Exercise 08 of Machine Learning [IN 2064]

Name: Yiman Li
Matr-Nr: 03724352
cooperate with Kejia Chen(03729686)

Problem 1

According to [1], both perceptron algorithm and support machine vector are solutions that separate the classes exactly.
However, the solution that perceptron algorithm finds will be dependent on the arbitrary initial values chosen for $\boldsymbol{w}$ and b as well as on the order in which the data points are presented. If there are multiple solutions all of which classify the training data set exactly, the we should try to find the one that will give the smallest generalization error.
On the other hand, the support vector machine approaches this problem through the concept of the margin, which is defined to be the smallest distance between the decision boundary and any of the samples.

Problem 2

a) Recall that $\boldsymbol{X} \in \mathcal{R}^{N\times D}$, $\boldsymbol{y} \in \mathcal{R}^{N}$, so we have

$$\begin{aligned} g(\boldsymbol{\alpha}) &= \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N} y_i y_j \alpha_i \alpha_j \boldsymbol{x_i}^T \boldsymbol{x_j} \\ &= \frac{1}{2}\boldsymbol{\alpha}^T \boldsymbol{Q}\boldsymbol{\alpha} + \boldsymbol{\alpha}^T \mathbf{1}_N \\ &= \frac{1}{2}\boldsymbol{\alpha}^T(-\boldsymbol{X}\boldsymbol{X}^T \odot \boldsymbol{y}\boldsymbol{y}^T)\boldsymbol{\alpha} + \boldsymbol{\alpha}^T \mathbf{1}_N \end{aligned} \tag{1}$$

so we get

$$\boldsymbol{Q} = -\boldsymbol{X}\boldsymbol{X}^T \odot \boldsymbol{y}\boldsymbol{y}^T \tag{2}$$

b) Firstly we have

$$\boldsymbol{Q} = -\begin{bmatrix} \boldsymbol{x}_1^T\boldsymbol{x}_1 y_1 y_1 & \boldsymbol{x}_1^T\boldsymbol{x}_2 y_1 y_2 & \cdots & \boldsymbol{x}_1^T\boldsymbol{x}_N y_1 y_N \\ \boldsymbol{x}_2^T\boldsymbol{x}_1 y_2 y_1 & \boldsymbol{x}_2^T\boldsymbol{x}_2 y_2 y_2 & \cdots & \boldsymbol{x}_2^T\boldsymbol{x}_N y_2 y_N \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_N^T\boldsymbol{x}_1 y_N y_1 & \boldsymbol{x}_N^T\boldsymbol{x}_2 y_N y_2 & \cdots & \boldsymbol{x}_N^T\boldsymbol{x}_N y_N y_N \end{bmatrix} \tag{3}$$

note that

$$\boldsymbol{Q}_{(i,j)} = -(\boldsymbol{x}_i^T\boldsymbol{x}_j y_i y_j) = -(\boldsymbol{x}_j^T\boldsymbol{x}_i y_j y_i) = \boldsymbol{Q}_{(j,i)} \tag{4}$$

which means that $\boldsymbol{Q}$ is symmetric. Since $\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{y}\boldsymbol{y}^T$ are all $\boldsymbol{Gram\ Matrix}$, so $-\boldsymbol{Q}$ are always negative (semi-)definite matrix. Considering the linear constraints on $\boldsymbol{\alpha}$, so the function $g(\boldsymbol{\alpha})$ is concave, and the local maximizer of g is just its global maximizer. So we can search for a local maximizer of g to find its maximum.

Problem 3

Intuitively, the misclassification rate of LOOCV can be understood as below:

$$\epsilon = \frac{Number\ of\ misclassification\ samples}{Number\ of\ all\ samples} \tag{5}$$

Since we use support vector machine here with the number of training samples being s, so the number of misclaasification samples must be smaller than s, so we have

$$\epsilon \leq \frac{s}{N} \tag{6}$$

Problem 4

To be see in the end.

Problem 5

Firstly, we have

$$\begin{aligned} k(\boldsymbol{x}_m^T, \boldsymbol{x}_n^T) &= \sum_{i=1}^{N} a_i (\boldsymbol{x}_m^T \boldsymbol{x}_n)^i + a_0 \\ &= \sum_{i=1}^{N} a_i (\boldsymbol{x}_n^T \boldsymbol{x}_m)^i + a_0 \\ &= k(\boldsymbol{x}_n^T, \boldsymbol{x}_m^T) \end{aligned} \tag{7}$$

Which means the matrix is symmetric. Then the kernel matrix

$$\boldsymbol{K} = \begin{bmatrix} \sum_{i=1}^{N} a_i (\boldsymbol{x}_1^T \boldsymbol{x}_1)^i + a_0 & \sum_{i=1}^{N} a_i (\boldsymbol{x}_1^T \boldsymbol{x}_2)^i + a_0 & \cdots & \sum_{i=1}^{N} a_i (\boldsymbol{x}_1^T \boldsymbol{x}_N)^i + a_0 \\ \sum_{i=1}^{N} a_i (\boldsymbol{x}_2^T \boldsymbol{x}_1)^i + a_0 & \sum_{i=1}^{N} a_i (\boldsymbol{x}_2^T \boldsymbol{x}_2)^i + a_0 & \cdots & \sum_{i=1}^{N} a_i (\boldsymbol{x}_2^T \boldsymbol{x}_N)^i + a_0 \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{N} a_i (\boldsymbol{x}_N^T \boldsymbol{x}_1)^i + a_0 & \sum_{i=1}^{N} a_i (\boldsymbol{x}_N^T \boldsymbol{x}_2)^i + a_0 & \cdots & \sum_{i=1}^{N} a_i (\boldsymbol{x}_N^T \boldsymbol{x}_N)^i + a_0 \end{bmatrix} \tag{8}$$

$$= a_1 \boldsymbol{X}\boldsymbol{X}^T + a_2 (\boldsymbol{X}\boldsymbol{X}^T)^2 + \cdots + a_N (\boldsymbol{X}\boldsymbol{X}^T)^N + a_0$$

which is the positive sum of a series of $\boldsymbol{Gram\ Matrix}$ plus a constant a_0. Since $\boldsymbol{Gram\ Matrix}$ is alway positive semi-definite, so we have that the kernel matrix is also positive semi-definite for any input data $\boldsymbol{X}$.
Since the function here gives rise to a symmetric and positive semi-definite kernel matrix $\boldsymbol{K}$, so we can say that the function k here is a valid kernel.

Problem 6

For $x_1, x_2 \in (0, 1)$, consider a kernel function in an infinite-dimensional feature space, we have

$$\lim_{n=\infty} x_1^n = 0\ ,\ \ \lim_{n=\infty} x_2^n = 0\ ,\ \ \lim_{n=\infty} (x_1 x_2)^n = 0 \tag{9}$$

so use the feature transformation function

$$\begin{aligned} \boldsymbol{\phi}(x) &= 1 + x^1 + x^2 + \cdots + x^n \\ &= \frac{1 - x^n}{1 - x} \\ &\approx \frac{1}{1 - x} \end{aligned} \tag{10}$$

then we have

$$\begin{aligned} k(x_1, x_2) &= \phi(x_1)^T \phi(x_2) \\ &= 1 + x_1 x_2 + (x_1 x_2)^2 + \cdots + (x_1 x_2)^n \\ &= \frac{1 - (x_1 x_2)^n}{1 - x_1 x_2} \\ &\approx \frac{1}{1 - x_1 x_2} \end{aligned} \tag{11}$$

References

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007.