Assignment 2: CS-E4830 Kernel Methods in Machine Learning 2022

The **deadline** for this assignment is **Monday 04.04.2022 at 4pm**. If you have **questions** about the assignment, you can ask them in the corresponding Zulip stream. We will have an exercise session regrading this assignment on 31.03.22 at 4:15 pm in TU1 Saab Auditorium.

Please follow the **submission instructions** corresponding to this assignment as given in MyCourses `https://mycourses.aalto.fi/course/view.php?id=32426&section=1`

Pen & Paper (8 points)

Kernel centering

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function and $\phi : \mathcal{X} \to F$ a feature map associated with this kernel. Let $S = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be the set of training inputs.

Centering the data in the feature space moves the origin of the feature space to the center of mass of the training features $\frac{1}{N}\sum_{i=1}^{N} \phi(\mathbf{x}_i)$ and generally helps to improve the performance. After centering, the feature map is given by: $\phi_c(\mathbf{x}) = \phi(\mathbf{x}) - \frac{1}{N}\sum_{i=1}^{N} \phi(\mathbf{x}_i)$. We will see in this question that centering can be performed implicitly by transforming the kernel values.

Task 1: (2 points)

Show that

$$k_c(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) - \frac{1}{N}\sum_{p=1}^{N} k(\mathbf{x}_p, \mathbf{x}_j) - \frac{1}{N}\sum_{q=1}^{N} k(\mathbf{x}_i, \mathbf{x}_q) + \frac{1}{N^2}\sum_{p=1}^{N}\sum_{q=1}^{N} k(\mathbf{x}_p, \mathbf{x}_q),$$

where $k_c(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi_c(\mathbf{x}_i), \phi_c(\mathbf{x}_j) \rangle$ is the kernel value after centering.

Task 2 (3 points):

Consider the binary classification as discussed in Lecture 4 and shown in Figure 1, where the probability densities, $p(x, C_1)$ and $p(x, C_2)$ for the two classes are known.

1. (1 point) For the point $\hat{x}$, compute the probability that it belongs to C_1, i.e., $P(y = C_1 | X = \hat{x})$.
2. (2 points) Prove that the probability of the minimum misclassification error satisfies this inequality:

$$P(\text{Minimum misclassification error}) \leq \int_{x \in \mathcal{X}} \left(p(x, C_1)p(x, C_2)\right)^{1/2} dx$$

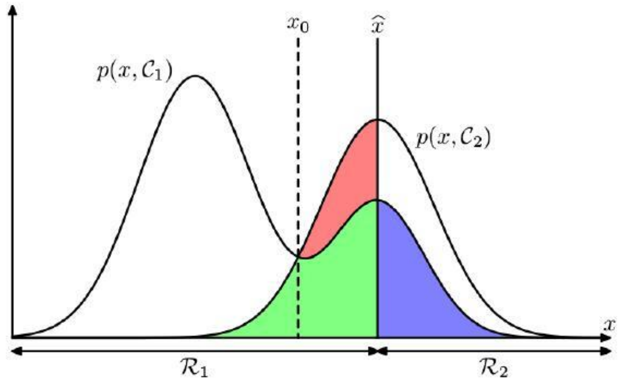

Figure 1: Data distribution for a binary classification problem

Hint : In the proof you can apply the following inequality, for any $a \geq 0$ and $b \geq 0$ we have

$$\min(a, b) \leq (ab)^{1/2}.$$

Multiclass classification

Recall from Lecture 4, where the Bayes classifier has been introduced. In those slides a decision rule to predict the classes, C_1 and C_2 has been presented. That rule selects that class which has the greater conditional probability at a given x, namely

$$\arg\max_k P(y = C_k|X = x), k = 1, 2$$

The above setup can deal with two classes.

Task 3: (1 point)

Let $\mathbf{x}_i \in \mathcal{R}^d$ be an input example, and $\mathbf{w}_k \in \mathcal{R}^d, k = 1, \ldots, K$ a set of parameter vectors assigned to each class in the multi-class classification. Let the probability $P(Y_i = k|X = x_i)$ of a class with respect to $\mathbf{x}_i$ be given by $\frac{1}{Z}\exp(\langle \mathbf{w}_k, \mathbf{x}_i \rangle)$, where Z is a normalization factor to guarantee that $\frac{1}{Z}\exp(\langle \mathbf{w}_k, \mathbf{x}_i \rangle)$ is a probability.

The task is to suggest a multi-class decision function for this concrete probability model, and derive the value of Z for a fixed number of classes.

Task 4: (2 points)

Consider a random variable ϵ that takes the values $\{-1, +1\}$ with equal probability. Show that

$$\mathbb{E}[e^{\lambda\epsilon}] \leq e^{\frac{\lambda^2}{2}} \text{ for all } \lambda \in \mathbb{R}$$

where $\mathbb{E}[.]$ denotes the expectation w.r.t the random variable ϵ.

Hint : Use power series expansion of the exponential function.

Programming (8 points)

Solve the programming tasks in JupyterHub (`https://jupyter.cs.aalto.fi`). The instructions for that are given in MyCourses: `https://mycourses.aalto.fi/course/view.php?id=32426&section=4`.