Deep Learning KU (DAT.C302UF), WS24
Assignment 1
Maximum Likelihood Estimation, Decision Theory

Thomas Wedenig
thomas.wedenig@tugraz.at

Teaching Assistants:	Patrick Ebnicher, Hade Mohamed
Points to achieve:	10 pts
Deadline:	30.10.2024 23:59
Hand-in procedure:	This is a **solo assignment**. No teams allowed. Submit **your report (PDF)** to the TeachCenter. You do not have to add the cover letter since there are no teams allowed.
Plagiarism:	If detected, 0 points for all parties involved. If this happens twice, we will grade the group with "Ungültig aufgrund von Täuschung"

Supervised Learning – The Setup

Assume we are given a dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i \in \mathbb{R}^d$, and $y_i \in \{0, 1\}$, $\forall i \in \{1, \ldots, n\}$. We can think of $\mathbf{x}_i$ as a feature vector and of y_i as the corresponding *class* (i.e., this is a binary classification problem). Each $(\mathbf{x}_i, y_i)$ tuple is assumed to be an i.i.d. sample from some true, unknown joint distribution $p^*(\mathbf{x}, y)$.

We wish to learn a *discriminative* model that predicts the probability for the binary event y, given a particular feature vector $\mathbf{x}$, i.e., $p_\theta(y \mid \mathbf{x})$.

ovo zelimo naci

Task 1 – Maximum Likelihood Estimation [5 Points]

1. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ and $\mathbf{y} = (y_1, \ldots, y_n)^T$. Write down the likelihood of the entire dataset under our model $p_\theta(\mathbf{y} \mid \mathbf{X})$. Express this in terms of the single-sample likelihoods $p_\theta(y_i \mid \mathbf{x}_i)$ and use the i.i.d. assumption. Write down the negative log-likelihood $\mathrm{NLL}(\theta) = -\log(p_\theta(\mathbf{y} \mid \mathbf{X}))$ as well, again in terms of single-sample likelihoods.

2. Consider the *empirical* distribution induced by $\mathcal{D}$, given by $p_\mathcal{D}(\mathbf{x}, y) = p_\mathcal{D}(\mathbf{x}) p_\mathcal{D}(y \mid \mathbf{x})$ with

$$p_\mathcal{D}(y \mid \mathbf{x}_i) = \begin{cases} 1 & \text{if } y = y_i \\ 0 & \text{else} \end{cases} \qquad p_\mathcal{D}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \delta(\mathbf{x} - \mathbf{x}_i)$$

   where $\delta(\cdot)$ is the Dirac delta function centered around 0.

   Show that the Maximum-Likelihood estimator minimizes the *expected KL-Divergence* between the empirical distribution and the model distribution:

$$\underset{\theta}{\operatorname{argmin}}\ \mathrm{NLL}(\theta) = \underset{\theta}{\operatorname{argmin}}\ \mathbb{E}_{\mathbf{x} \sim p_\mathcal{D}(\mathbf{x})} \left[D_{\mathbb{KL}}(p_\mathcal{D}(\cdot \mid \mathbf{x}),\, p_\theta(\cdot \mid \mathbf{x})) \right]$$

where

$$D_{\mathbb{KL}}(p_{\mathcal{D}}(\cdot \,|\, \mathbf{x}),\, p_\theta(\cdot \,|\, \mathbf{x})) = \mathbb{E}_{y \sim p_{\mathcal{D}}(\cdot \,|\, \mathbf{x})}\left[\log\left(\frac{p_{\mathcal{D}}(y \,|\, \mathbf{x})}{p_\theta(y \,|\, \mathbf{x})}\right)\right]$$

3. Show that the Maximum-Likelihood estimator also minimizes the *expected cross-entropy* between the empirical distribution $p_{\mathcal{D}}(y \,|\, \mathbf{x})$ and the model distribution $p_\theta(y \,|\, \mathbf{x})$, i.e.,

$$\underset{\theta}{\operatorname{argmin}}\ \mathrm{NLL}(\theta) = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{\mathbf{x} \sim p_{\mathcal{D}}(\mathbf{x})}\left[\mathbb{H}_{\mathrm{ce}}(p_{\mathcal{D}}(\cdot \,|\, \mathbf{x}), p_\theta(\cdot \,|\, \mathbf{x}))\right]$$

where $\mathbb{H}_{\mathrm{ce}}(p_{\mathcal{D}}(\cdot \,|\, \mathbf{x}), p_\theta(\cdot \,|\, \mathbf{x}))$ denotes the *cross-entropy* between the input distributions, defined by

$$\mathbb{H}_{\mathrm{ce}}(p_{\mathcal{D}}(\cdot \,|\, \mathbf{x}), p_\theta(\cdot \,|\, \mathbf{x})) = \mathbb{E}_{y \sim p_{\mathcal{D}}(\cdot \,|\, \mathbf{x})}\left[-\log(p_\theta(y \,|\, \mathbf{x}))\right]$$

4. In general, given distributions $p(\mathbf{z})$ and $q(\mathbf{z})$, show the relationship between $D_{\mathbb{KL}}(p, q)$, $\mathbb{H}_{\mathrm{ce}}(p, q)$ and $\mathbb{H}(p)$, where $\mathbb{H}(p) = \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})}\left[-\log(p(\mathbf{z}))\right]$ is the *entropy* of p. Hint: Start by writing down the definition of $D_{\mathbb{KL}}(p, q)$. Also recall that the expectation operator is *linear*, i.e., $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$.

Task 2 – Decision Theory [5 Points]

For all following tasks, assume we have access to the *true* posterior $p^*(y \,|\, \mathbf{x})$, for each $\mathbf{x} \in \mathbb{R}^d$.

1. We define a *loss function*

$$\mathcal{L}(y, \hat{y}) = \begin{cases} 1 & \text{if } y \neq \hat{y} \\ 0 & \text{else} \end{cases}$$

where y is the *true, observed* label, and $\hat{y}$ is the model's prediction. This function is the so-called *zero-one* loss. Write down the decision function $f : \mathbb{R}^d \to \{0, 1\}$ that minimizes

$$\mathbb{E}_{(\mathbf{x}, y) \sim p^*(\mathbf{x}, y)}\left[\mathcal{L}(y, f(\mathbf{x}))\right]$$

i.e., the expected loss over the data generating distribution.

2. Does there exist a *different*[1] decision function f' that – in expectation over $p^*(\mathbf{x}, y)$ – makes *fewer misclassifications*? Explain why/why not.

3. If f had access to the marginal $p^*(\mathbf{x})$, could we construct a decision function that achieves a lower expected loss? Explain why/why not.

4. We define a *new* loss function $\mathcal{L}(y, \hat{y}) = L_{y, \hat{y}}$ with

$$L = \begin{bmatrix} 0 & 1 \\ 10 & 0 \end{bmatrix}$$

where matrix indexing is 0-based. For example, $L_{0,1} = 1$ and $L_{1,0} = 10$. Using this new loss function, again write down the definition of the decision function $g : \mathbb{R}^d \to \{0, 1\}$ that minimizes the expected loss

$$\mathbb{E}_{(\mathbf{x}, y) \sim p^*(\mathbf{x}, y)}\left[\mathcal{L}(y, g(\mathbf{x}))\right].$$

5. For a particular $\mathbf{x}$, assume the true posterior is $p^*(y = 0 \,|\, \mathbf{x}) = 0.9$ and $p^*(y = 1 \,|\, \mathbf{x}) = 0.1$. What is the output of $f(\mathbf{x})$ and $g(\mathbf{x})$? Explain any differences in their decision.

6. Assume that y encodes if a patient (with feature vector $\mathbf{x}$) has a disease ($y = 1$), or is healthy ($y = 0$). In words, briefly describe what the matrix L encodes in this case.

[1] i.e., $\exists \mathbf{x} \in \mathbb{R}^d : f'(\mathbf{x}) \neq f(\mathbf{x})$