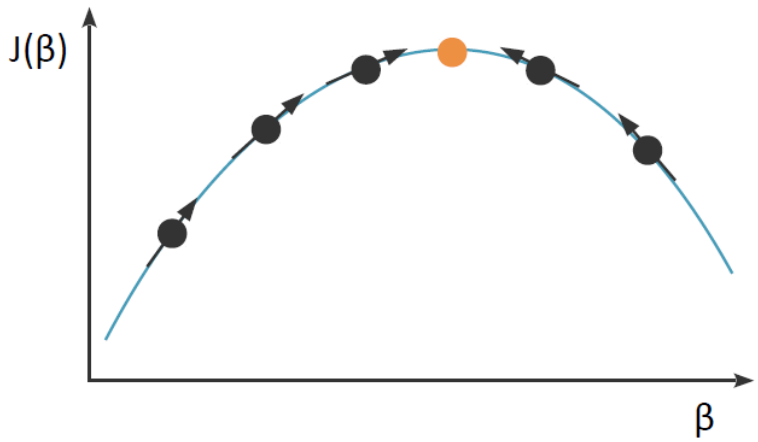
And solving for zero in the gradient formula :

$$\sum_{i=1}^{n} \left(y^i - \frac{1}{1+e^{-\beta^T \boldsymbol{x}^i}} \right) = 0$$

However, this expression cannot be solved analytically.
Since $\mathcal{L}$ is concave, hence there is no local minima, we can use the gradient ascent method.

Gradient ascent method For a function J concave in β :

- Update rule : $\beta^{(t+1)} \leftarrow \beta^{(t)} + \eta \nabla_\beta J(\beta^{(t)})$
- Iterate until change is inferior to a chosen margin ϵ
- η is the learning rate



Other methods remain possible such as the Newton method, conjugate gradient ascent or IRLS.

4 Summary

The main points to keep in mind from this lesson are the following :

- MAP estimate : $\hat{\theta}_{\text{MAP}} = \arg\max_{\theta} p(\theta|\mathcal{X})$
- MLE : $\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} p(\mathcal{X}|\theta)$
- Bayes estimate : $\hat{\theta}_{\text{Bayes}} = \mathbb{E}[\theta|\mathcal{X}] = \int \theta p(\theta|\mathcal{X}) d\theta$
- Assuming a Gaussian error, maximizing the likelihood (MLE) is equivalent to minimizing the RSS (residual sum of squares).
- Linear regression MLE : $\hat{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T y$
- Logistic regression : to solve with gradient ascent.