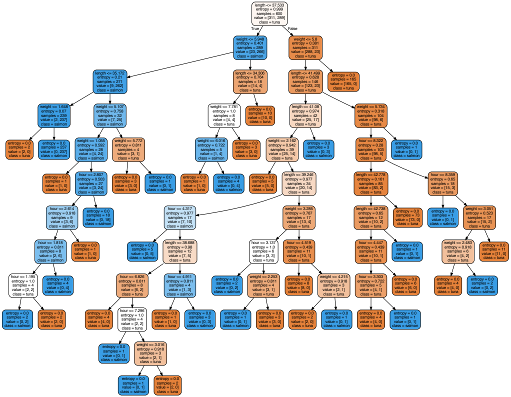
Fondamentaux théoriques du machine learning



Risks and risk decompositions

Risks and risk decompositions

Deterministic bound on the estimation error

We consider the best estimator in the hypothesis space F.

$$f_a = \arg\min_{h \in F} R(h)$$

Exercice 1 : Let us show that

$$R(f_n) - R(f_a) \leq 2 \sup_{h \in F} |R(h) - R_n(h)| \tag{1}$$

Deterministic bound on the estimation error

$$f_a = \arg\min_{h \in F} R(h)$$

$$\begin{aligned}
R(f_n) - R(f_a) &= \big(R(f_n) - R_n(f_n)\big) \\
&+ \big(R_n(f_n) - R_n(f_a)\big) \\
&+ \big(R_n(f_a) - R(f_a)\big) \\
&\leq |R(f_n) - R_n(f_n)| \\
&+ \big(R_n(f_n) - R_n(f_a)\big) \\
&+ |R_n(f_a) - R(f_a)| \\
&\leq 2 \sup_{h \in F} |R(h) - R_n(h)| \\
&+ \big(R_n(f_n) - R_n(f_a)\big)
\end{aligned} \tag{2}$$

But by definition f_n minimizes R_n, so $\big(R_n(f_n) - R_n(f_a)\big) \leq 0$.

Example 1

Exercice 2 : We observe the data $(1, 0)$. We model these data with a Bernoulli distribution of parameter p.

- What is the likelihood of these observations as a function of p ?
- What is the value $\hat{p}$ that maximizes this likelihood ?

Example 2

Exercice 3 : We observe the data $(1, 0, 1)$ (same hypotheses)

- What is the likelihood of these observations as a function of p ?
- What is the value $\hat{p}$ that maximizes this likelihood ?

Link with logistic regression

We consider a binary classification problem, with $\mathcal{Y} = \{0, 1\}$.
Let us now consider the probabilistic model such that

$$p_\theta(1|x) = \sigma(\theta^T x)$$

Equivalently, this model can be written (remember that $y = 0$ or $y = 1$)

$$p_\theta(y|x) = \left(\sigma(\theta^T x)\right)^y \left(1 - \sigma(\theta^T x)\right)^{1-y} \tag{3}$$

Exercice 4 : Show that the parameter θ with maximum likelihood is the logistic regression estimator θ_{logit} (cross entropy version).

We know that $\forall z \in \mathbb{R}, \sigma(-z) = 1 - \sigma(z)$.

$$
\begin{aligned}
R_n(\theta) &= -\frac{1}{n}\sum_{i=1}^{n} \log\left(p_\theta(y_i|x_i)\right) \\
&= -\frac{1}{n}\sum_{i=1}^{n} \log\left(\left(\sigma(\theta^T x_i)\right)^{y_i}\left(1-\sigma(\theta^T x_i)\right)^{1-y_i}\right) \\
&= -\frac{1}{n}\sum_{i=1}^{n} y_i \log\left(\sigma(\theta^T x_i)\right) + (1-y_i)\log\left(\sigma(-\theta^T x_i)\right) \qquad (4) \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i \log\left(1+e^{-\theta^T x_i}\right) + (1-y_i)\log\left(1+e^{\theta^T x_i}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} l(\theta^T x_i, y_i)
\end{aligned}
$$

- Y_{pred} is the random variable representing this prediction (proportional)
- Y is the random variable representing the class, in this node (empirical distribution)

$$\begin{aligned} P(Y_{pred} \neq Y) &= \sum_{l=1}^{L} P\big(Y_{pred} \neq Y|Y = l\big)P(Y = l) \\ &= \sum_{l=1}^{L} \Big(1 - P\big(Y_{pred} = Y|Y = l\big)\Big)P(Y = l) \qquad (5) \\ &= \sum_{l=1}^{L}(1 - p_n^l)p_n^l \end{aligned}$$

Homogeneity criterion for classification : Gini impurity

$$H(n) = \sum_{l=1}^{L} p_n^l (1 - p_n^l) \tag{6}$$

If we predict the classes in node n according to the proportions of the labels in n, then the Gini impurity is the probability of making a mistake, given that we are in node n.