Exercise Sheet 4 - Logistic Regression

Prof. Dr. Paul Swoboda

NOTE: For this exercise sheet, you must submit a jupyter notebook and a PDF.

Exercise 1: Hessian of logistic regression (1 Points)

Consider binary logistic regression with training data $(x_i, y_i)_{i=1}^N$, where $x_i \in \mathbb{R}^D$ and $y_i \in \{0, 1\}$. We model

$$p(y_i = 1 \mid x_i, w) = \sigma(w^\top x_i), \qquad p(y_i = 0 \mid x_i, w) = 1 - \sigma(w^\top x_i),$$

where $\sigma(t) = \frac{1}{1+e^{-t}}$ is the logistic sigmoid function.

The negative log-likelihood is

$$L(w; X, y) = -\sum_{i=1}^{N} \Big(y_i \log p(y_i = 1 \mid x_i, w) + (1 - y_i) \log p(y_i = 0 \mid x_i, w) \Big).$$

Compute the Hessian $\nabla_w^2 L(w; X, y)$ of the logistic regression loss with respect to w, and express your result in a matrix form. (A compact matrix form is also acceptable.)

Result should be equivalent to:
Let

$$p_i(w) := p(y_i = 1 \mid x_i, w) = \sigma(w^\top x_i) \quad \textit{for } i = 1, \ldots, N.$$

Define the diagonal matrix

$$S(w) := \operatorname{diag}\big(p_1(w)(1 - p_1(w)), \ldots, p_N(w)(1 - p_N(w))\big) \in \mathbb{R}^{N \times N}.$$

Then the Hessian can be written compactly as

$$\nabla_w^2 L(w; X, y) = X^\top S(w) X.$$

Exercise 2: Kullback–Leibler Divergence (2 Points)

Let p and q be probability distributions on $\mathcal{X}$. The Kullback–Leibler (KL) divergence of q from p is defined as

$$D_{\mathrm{KL}}(p\|q) := \sum_{x\in\mathcal{X}} p(x)\,\log\frac{p(x)}{q(x)}.$$

1. Show that $D_{\mathrm{KL}}(p\|q) \geq 0$ and that $D_{\mathrm{KL}}(p\|q) = 0$ if and only if $p = q$.
2. Give an explicit example of two distributions p and q for which

   $$D_{\mathrm{KL}}(p\|q) \neq D_{\mathrm{KL}}(q\|p),$$

   i.e. show that the KL divergence is in general not symmetric.

Exercise 3: Implement Newton-Method for logistic regression (2 Points)

In the provided Jupyter notebook, implement logistic regression training using Newton's method. Use your results from exercise 1 in your implementation.

Hint: when you need to invert the Hessian, you should use $(\varepsilon \cdot I + H)^{-1}$