TECHNISCHE UNIVERSITÄT MÜNCHEN
Fakultät für Elektrotechnik und Informationstechnik
Lehrstuhl für Datenverarbeitung
Prof. Dr. Martin Kleinsteuber

Information Retrieval in High Dimensional Data

Lab #3: Theoretical Exercises, 26.04.2018

Logistic Regression

Task 1. Consider the binary classification problem of assigning a label $y \in \{-1, 1\}$ to a data sample $\mathbf{x} \in \mathbb{R}^p$ by means of Logistic Regression. You are given a training set $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\}$ of labeled data. Recall that the loss function is given by

$$L(\mathbf{w}, b) = \sum_{i=1}^{N} \log(1 + \exp(-y_i(\mathbf{w}^\top \mathbf{x}_i + b))).$$

a) Compute the gradient $\nabla_{\mathbf{w},b} L$.

Solution: Applying the chain rule, we get

$$\begin{aligned}
\nabla_b L &= \sum_{i=1}^{N} \frac{\nabla_b (1 + \exp(-y_i(\mathbf{w}^\top \mathbf{x}_i + b)))}{1 + \exp(-y_i(\mathbf{w}^\top \mathbf{x}_i + b))} \\
&= -\sum_{i=1}^{N} y_i \frac{\exp(-y_i(\mathbf{w}^\top \mathbf{x}_i + b))}{1 + \exp(-y_i(\mathbf{w}^\top \mathbf{x}_i + b))} \\
&= -\sum_{i=1}^{N} \frac{y_i}{1 + \exp(y_i(\mathbf{w}^\top \mathbf{x}_i + b))}.
\end{aligned}$$

Accordingly, we get

$$\begin{aligned}
\nabla_{\mathbf{w}} L &= \sum_{i=1}^{N} \frac{1}{1 + \exp(-y_i(\mathbf{w}^\top \mathbf{x}_i + b))} \nabla_{\mathbf{w}} (1 + \exp(-y_i(\mathbf{w}^\top \mathbf{x}_i + b))) \\
&= -\sum_{i=1}^{N} y_i \frac{\exp(-y_i(\mathbf{w}^\top \mathbf{x}_i + b))}{1 + \exp(-y_i(\mathbf{w}^\top \mathbf{x}_i + b))} \mathbf{x}_i \\
&= -\sum_{i=1}^{N} \frac{y_i}{1 + \exp(y_i(\mathbf{w}^\top \mathbf{x}_i + b))} \mathbf{x}_i.
\end{aligned}$$

b) Assume that the two classes of the training set are linearly separable, i.e. there is a weight vector $\mathbf{w}_s \in \mathbb{R}^p$ and a bias $b_s \in \mathbb{R}$ such that

$$y_i(\mathbf{w}_s^\top \mathbf{x}_i + b_s) > 0 \quad \forall i$$

holds. Show that, under this assumption, the loss function has no global minimum $(\mathbf{w}^*, b^*) \in \mathbb{R}^{p+1}$.

Solution: A global minimum of L is a pair $(\mathbf{w}^*, b^*) \in \mathbb{R}^{p+1}$ such that

$$L(\mathbf{w}, b) \geq L(\mathbf{w}^*, b^*) \ \ \forall (\mathbf{w}, b) \in \mathbb{R}^{p+1}$$

holds. Furthermore, for non-empty training sets, L is strictly positive, so that we can conclude

$$L(\mathbf{w}^*, b^*) = \varepsilon > 0.$$

Assume that such a point exist. Let us define

$$z_i = y_i(\mathbf{w}_s^\top \mathbf{x}_i + b_s).$$

Observe that z_i is strictly positive for every i. Consider the function

$$f(h) = \sum_{i=1}^{N} \log(1 + \exp(-h z_i)).$$

Since every summand approaches 0 as h approaches ∞, so does $f(h)$, i.e.

$$\lim_{h \to \infty} f(h) = 0.$$

Observing the equality

$$f(h) = L(h\mathbf{w_s}, hb_s),$$

this means that for any $\varepsilon > 0$, we can find an $h \in \mathbb{R}$ and set $(\mathbf{w}, b) = (h\mathbf{w_s}, hb_s)$, such that

$$L(\mathbf{w}, b) < \varepsilon$$

holds, which contradicts the assumption of $(\mathbf{w}^*, b^*)$ with $L(\mathbf{w}^*, b^*) = \varepsilon$ being a global minimum.

Note that the hyperplane described by $(\mathbf{w}_s, b_s)$ does not have to be optimal in any sense. Depending on the algorithm this can lead to a perpetual increase of the norm of "non-ideal" hyperplane descriptors.

c) To avoid the scenario in b), the norm of $(\mathbf{w}, b)$ can be penalized by adding a squared norm regularizer. Consider the modified loss function

$$\tilde{L}(\mathbf{w}, b) = L(\mathbf{w}, b) + \lambda(\|\mathbf{w}\|^2 + b^2),$$

where $\lambda > 0$ is a real-valued constant. Compute the gradient $\nabla_{\mathbf{w},b}\tilde{L}$.

Solution: Due to the linearity of the derivative, we have

$$\nabla_b \tilde{L} = \nabla_b L + 2\lambda b,$$

and

$$\nabla_{\mathbf{w}} \tilde{L} = \nabla_{\mathbf{w}} L + 2\lambda \mathbf{w}.$$