Introduction à l'apprentissage automatique - TD1

Caio Corro

1 Probabilistic losses

1. Let $D = \left\{(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})\right\}_{i=1}^{n}$ be a dataset. We assume we train a discriminative model and write $p_\theta(\boldsymbol{x}|\boldsymbol{y})$ the conditional model distribution, where θ is the set of model parameters. A good model should be a model that maximizes the probability of the dataset defined as:

$$p_\theta(D) = p_\theta(\boldsymbol{x}^{(i)}, \boldsymbol{y}^{(i)})$$

   Assuming the dataset contains a set of i.i.d. samples (independent and identically distributed), derive the learning problem as the minimization of an additively separable loss function (that is, the object should be a sum over datapoints). **Hint:** negative log-likelihood.

2. We assume a linear model for regression parameterized by $\theta = \{\boldsymbol{a}, b\}$. Assume that we are building a probabilistic model where the output distribution is a Gaussian whose mean is given by the scoring function s_θ and the variance σ^2 is a prefixed value. Show that minimizing the negative log-likelihood in this model is equivalent to learning a linear regression model via the squared error loss.
   The PDF of a gaussian with mean μ and variance σ^2 is defined as:

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{\sigma^2}\right).$$

3. We assume a probabilistic linear model for binary classification, where $Y = \{0, 1\}$.

   (a) Remember that we have:

$$p_\theta(y = 1|\boldsymbol{x}) = \sigma(s_\theta(\boldsymbol{x}))$$
$$p_\theta(y = 0|\boldsymbol{x}) = 1 - \sigma(s_\theta(\boldsymbol{x}))$$

   where σ is the sigmoid function and $s_\theta(\boldsymbol{x}) = \langle \boldsymbol{a}, \boldsymbol{x} \rangle + b$ is the scoring function. Show that we can write:

$$p_\theta(y|x) = \frac{\exp\left(y \times s_\theta(\boldsymbol{x})\right)}{1 + \exp(s_\theta(\boldsymbol{x}))}$$

   (b) Derive the negative log-likelihood loss.

4. We assume a multiclass classification problem with k classes, where $Y = E(k)$.

   - The scoring function $s_\theta : \mathbb{R}^d \to \mathbb{R}^k$ is defined as $s_\theta(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ where $\theta = \{\boldsymbol{A}, \boldsymbol{b}\}$ are the parameters of the model.
   - The probabilistic prediction function is the softmax function: $\widehat{\boldsymbol{y}}(\boldsymbol{w}) = \text{softmax}(\boldsymbol{w})$ defined as:

$$\boldsymbol{\mu} = \text{softmax}(\boldsymbol{w}) \quad \Leftrightarrow \quad \mu_i = \frac{\exp(w_i)}{\sum_j \exp(w_j)}$$

     The output of the softmax can be interpreted as the parameters of a discrete distribution over outputs, i.e. the model distribution is defined as:

$$p_\theta(\boldsymbol{y}|\boldsymbol{x}) = \langle \boldsymbol{y}, \text{softmax}(s_\theta(\boldsymbol{x})) \rangle.$$

     Derive the negative log-likelihood loss.

2 Gradient computation (part 1)

1. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function defined as $f(\boldsymbol{a}) = \frac{1}{2}\|\boldsymbol{a}\|_2^2$. Compute $\nabla f(\boldsymbol{a})$ using the definition of partial derivatives and gradients.
2. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function defined as $f(\boldsymbol{a}) = \frac{1}{2}\|\boldsymbol{X}\boldsymbol{a} - \boldsymbol{y}\|_2^2$. Compute $\nabla f(\boldsymbol{a})$ using the definition of partial derivatives and gradients.
3. Let $f : \mathbb{R} \to \mathbb{R}$ be the softplus function, defined as $f(w) = \log(1 + \exp(w))$ Compute the derivative of f. What is this function?
4. The binary negative log-likelihood loss $\ell_{nll} : \{0, 1\} \times \mathbb{R} \to \mathbb{R}_+$ is defined as $\ell_{nll}(y, w) = -yw + \log(1 + \exp(w))$. Compute $\nabla_w \ell_{nll}(y, w)$. How can you interpret this term?
5. Let $f : \mathbb{R}^k \to \mathbb{R}$ be the log-partition function, defined as $f(\boldsymbol{w}) = \log \sum_i \exp(w_i)$ Compute the gradient of f. What is this function?
6. The multiclass classification negative log-likelihood loss $\ell_{nll} : E(k) \times \mathbb{R}^k \to \mathbb{R}_+$ is defined as $\ell_{nll}(\boldsymbol{y}, \boldsymbol{w}) = -\langle \boldsymbol{w}, \boldsymbol{y} \rangle + \log \sum_i \exp(w_i)$. Compute $\nabla_{\boldsymbol{w}} \ell_{nll}(y, w)$. How can you interpret this term?

3 Impact of regularization (by S. Lall & S. Boyd)

We assume a learning problem with an objective of the following form:

$$\min_\theta \mathcal{L}(\theta) + \beta r(\theta)$$

where θ are the parameters of the model, $\mathcal{L}(\theta)$ is the total loss on the training data, $r(\theta)$ is a regularization term on the parameters and $\beta \geq 0$ is the regularization weight. The specific form of $\mathcal{L}$ and r is not important for this exercise.

Let $\beta, \bar{\beta} \in \mathbb{R}$ and $\theta^*, \bar{\theta}^*$ such that:

$$\begin{aligned} 0 < \beta \leq \bar{\beta} \\ \theta^* = \arg\min_\theta \mathcal{L}(\theta) + \beta r(\theta) \\ \bar{\theta}^* = \arg\min_\theta \mathcal{L}(\theta) + \bar{\beta} r(\theta) \end{aligned}$$

1. Show that $r(\theta^*) \geq r(\bar{\theta}^*)$, i.e. increasing the regularization term will never make the regularization larger.
2. Show that $\mathcal{L}(\theta^*) \leq \mathcal{L}(\bar{\theta}^*)$, i.e. increasing the regularization term will never make the loss lower.
3. What can you deduce from this regarding overfitting?