MALTA Exercises

Giovanni Neglia and Samir M. Perlaza

October 9, 2024

Notation

Let $\alpha \in (0,1)$ and $\beta \in [0,1]$ be two fixed reals. The binary **relative entropy** (or Kullback-Leibler divergence) of the probability distribution represented by the vector $(\alpha, 1-\alpha)$ with respect to the probability distribution represented by the vector $(\beta, 1-\beta)$, denoted by $\mathsf{D}(\alpha\|\beta)$, is

$$\mathsf{D}(\alpha\|\beta) = \beta \log\left(\frac{\beta}{\alpha}\right) + (1-\beta)\log\left(\frac{1-\beta}{1-\alpha}\right). \tag{1}$$

The binary **entropy** of the probability distribution represented by the vector $(\beta, 1-\beta)$, denoted by $\mathsf{H}(\beta)$, is

$$\mathsf{H}(\beta) = -\beta\log(\beta) - (1-\beta)\log(1-\beta), \tag{2}$$

where $0\log(0)$ is assumed to be zero in both (1) and (2).

Exercises Week 5

Ex. 1 — Consider a dataset

$$S = ((\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)) \in \left(\mathbb{R}^d \times \{-1,1\}\right)^m,$$

and a function $h_{\boldsymbol{w}} : \mathbb{R}^d \to (0,1)$ in the class of functions $\mathcal{H}_{sig}$ of the form

$$h_{\boldsymbol{w}}(\boldsymbol{x}) = \frac{1}{1+\exp\left(-\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}\right)}.$$

In class, it was shown that if for all $t \in \{1,2,\ldots,m\}$, the value $h_{\boldsymbol{w}}(\boldsymbol{x}_t)$ is interpreted as the posterior probability

$$\mathrm{Prob}\left(Y = 1 | \boldsymbol{X} = \boldsymbol{x}_t\right),$$

then the likelihood of the dataset S, denoted by $P_{\boldsymbol{w}}(S)$, is

$$P_{\boldsymbol{w}}(S) = \left(\prod_{t\in\{i\in[m]:y_i=1\}} h_{\boldsymbol{w}}(\boldsymbol{x}_t)\right)\left(\prod_{t\in\{i\in[m]:y_i=-1\}} 1 - h_{\boldsymbol{w}}(\boldsymbol{x}_t)\right), \tag{3}$$

which leads to the equality

$$-\frac{1}{m}\log\left(P_{\boldsymbol{w}}(S)\right) = \frac{1}{m}\sum_{t=1}^{m}\log\left(1+\exp\left(-y_t\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}\right)\right). \tag{4}$$

1. Using the equality in (4), explain the connection between empirical risk minimization and maximum likelihood estimation in the context of logistic regressions.
2. The empirical risk minimization was claimed to be a convex optimization problem. Provide a proof.

3. The names of the labels do not influence the learning process. Assume that the dataset is transformed (by re-labeling) into

$$S = ((\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)) \in \left(\mathbb{R}^d \times \{0, 1\}\right)^m.$$

Show that in such a case, the following holds:

$$\log\left(P_{\boldsymbol{w}}\left(S\right)\right) = \sum_{t=1}^{m} y_t \log\left(h_{\boldsymbol{w}}\left(\boldsymbol{x}_t\right)\right) + (1 - y_t)\log\left(1 - h_{\boldsymbol{w}}\left(\boldsymbol{x}_t\right)\right). \quad (5)$$

Provide an interpretation on the right-hand side of (5) in terms of empirical risk minimization: What is the loss function?

4. The equality in (5) can be written in terms of information measures as follows:

$$-\log\left(P_{\boldsymbol{w}}\left(S\right)\right) = \sum_{t=1}^{m} \mathsf{D}\left(y_t \| h_{\boldsymbol{w}}\left(\boldsymbol{x}_t\right)\right) + H(y_t). \quad (6)$$

Prove the equality in (6) and give an interpretation on the right-hand side of (6) in terms of empirical risk minimization: What is the loss function? **Hint:** Note that for all $t \in \{1, 2, \ldots, m\}$, $H(y_t) = 0$.