Logistic regression

1 Warm-up

The *logistic model* assumes that the random variables $(X, Y) \in \mathbb{R}^d \times \{0, 1\}$ are such that

$$\mathbb{P}(Y = 1|X) = \frac{\exp(\langle \beta^*, X \rangle)}{1 + \exp(\langle \beta^*, X \rangle)},$$

with $\beta^* \in \mathbb{R}^d$. In this case, $\mathbb{P}(Y = 1|X) > 1/2$ if and only if $\langle \beta^*, X \rangle > 0$, so the frontier between $\{x\,;\,h_*(x) = 1\}$ and $\{x\,;\,h_*(x) = 0\}$ is an hyperplane, with orthogonal direction β^*.

1. In this question only, $\beta^* = (\beta_0, \beta_1) \in \mathbb{R} \times \mathbb{R}_*$ and $X_i = (1, x_i)$ for all $1 \leqslant i \leqslant n$.

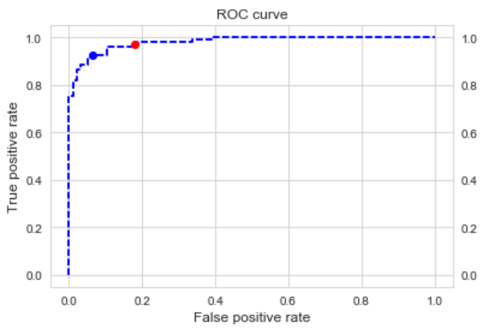
   (a) Provide the value x_* of x_i such that $\mathbb{P}(Y_i = 1|X_i) = 1/2$.

   By definition, $\mathbb{P}(Y_i = 1|X_i) = 1/2$ *if and only if* $\beta_0 + \beta_1 x_i = 0$ *i.e. if* $x_i = -\beta_0/\beta_1$.

   (b) Another classifier could be defined by choosing a threshold $\tilde{p} \in (0, 1)$ and defining $\tilde{h}(X_i) = 1$ if and only if $\mathbb{P}(Y_i = 1|X_i) > \tilde{p}$. Provide $\tilde{x}$ such that $\mathbb{P}(Y_i = 1|X_i) = \tilde{p}$. Explain a practical interest to choose $\tilde{p} < 1/2$.

   By definition, $\mathbb{P}(Y_i = 1|X_i) = \tilde{p}$ *if and only if* $(1 - \tilde{p})\mathrm{e}^{\beta_0 + \beta_1 x_i} = \tilde{p}$ *i.e. if* $\beta_0 + \beta_1 x_i = \log(\tilde{p}/(1 - \tilde{p}))$.

2. The usual logistic regression classifier is defined by $h_n : x \mapsto 1$ is $x^\top \hat{\beta}_n > 0$ and 0 otherwise, where $\hat{\beta}_n$ is an estimator of β. Therefore $h_n(X) = 1$ if and only if $\mathbb{P}(Y = 1|X) > 1/2$. Other classifiers can be defined by setting $h_n(X) = 1$ if and only if $\mathbb{P}(Y = 1|X) > p_*$ for a chosen $p_* \in (0, 1)$. Two classifiers were built with $p_* = 0.5$ and $p_* = 0.2$, associate each classifier with its point on ROC curve displayed above.



The red dot corresponds to $p_* = 0.2$ *as decreasing* p_* *leads to more individual classified in group 1 which can only increase the true positive rate and the false positive rate.*

2 Softmax regression

Assume that the observation Y takes values in $\{1,\ldots,M\}$ and that $X \in \mathbb{R}^d$. The negative loglikelihood to be minimized to estimate the parameters of the model is given by:

$$\theta \mapsto \ell_n^{\mathrm{multi}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{M} \mathbb{1}_{Y_i=k} \log \mathbb{P}_\theta(Y_i = k|X_i)\,,$$

where $\{(X_i, Y_i)\}_{1\leqslant i\leqslant n}$ are i.i.d. observations with the same law as (X, Y).

1. Explain the construction of $\mathbb{P}_\theta(Y_i = k|X_i)$, $1 \leqslant i \leqslant n$ for a softmax regression model with parameters $\omega_m \in \mathbb{R}^d$ for $1 \leqslant m \leqslant M$. In this case, $\theta = \{\omega_1, \ldots, \omega_M\}$.

   In a softmax regression setting, we assume, for $1 \leqslant m \leqslant M$ and $1 \leqslant i \leqslant n$, that

   $$\mathbb{P}_\theta(Y_i = k|X_i) = \frac{\mathrm{e}^{\omega_k^\top X_i}}{\sum_{\ell=1}^{n} \mathrm{e}^{\omega_\ell^\top X_i}}\,.$$

2. In the setting of the softmax regression function, compute $\theta \mapsto \nabla_\theta \ell_n^{\mathrm{multi}}(\theta)$.

   It is enough to compute the partial derivative of $\theta \mapsto \log \mathbb{P}_\theta(Y_i = k|X_i)$ with respect to each ω_j, $1 \leqslant j \leqslant M$. For all $1 \leqslant k \leqslant M$,

   $$\log \mathbb{P}_\theta(Y_i = k|X_i) = \omega_k^\top X_i - \log\left(\sum_{\ell=1}^{n} \mathrm{e}^{\omega_\ell^\top X_i}\right)\,.$$

   Therefore, for all $1 \leqslant j \leqslant M$,

   $$\partial_{\omega_j} \log \mathbb{P}_\theta(Y_i = k|X_i) = X_i \mathbb{1}_{j=k} - \frac{\mathrm{e}^{\omega_j^\top X_i}}{\sum_{\ell=1}^{n} \mathrm{e}^{\omega_\ell^\top X_i}} X_i\,.$$

3 Maximum likelihood estimation

The unknown parameter β^* may be estimated by maximizing the conditional likelihood of the observations given the input data:

$$\widehat{\beta}_n \in \mathrm{argmax}_{\beta\in\mathbb{R}^d} \prod_{i=1}^{n}\left[\left(\frac{\exp\left(\langle \beta,_i \rangle\right)}{1+\exp\left(\langle \beta, x_i\rangle\right)}\right)^{Y_i}\left(\frac{1}{1+\exp\left(\langle \beta, x_i\rangle\right)}\right)^{1-Y_i}\right],$$

to define the empirical classifier

$$\widehat{h}_n : x \mapsto \mathbb{1}_{\langle \widehat{\beta}_n, x\rangle > 0}\,.$$

1. Compute the gradient and the Hessian H_n of

   $$\ell_n : \beta \mapsto -\sum_{i=1}^{n}\left[Y_i\langle x_i, \beta\rangle - \log(1 + \exp(\langle x_i, \beta\rangle))\right]\,.$$

   What can be said about the function ℓ_n when for all $\beta \in \mathbb{R}^d$, $H_n(\beta)$ is nonsingular? This assumption is supposed to hold in the following questions.

   Since for all $u \in \mathbb{R}^d$, $\nabla_\beta\langle u, \beta\rangle = u$,

   $$\nabla \ell_n(\beta) = -\sum_{i=1}^{n} Y_i x_i + \sum_{i=1}^{n}\frac{\exp(\langle x_i, \beta\rangle)}{1 + \exp(\langle x_i, \beta\rangle)} x_i\,.$$

On the other hand, for all $1 \leqslant i \leqslant n$ and all $1 \leqslant j \leqslant d$,

$$\partial_j \left(\frac{\exp(\langle x_i, \beta \rangle)}{1+\exp(\langle x_i, \beta \rangle)} x_i \right) = \frac{\exp(\langle x_i, \beta \rangle)}{(1+\exp(\langle x_i, \beta \rangle))^2} x_{ij} x_i \,,$$

where x_{ij} is the jth component of x_i. Then

$$\left(H_n(\beta)\right)_{\ell j} = \sum_{i=1}^n \frac{\exp(\langle x_i, \beta \rangle)}{(1+\exp(\langle x_i, \beta \rangle))^2} x_{ij} x_{i\ell} \,,$$

that is,

$$H_n(\beta) = \sum_{i=1}^n \frac{\exp(\langle x_i, \beta \rangle)}{(1+\exp(\langle x_i, \beta \rangle))^2} x_i x_i^\top \,.$$

$H_n(\beta)$ is a semi positive definite matrix, which implies that $\beta \mapsto \ell_n(\beta)$ is convex. If we assume that H_n is nonsingular, ℓ_n is strictly convex.

2. Prove that there exists $\widetilde{\beta}_n \in \mathbb{R}^d$ such that $\|\widetilde{\beta}_n - \beta^*\| \leq \|\widehat{\beta}_n - \beta^*\|$ and

$$\widehat{\beta}_n - \beta^* = -H_n(\widetilde{\beta}_n)^{-1} \nabla \ell_n(\beta^*) \,.$$

Using a Taylor expansion between $\beta^\star$ and $\widehat{\beta}_n$, there exists $\tilde{\beta}_n \in B(\beta^\star, \|\widehat{\beta}_n - \beta^\star\|)$ such that

$$\nabla \ell_n(\widehat{\beta}_n) = \nabla \ell_n(\beta^\star) + H_n(\tilde{\beta}_n)(\hat{\beta}_n - \beta^\star) \,.$$

By definition, $\nabla \ell_n(\widehat{\beta}_n) = 0$. Therefore,

$$\widehat{\beta}_n - \beta^\star = -H_n(\tilde{\beta}_n)^{-1} \nabla \ell_n(\beta^\star) \,,$$

where $H_n(\tilde{\beta}_n)^{-1}$ exists since $H_n(\beta)$ is assumed to be non-singular for all β.

In the following it is assumed that the $(x_i)_{1 \leqslant i \leqslant n}$ are uniformly bounded, $\widehat{\beta}_n \to \beta^*$ a.s. and that there exists a continuous and nonsingular function H such that $n^{-1} H_n(\beta)$ converges to $H(\beta)$, uniformly in a ball around β^*.

3. Define for all $1 \leqslant i \leqslant n$, $p_i(\beta) = e^{\langle x_i, \beta \rangle} / \left(1 + e^{\langle x_i, \beta \rangle}\right)$. Check that

$$\begin{aligned}
\mathbb{E}\left[e^{-n^{-1/2}\langle t, \nabla \ell_n(\beta^*)\rangle}\right] &= \prod_{i=1}^n \left(1 - p_i(\beta^*) + p_i(\beta^*) e^{\langle t, x_i \rangle / \sqrt{n}}\right) e^{-p_i(\beta^*)\langle t, x_i \rangle / \sqrt{n}} \,, \\
&= \exp\left(\frac{1}{2} t^T \left(n^{-1} H_n(\beta^*)\right) t + O(n^{-1/2})\right) \,.
\end{aligned}$$

For all $t \in \mathbb{R}^d$,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(-\frac{1}{\sqrt{n}} \langle t, \nabla \ell_n(\beta^\star) \rangle\right)\right] &= \prod_{i=1}^n \mathbb{E}\left[\exp\left(\frac{1}{\sqrt{n}} (Y_i - p_i(\beta^\star)) \langle x_i, t \rangle\right)\right] \,, \\
&= \prod_{i=1}^n \left[\left(1 - p_i(\beta^\star) + p_i(\beta^\star) \exp\left(\frac{1}{\sqrt{n}} \langle x_i, t \rangle\right)\right) \exp\left(-\frac{p_i(\beta^\star)}{\sqrt{n}} \langle x_i, t \rangle\right)\right] \,.
\end{aligned}$$

Note that

$$\log\left(1 - p_i + p_i \exp(u/\sqrt{n})\right) = \log\left(1 + p_i \frac{u}{\sqrt{n}} + p_i \frac{u^2}{2n} + O\left(n^{-3/2}\right)\right) = p_i \frac{u}{\sqrt{n}} + \frac{p_i u^2}{2n} - \frac{p_i^2 u^2}{2n} + O\left(n^{-3/2}\right) \,.$$

Finally,

$$\mathbb{E}\left[\exp\left(-\frac{1}{\sqrt{n}}\langle t, \nabla \ell_n(\beta^\star)\rangle\right)\right] = \exp\left(\underbrace{\frac{1}{2n}\sum_{i=1}^{n} p_i(\beta^\star)(1-p_i(\beta^\star))\langle t, x_i\rangle^2}_{t^T H_n(\beta^\star) t} + O(n^{-1/2})\right).$$

4. What is the asymptotic distribution of $-n^{-1/2}\nabla \ell_n(\beta^*)$ and of $\sqrt{n}(\widehat{\beta}_n - \beta^*)$?

   Recall that for a multivariate random variable X, the moment-generating function is defined as

   $$t \mapsto M_X(t) = \mathbb{E}\left[\exp\left(\langle t, X\rangle\right)\right].$$

   In particular, we know that if $X \sim \mathcal{N}(\mu, \Sigma)$ then

   $$t \mapsto M_X(t) = \mathbb{E}\left[\exp\left(\langle t, \mu + \frac{1}{2}\Sigma t\rangle\right)\right].$$

   If, for all t, $M_{X_n}(t) \to M_X(t)$ then X_n converges to X in distribution.
   For all $t \in \mathbb{R}^d$, since $n^{-1}H_n(\beta^\star) \to_{n\to\infty} H(\beta^\star)$,

   $$\mathbb{E}\left[\exp\left(-\frac{1}{\sqrt{n}}\langle t, \nabla \ell_n(\beta^\star)\rangle\right)\right] \to_{n\to\infty} \exp\left(\frac{1}{2}t^T H(\beta^\star) t\right).$$

   Therefore, $-\nabla \ell_n(\beta^\star)/\sqrt{n}$ converges in distribution to $Z \sim \mathcal{N}(0, H(\beta^\star))$. On the other hand,

   $$\sqrt{n}(\widehat{\beta}_n - \beta^\star) = -\left(\frac{1}{n}H_n(\tilde{\beta}_n)\right)^{-1}\frac{1}{\sqrt{n}}\nabla \ell_n(\beta^\star).$$

   As for all $n \geqslant 1$, $\tilde{\beta}_n \in B(\beta^\star, \|\widehat{\beta}_n - \beta^\star\|)$, $\tilde{\beta}_n$ converges to $\beta^\star$ almost surely as n grows to infinity. Hence, almost surely

   $$\left(\frac{1}{n}H_n(\tilde{\beta}_n)\right)^{-1} \to H(\beta^\star)^{-1}$$

   and, by Slutsky lemma, $\sqrt{n}(\widehat{\beta}_n - \beta^\star)$ converges in distribution to $Z \sim \mathcal{N}(0, H(\beta^\star)^{-1})$.

5. For all $1 \leqslant j \leqslant d$ and all $\alpha \in (0,1)$, propose a confidence interval $\mathcal{I}_{n,\alpha}$ such that $\beta^*_j \in \mathcal{I}_{n,\alpha}$ with asymptotic probability $1-\alpha$.

   According to the last question, $\sqrt{n}(\widehat{\beta}_j - \beta^\star_j)$ converges in distribution to a centered Gaussian random variable with variance $(H(\beta^\star)^{-1})_{jj}$. On the other hand, almost surely,

   $$\widehat{\sigma}^2_{n,j} = (nH_n(\widehat{\beta}_n)^{-1})_{jj} \to_{n\to\infty} (H(\beta^\star)^{-1})_{jj}.$$

   Then,

   $$\sqrt{\frac{n}{\widehat{\sigma}^2_{n,j}}}(\widehat{\beta}_{n,j} - \beta^\star_j) \to_{n\to\infty} \mathcal{N}(0,1).$$

   An asymptotic confidence interval $\mathcal{I}_{n,\alpha}$ of level $1-\alpha$ is then given by

   $$\mathcal{I}_{n,\alpha} = \left[\widehat{\beta}_{n,j} - z_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2_{n,j}}{n}}, \widehat{\beta}_{n,j} + z_{1-\alpha/2}\sqrt{\frac{\widehat{\sigma}^2_{n,j}}{n}}\right],$$

   where $z_{1-\alpha/2}$ is the quantile of order $1-\alpha/2$ of $\mathcal{N}(0,1)$.