Introduction à l'apprentissage automatique - TD4

Caio Corro

1 Statistical consistency

1. We assume a binary classification problem with $Y = \{-1, 1\}$. We focus on the pointwise setting and denote $w = s(\boldsymbol{x}) \in \mathbb{R}$ the output of the scoring function and $\mu \in [0, 1]$ the Bernoulli parameter of the conditional data distribution, *i.e.* $p(\mathbf{y} = 1|\mathbf{x} = \boldsymbol{x}) = \mu$ and $p(\mathbf{y} = -1|\mathbf{x} = \boldsymbol{x}) = 1 - \mu$.
   Are the following losses classification calibrated? Strictly proper when the sigmoid function is used to map weight w to a Bernoulli parameter? Justify (without using the sufficient conditions theorem).

   (a) Negative log-likelihood loss: $\ell_{(\text{nll})}(y, w) = \log(1 + \exp(-yw))$

   (b) Quadratic error: $\ell_{(\text{quad})}(y, w) = \frac{1}{2}(yw - 1)^2$

   (c) Exponential loss: $\ell_{(\text{exp})}(y, w) = \exp(-yw)$

   (d) Hinge loss: $\ell_{(\text{hinge})}(y, w) = \max(0, 1 - yw)$

   (e) Perceptron loss: $\ell_{(\text{perc.})}(y, w) = \max(0, -yw)$

2. The hard sigmoid function is defined as:

$$\overline{\sigma}(w) = \begin{cases} 0 & \text{if } w \leq -1, \\ \frac{1}{2}(w + 1) & \text{if } w \in [-1, 1], \\ 1 & \text{if } w \geq 1, \end{cases}$$

   This function can be used instead of the standard sigmoid to transform a score $w \in \mathbb{R}$ to a Bernoulli parameter in $[0, 1]$.

   (a) What is one advantage of the hard sigmoid function compared to the standard sigmoid function? (in term of "expressivity")

   (b) Show that the quadratic error loss is classification calibrated for the hard sigmoid prediction function.

3. We assume a multiclass classification problem with $Y = \{1, ..., k\}$. We focus on the pointwise setting and denote $\boldsymbol{w} = s(\boldsymbol{x}) \in \mathbb{R}^k$ the output of the scoring function and $\lambda \in \triangle(k)$ the discrete distribution parameters of the conditional data distribution, *i.e.* $p(\mathbf{y} = y|\mathbf{x} = \boldsymbol{x}) = \lambda_y$.

   (a) The multiclass classification negative log-likelihood loss function is defined as $\ell_{\text{nll}}(y, \boldsymbol{w}) = -w_y + \log \sum_i \exp w_i$. Is it classification calibrated? Strictly proper when the softmax is used to map $\boldsymbol{w}$ to the parameters of a discrete probability distribution? Justify.

   (b) The one-vs-all loss function is defined as $\ell_{\text{ova}}(y, \boldsymbol{w}) = -w_y + \sum_i \log(1 + \exp w_i)$. How can you interpret this loss?

   (c) Prove that $\ell_{\text{ova}}(y, \boldsymbol{w}) \geq \ell_{\text{nll}}(y, \boldsymbol{w})$

   (d) Is the one-vs-all loss function classification calibrated? Strictly proper? Justify.