Foundations of Machine Learning – Homework assignment 2

CentraleSupélec MSc AI
yannick.le-cacheux@centralesupelec.fr

November 2024

As a reminder, this "homework" is entirely optional. As such, you may disregard the points awarded per exercise – they are used for another class.

You can email me if you want the solution. You are also welcome to send me your solution if you want some feedback.

1 SVMs and kernels [40 points]

The *kernel trick* in SVMs consists in obtaining the same decision boundary as if we applied the SVM in a higher dimensional space, without actually having to compute the coordinates of the training points in this space. In order to achieve this, we can reduce the projections of points into high-dimensional space and the computation of dot products in this space to a single, faster operation.

A function $k(\cdot, \cdot)$ with two inputs of same dimension D written $\mathbf{x}$ and $\mathbf{x}'$ is said to be a *kernel* if there exists a function $\boldsymbol{\phi}(\cdot) : \mathbb{R}^D \rightarrow \mathbb{R}^H$ such that

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\phi}(\mathbf{x}') \tag{1}$$

1.1 [10 points]. Prove that the function $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}')^2$ is a kernel.
1.2 [12 points]. Prove that if k_1 and k_2 are valid kernels, $k_1 + k_2$ as well as $k_1 \cdot k_2$ and $\alpha \cdot k_1$ with $\alpha > 0$ are valid kernels.
1.3 [8 points]. Prove that $k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\top \mathbf{x}' + c)^d$ is a valid kernel for any $c \geq 0$ and any $d \in \mathbb{N}$.
1.4 [10 points]. Prove that this corresponds to polynomial features.

2 Evaluation and metrics [30 points]

2.1 Precision and recall [10 points]. A fellow machine learning practitioner has designed a model with a precision of 0.9 and a recall of 0.8. Based on this information only, can you estimate the probability that an outcome is negative if the model has predicted that it is positive? What about the probability that the outcome is negative if the model has predicted that it is negative? Provide the calculation(s) and the result(s) if yes, and explain why not if no.

2.2 Sigmoid and softmax functions [10 points]. In logistic regression, based on the "raw score" $\mathbf{w}^\top \mathbf{x}$, we estimate the probability that the outcome is positive with

$$p(y = 1 | \mathbf{x}) = \frac{1}{1 + e^{-\mathbf{w}^\top \mathbf{x}}} \tag{2}$$

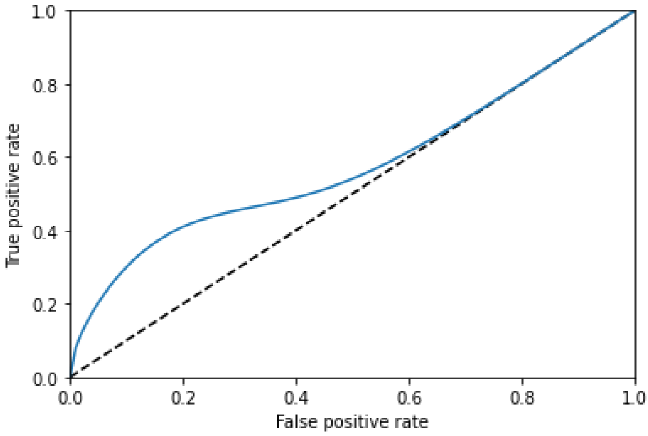

Figure 1: A Receiver Operating Characteristic (ROC) curve.

In deep learning, in a multi-class setting with K different classes, we usually have a “raw” activation score a_k for each class k, and the probability that an instance $\mathbf{x}$ belongs to class k is obtained using the softmax function

$$p(y = k|\mathbf{x}) = \frac{e^{a_k}}{\sum_{i=1}^{K} e^{a_i}} \tag{3}$$

Prove that if $K = 2$ and we use a fully connected network with no hidden layer, this is equivalent to making predictions with a logistic regression model.

2.3 ROC curve interpretation [10 points]. Figure 1 (on the next page) shows the Receiver Operating Characteristic (ROC) curve of a model predicting whether a *prospect* (a potential client) will buy an insurance product. Based on this curve only, do you think the model is better at identifying prospects more likely to buy than average, less likely to buy than average, or is not better at one than the other? Briefly explain why.

3 Information theory [30 points]

For a random variable x sampled from a discrete probability distribution with (a finite number of) outcomes in $\mathcal{X}$, the *entropy* of x is defined as

$$H[x] = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{4}$$

The *joint entropy* of the joint distribution $p(x, y)$ of two random variables x and y with outcomes in $\mathcal{X}$ and $\mathcal{Y}$ is similarly defined as

$$H[x, y] = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(x, y) \tag{5}$$

Finally, the *conditional entropy* of y given x is defined as

$$H[y|x] = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log p(y|x) \tag{6}$$

3.1 [10 points]. Prove that $H[x,y] = H[y|x] + H[x]$. How do you intuitively explain this?
3.2 [10 points]. Prove that $H[x,y] \leq H[x] + H[y]$, or equivalently prove that $H[y|x] \leq H[y]$. How do you intuitively explain this?
Hint: Jensen's inequality may be useful here. For a convex function $\phi : \mathbb{R} \to \mathbb{R}$,

$$\phi\left(\frac{\sum_i a_i x_i}{\sum_i a_i}\right) \leq \frac{\sum_i a_i \phi(x_i)}{\sum_i a_i} \tag{7}$$

3.3 [10 points]. Prove that $H[x,y] = H[y] + H[x]$ if and only if y and x are independent. How do you intuitively explain this?
Hint: there is a special case of Jensen's inequality for *strictly* convex functions.

4 *Bonus question [+10 points]*

4.1 Expectation-maximization for mixtures of Gaussians ***[+10 points].*** If we model a set of N observations $\mathbf{X} = (\mathbf{x}_1, \dots \mathbf{x}_N)^\top \in \mathbb{R}^{N \times D}$ as a mixture of K gaussians, the log-likelihood of the data may be expressed as

$$\log p(\mathbf{X}|\boldsymbol{\pi}, \mathbf{M}, \boldsymbol{S}) = \sum_{n=1}^{N} \log\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \tag{8}$$

Using Equation (8), derive a condition that every $\boldsymbol{\mu}_k$ must meet if the log-likelihood is maximized. How do you interpret this?

4.2 Kullback-Leibler divergence in variational auto-encoders ***[+10 points].*** In variational auto-encoders, we assume that the ground-truth prior distribution p over elements $\mathbf{z} \in \mathbb{R}^H$ of the latent space is a standard multivariate normal distribution $\mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I})$, and the estimated posterior distribution q is a multivariate normal distribution with mean $\boldsymbol{\mu}$ and a diagonal covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\sigma}^2 \odot \mathbf{I} = \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_H^2 \end{pmatrix}$:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}) \tag{9}$$
$$q(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2 \odot \mathbf{I}) \tag{10}$$

Prove that under these assumptions, their Kullback-Leibler divergence is given by

$$D_{\mathrm{KL}}[q||p] = \frac{1}{2}\sum_{h=1}^{H}(\sigma_h^2 + \mu_h^2 - \log \sigma_h^2 - 1) \tag{11}$$