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

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

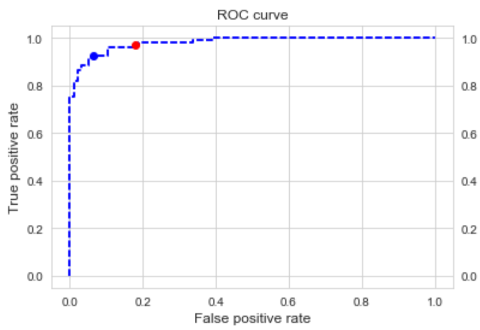
with $\beta^* \in \mathbb{R}^d$. In this case, $\mathbb{P}(Y = 1|X) > 1/2$ if and only if $\langle \beta^*, X\rangle > 0$, so the frontier between $\{x\,;\, h_*(x) = 1\}$ and $\{x\,;\, h_*(x) = 0\}$ is an hyperplane, with orthogonal direction β^*.

1. In this question only, $\beta^* = (\beta_0, \beta_1) \in \mathbb{R}^2$ and $X_i = (1, x_i)$ for all $1 \leqslant i \leqslant n$.

   (a) Provide the value x_* of x_i such that $\mathbb{P}(Y_i = 1|X_i) = 1/2$. The logistic Bayes classifier is therefore defined by $h_*(X_i) = 1$ if and only if $x_i > x_*$.

   (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 $\tilde{h}(X_i) = 1$ if and only if $x_i > \tilde{x}$. Explain a practical interest to choose $\tilde{p} < 1/2$.

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.



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 $1 \leqslant m \leqslant M$.
2. In the setting of the softmax regression function, compute $\theta \mapsto \nabla_\theta \ell_n^{\mathrm{multi}}(\theta)$.

3 Maximum likelihood estimation

The unknown parameter β^* may be estimated by maximizing the conditional likelihood of Y given X

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

to define the empirical classifier

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

In the following, $\{(x_i, Y_i)\}_{1\leqslant i \leqslant n}$ are assumed to be i.i.d. with the same distribution as (X, Y).

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.

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

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

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

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

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