CS-E4830 - Kernel Methods in Machine Learning D
Exercise 2

Alex Herrero Pons

Autumn Term Course 2020-2021

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.

Question 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.

Solution.
Knowing that

$$\phi_c(\mathbf{x}) = \phi(\mathbf{x}) - \frac{1}{N}\sum_{i=1}^{N}\phi(\mathbf{x}_i)$$

we can obtain

$$k_c(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi_c(\mathbf{x}_i), \phi_c(\mathbf{x}_j) \rangle = \left\langle \phi(\mathbf{x}_i) - \frac{1}{N}\sum_{p=1}^{N}\phi(\mathbf{x}_p), \phi(\mathbf{x}_j) - \frac{1}{N}\sum_{q=1}^{N}\phi(\mathbf{x}_q) \right\rangle$$

Then, applying that $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$ and $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$ we can expand:

$$k_c(\mathbf{x}_i, \mathbf{x}_j) = \left\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) - \frac{1}{N}\sum_{q=1}^{N}\phi(\mathbf{x}_q) \right\rangle - \left\langle \frac{1}{N}\sum_{p=1}^{N}\phi(\mathbf{x}_p), \phi(\mathbf{x}_j) - \frac{1}{N}\sum_{q=1}^{N}\phi(\mathbf{x}_q) \right\rangle =$$

$$= \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle - \left\langle \phi(\mathbf{x}_i), \frac{1}{N} \sum_{q=1}^{N} \phi(\mathbf{x}_q) \right\rangle - \left\langle \frac{1}{N} \sum_{p=1}^{N} \phi(\mathbf{x}_p), \phi(\mathbf{x}_j) \right\rangle - \left\langle \frac{1}{N} \sum_{p=1}^{N} \phi(\mathbf{x}_p), \frac{1}{N} \sum_{q=1}^{N} \phi(\mathbf{x}_q) \right\rangle =$$

$$= \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle - \frac{1}{N} \sum_{q=1}^{N} \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_q) \rangle - \frac{1}{N} \sum_{p=1}^{N} \langle \phi(\mathbf{x}_p), \phi(\mathbf{x}_j) \rangle - \frac{1}{N^2} \sum_{p=1}^{N} \sum_{q=1}^{N} \langle \phi(\mathbf{x}_p), \phi(\mathbf{x}_q) \rangle =$$

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

$\square$

Question 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., $P(y = C_1 | X = \hat{x})$.

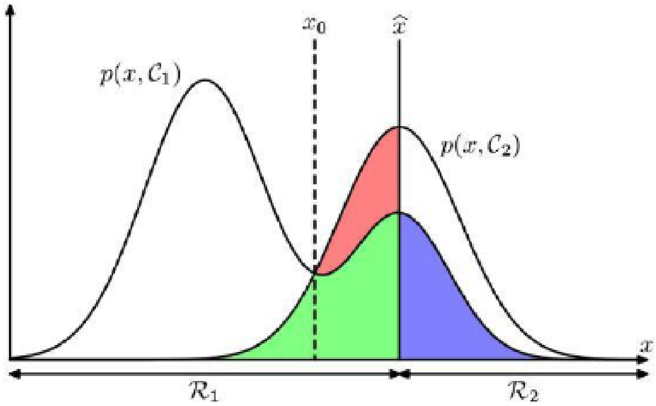

Figure 1: Data distribution for a binary classification problem

Solution.

$$P(y = C_1 | X = \hat{x}) = \frac{p(\hat{x}, C_1)}{p(\hat{x}, C_1) + p(\hat{x}, C_2)}$$

2. (2 points) Prove that the probability of the minimum misclassification error satisfies this inequality:

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

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

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

Solution.
The probability of the minimum misclassification error is represented in Figure 1 by the green and blue zones. We can observe that this can be computed as:

$$P(\text{Minimum misclassification error}) = \int_{x \in \mathcal{X}} \frac{\min\left(p(x, C_1), p(x, C_2)\right)}{p(x, C_1) + p(x, C_2)} dx$$

Knowing that $p(x, C_1) + p(x, C_2) \geq 0$:

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

And using the given hint ($\min(a, b) \leq (ab)^{1/2}$):

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

$\square$

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.

Question 3 (1 points)

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.

Solution.
Having that

$$P(Y_i = k | X = x_i) = \frac{1}{Z} \exp(\langle \mathbf{w}_k, \mathbf{x}_i \rangle),$$

we know that,

$$\sum_{k=1}^{K} P(Y_i = k | X = x_i) = \frac{1}{Z} \sum_{k=1}^{K} \exp(\langle \mathbf{w}_k, \mathbf{x}_i \rangle) = 1.$$

Therefore,

$$Z = \sum_{k=1}^{K} \exp(\langle \mathbf{w}_k, \mathbf{x}_i \rangle).$$

Knowing this we obtain that the multi-class decision function can be expressed as:

$$P(Y_i = k | X = x_i) = \frac{\exp(\langle \mathbf{w}_k, \mathbf{x}_i \rangle)}{\sum_{l=1}^{K} \exp(\langle \mathbf{w}_l, \mathbf{x}_i \rangle)}$$

Question 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.

Solution.
The power series expansion of the exponential function shows that:

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

Therefore, the expectation

$$\mathbb{E}[e^{\lambda\epsilon}] = \mathbb{E}\left[\sum_{n=0}^{\infty} \frac{(\lambda\epsilon)^n}{n!}\right] = \sum_{n=0}^{\infty} \frac{(-\lambda)^n}{n!} + \sum_{n=0}^{\infty} \frac{(\lambda)^n}{n!} = e^{-\lambda} + e^{\lambda} = \frac{1}{e^{\lambda}} + e^{\lambda} = \frac{1 + e^{\lambda 2}}{e^{\lambda}} =$$

$$\leq e^{\frac{\lambda^2}{2}}$$