CS-E4830 - Kernel Methods in Machine Learning

Homework Assignment 2 – Pen and Paper

Student: Quoc Tuan Vinh, Ngo (704526)

Question 1. "*Kernel Centering*"

Let $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ be a kernel function and $\emptyset : \mathcal{X} \rightarrow F$ a feature map associated with this kernel. Let $S = \{ x_1, \dots, x_l\}$ be the set of training inputs.

After centering, the feature map is given by: $\emptyset_c(\boldsymbol{x}) = \emptyset(\boldsymbol{x}) - \frac{1}{l}\sum_{i=1}^{l} \emptyset(\boldsymbol{x}_i)$.

Therefore, with $k_c(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \emptyset_c(\boldsymbol{x}_i), \emptyset_c(\boldsymbol{x}_j)\rangle$, we can do as following:

$$k_c(\boldsymbol{x}_i, \boldsymbol{x}_j) = \langle \emptyset_c(\boldsymbol{x}_i), \emptyset_c(\boldsymbol{x}_j)\rangle = \emptyset_c(\boldsymbol{x}_i)\, \emptyset_c(\boldsymbol{x}_j)^T$$

$$= \left(\emptyset(\boldsymbol{x}_i) - \frac{1}{l}\sum_{p=1}^{l} \emptyset(\boldsymbol{x}_p)\right)\left(\emptyset(\boldsymbol{x}_j) - \frac{1}{l}\sum_{q=1}^{l} \emptyset(\boldsymbol{x}_q)\right)^T$$

$$= \emptyset(\boldsymbol{x}_i)\emptyset(\boldsymbol{x}_j)^T - \frac{1}{l}\sum_{q=1}^{l} \emptyset(\boldsymbol{x}_i)\, \emptyset(\boldsymbol{x}_q)^T - \frac{1}{l}\sum_{p=1}^{l} \emptyset(\boldsymbol{x}_p)\, \emptyset(\boldsymbol{x}_j)^T + \frac{1}{l^2}\sum_{p=1}^{l} \emptyset(\boldsymbol{x}_p) \sum_{q=1}^{l} \emptyset(\boldsymbol{x}_q)^T$$

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

, which is what we need to prove.

Question 2. "*Multiclass (multinomial) classification*"

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

Thus, for a multiclass problem with a fixed number of K, the prediction is made upon the value k with the highest probability from Gibbs measure function. Therefore, we can define the decision function as below:

$$f_n(x) = a,\ \ with\ a \in \{1, \dots, K\}\ and\ a = \underset{k\,\in\{1,..,K\}}{\arg\max} \frac{1}{Z}\exp(\langle \boldsymbol{w}_k, \boldsymbol{x}_i\rangle)$$

Since Z functions as a normalization factor to make $\frac{1}{Z}\exp\left(\langle \boldsymbol{w}_k, \boldsymbol{x}_i \rangle\right)$ a probability, then:

$$\frac{1}{Z}\exp(\langle \boldsymbol{w}_k, \boldsymbol{x}_i \rangle) \leq 1 \Leftrightarrow \exp(\langle \boldsymbol{w}_k, \boldsymbol{x}_i \rangle) \leq Z$$

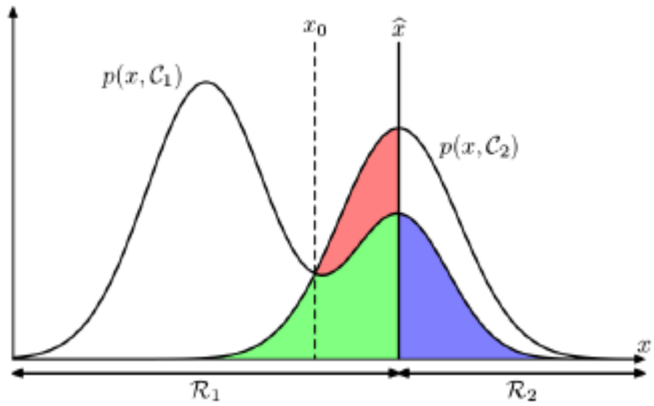
Also, with a fixed K, we can the sum of all components being set to 1:

$$\sum_{i=1}^{K} P(Y_i = k \mid X = x_i) = 1 \Leftrightarrow \sum_{i=1}^{K} \frac{1}{Z}\exp(\langle \boldsymbol{w}_k, \boldsymbol{x}_i \rangle) = 1 \Leftrightarrow \frac{1}{Z}\sum_{i=1}^{K} \exp(\langle \boldsymbol{w}_k, \boldsymbol{x}_i \rangle) = 1$$

$$\Leftrightarrow Z = \sum_{i=1}^{K} \exp(\langle \boldsymbol{w}_k, \boldsymbol{x}_i \rangle)$$

Question 3.

Given a binary classification problem, in which $p(x, C_1)$ and $p(x, C_2)$ are known. This graph below is extracted from lecture 5:



At a given point, we can calculate the misclassification error as below:

$$P(Misclassification\ Error) = \int_{-\infty}^{+\infty} p(misclassification\ error, x)dx$$

$$= \int_{\mathcal{R}_1} p(x, C_2)dx + \int_{R_2} p(x, C_1)dx$$

The transformation above explains that the misclassification error is equal to the sum of misclassifying in each sub-graph: In the sub-graph belonging to $\mathcal{R}_1$, the error is the probability of classifying as C_2; meanwhile in the sub-graph belonging to $\mathcal{R}_2$, the error is the probability of classifying as C_1.

Let x_0 denote the point that achieve that minimum classification error, then at $x = x_0$, our error is either the misclassification error of either C_1 or C_2. Thus, the probability of the minimum classification error can be explained as:

$$P(Minimum\ Misclassification\ Error) = \int_{-\infty}^{x_0} p(x, C_2)dx + \int_{x_0}^{+\infty} p(x, C_1)dx$$

$$= \min\left(\int_{-\infty}^{x_0} p(x, C_2)dx\,, \int_{x_0}^{+\infty} p(x, C_1)dx\right) \leq \int_{x \in \mathcal{X}} \left(p(x, C_1)p(x, C_2)\right)^{\frac{1}{2}} dx$$

Which is what we need to prove.