Homework 4

姓名：毕秋宇　　**学号**：171860624

2019 年 5 月 1 日

第一部分　作业

Problem 1 (Kernel Methods)

From Mercer theorem, we know a two variables function $k(\cdot,\cdot)$ is a positive semi-definite kernel function if and only if for any N vectors $x_1, x_2, \cdots, x_N$, their kernel matrix is positive semi-definite. Assume $k_1(\cdot,\cdot)$ and $k_2(\cdot,\cdot)$ are positive definite kernel function for matrices K_1 and K_2. The element of kernel matrix K is denoted as $K_{ij} = k(x_i, x_j)$. Please proof the kernel function corresponding to the following matrices is positive semi-definite.

(1) [5pts] $K_3 = a_1K_1 + a_2K_2$ where $a_1, a_2 > 0$;

(2) [10pts] Assume $f(x) = \exp\{-\frac{\|x-\mu\|^2}{2\sigma^2}\}$ where μ and σ are real const. And K_4 is defined by $K_4 = f(X)^T f(X), where f(X) = [f(x_1), f(x_2), \cdots, f(x_N)]$;

(3) [10pts] $K_5 = K_1 \cdot K_2$ where '$\cdot$'means Kronecker product.

Solution

(1) By the definition, we denote that $K_{1ij} = k_1(x_i, x_j)$ and $K_{2ij} = k_2(x_i, x_j)$.
We know that K_1 and K_2 are positive semi-definite so there exists a vector $C = \{c_1, \cdots, c_n\}^T$ which makes

$$C^T K_1 C = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k_1(x_i, x_j) \geq 0,$$

$$C^T K_2 C = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k_2(x_i, x_j) \geq 0$$

We also know that $K_3 = a_1K_1 + a_2K_2$,
so we get

$$C^T K_3 C = C^T(a_1K_1 + a_2K_2)C = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j (a_1 k_1(x_i, x_j) + a_2 k_2(x_i, x_j)) \geq 0$$

(2) We know that $K_4 = f(X)^T f(X)$
So we get

$$C^T f(X)^T f(X) C = (f(X)C)^T f(X)C = (\sum_{i=1}^{n} f(x_i)c_i)^2 \geq 0$$

(3) According to the following characters of Kronecker product (espically theorem 3-3), we get that $K_5 = K_1 \cdot K_2$ is also positive semi-definite.

定理3 1) 设 $A \in C^{m\times m}, B \in C^{n\times n}$,则 $\mathrm{tr}(A\otimes B) = \mathrm{tr}(A)\mathrm{tr}(B) = \mathrm{tr}(B\otimes A)$;

2) 设 $\vec{x},\vec{y}$ 分别是 $A \in C^{m\times m}, B \in C^{n\times n}$ 的特征向量,则 $\vec{x}\otimes\vec{y}$ 是 $A\otimes B$ 的特征向量;

3) 设 $A \in C^{m\times m}, B \in C^{n\times n}$ 都是(半)正定矩阵,则 $A\otimes B$ 也是(半)正定矩阵.

证 1) 由引理1,$A\otimes B$ 的所有特征值为 $\lambda_i\mu_j (i = 1,\cdots,m, j = 1,\cdots,n)$,据定义4,

$$\mathrm{tr}(A\otimes B) = \sum_i\sum_j \lambda_i\mu_j = \Big(\sum_i \lambda_i\Big)\Big(\sum_j \mu_j\Big) = \mathrm{tr}(A)\mathrm{tr}(B) = \mathrm{tr}(B)\mathrm{tr}(A).$$

2) 设 $A\vec{x} = \lambda\vec{x}, B\vec{y} = \mu\vec{y}$,由性质5及性质1得,

$$(A\otimes B)(\vec{x}\otimes\vec{y}) = (A\vec{x})\otimes(B\vec{y}) = (\lambda\vec{x})\otimes(\mu\vec{y}) = (\lambda\mu)(\vec{x}\otimes\vec{y}),$$

故 $\vec{x}\otimes\vec{y}$ 是 $A\otimes B$ 的特征向量.

3) 设 $A \in C^{m\times m}$ 的所有特征值为 $\lambda_1,\lambda_2,\cdots,\lambda_m, B \in C^{n\times n}$ 的所有特征值为 $\mu_1,\mu_2,\cdots,\mu_n$. 由于 $A \in C^{m\times m}, B \in C^{n\times n}$ 都是(半)正定矩阵,从而 $\lambda_1,\lambda_2,\cdots,\lambda_m$ 及 $\mu_1,\mu_2,\cdots,\mu_n$ 全是(非负数)正数,由引理1,$A\otimes B$ 的所有特征值 $\lambda_i\mu_j$ 也是(非负数)正数,故 $A\otimes B$ 也是(半)正定矩阵.

☑ Problem 2 (SVM with Weighted Penalty)

Consider the standard SVM optimization problem as follows (i.e., formula (6.35)in book),

$$\min_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, 2, \cdots, m.$$

Note that in (2.1), for positive and negative examples, the "penalty"of the classification error in the objective function is the same. In the real scenario, the price of "punishment"is different for misclassifying positive and negative examples. For example, considering cancer diagnosis, misclassifying a person who actually has cancer as a healthy person, and misclassifying a healthy person as having cancer, the wrong influence and the cost should not be considered equivalent.

Now, we want to apply $k > 0$ to the "penalty"of the examples that were split in the positive case for the examples with negative classification results (i.e., false positive). For such scenario,

(1) [10pts] Please give the corresponding SVM optimization problem;

(2) [15pts] Please give the corresponding dual problem and detailed derivation steps, especially such as KKT conditions.

✎ Solution

(1) If misclassified case y_i is positive, punishment is ξ_i.

However if misclassified case y_i is negative, punishment is $k\xi_i$.

So the SVM optimization problem will be

$$\min_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\frac{(k+1) - y_i(k-1)}{2}\xi_i$$

$$\text{s.t. } y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0, i = 1, 2, \cdots, m.$$

(2) Using the method of lagrange multipliers, we get that function:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\frac{(k+1) - y_i(k-1)}{2}\xi_i + \sum_{i=1}^{m}\alpha_i(1 - \xi_i - y_i(w^T x_i + b)) - \sum_{i=1}^{m}\mu_i\xi_i$$

The derivation of w is

$$w - \sum_{i=1}^{m}\alpha_i y_i x_i$$

The derivation of b is

$$-\sum_{i=1}^{m}\alpha_i y_i$$

The derivation of ξ_i is

$$C\frac{(k+1)-y_i(k-1)}{2}-\alpha_i-\mu_i$$

That is to say

$$\sum_{i=1}^{m}\alpha_i y_i=0$$

$$\sum_{i=1}^{m}\alpha_i y_i x_i=w$$

$$C\frac{(k+1)-y_i(k-1)}{2}=\alpha_i+\mu_i$$

Then the dual problem is

$$\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j$$

$$\text{s.t. }\sum_{i=1}^{m}\alpha_i y_i=0,$$

$$0\le\alpha_i\le C\frac{(k+1)-y_i(k-1)}{2}, i=1,2,\cdots,m.$$

And the KKT conditions are

1 raw posibility : $\alpha_i\ge 0,\ \mu_i\ge 0$

2 dual posibility : $y_i(w^T x_i+b)-1+\xi_i\ge 0$

3 complementary slackness condition: $\xi_i\ge 0,\ \mu_i\xi_i=0$

4 Lagrangian stationarity: $\alpha_i(y_i(w^T x_i+b)-1+\xi_i)=0$

☑ Problem 3 (Nearest Neighbor)

Let $D=\{x_1,\cdots,x_n\}$ be a set of instances sampled completely at random from a p-dimensional unit ball B centered at the origin,

$$B=\{x:\|x\|^2\le 1\}\subset\mathbb{R}^p$$

Here, $\|x\|=\sqrt{\langle x,x\rangle}$ and $\langle\cdot,\cdot\rangle$ indicates the dot product of two vectors.
In this assignment, we consider to find the nearest neighbor for the origin. That is, we define the shortest distance between the origin and D as follows,

$$d^*:=\min_{1\le i\le n}\|x_i\|.$$

It can be seen that d^* is a random variable since x_i , $\forall i, 1\le i\le n$ are sampled completely at random.

(1) [5pts] Assume $p=2$ and $t\in[0,1]$, calculate $Pr(d^*\le t)$, i.e., the cumulative distribution function (CDF) of random variable d^*.

(2) [10pts] Show the general formula of CDF of random variable d^* for $p\in\{1,2,3,\cdots\}$. You may need to use the volume formula of sphere with radius equals to r,

$$V_p(r)=\frac{(r\sqrt{\pi})^p}{\Gamma(p/2+1)}.$$

Here, $\Gamma(1/2)=\sqrt{\pi},\Gamma(1)=1$, and $\Gamma(x+1)=x\Gamma(x),\forall x>0$. For $n\in\mathbb{N}^*$, $\Gamma(n+1)=n!$

(3) [10pts] Calculate the median of the value of random variable d^*, i.e., calculate the value of t that satisfies $Pr(d^*\le t)=1/2$.

Solution

(1) for one ball, the distance shorter than t is $\frac{\pi t^2}{\pi} = t^2$.
Then we assume that all the distance is bigger than t, the probability is $(1-t^2)^n$.
So $Pr(d^* \leq t) = 1-(1-t^2)^n$.

(2) for one ball, the distance shorter than t is $\frac{Vp(t)}{Vp(1)} = \frac{(r\sqrt{\pi})^t}{\Gamma(t/2+1)} = t^p$.
Then we assume that all the distance is bigger than t, the probability is $(1-t^p)^n$.
So $Pr(d^* \leq t) = 1-(1-t^p)^n$.

(3) According to the probability we get from (2).
$Pr(d^* \leq t) = 1-(1-t^p)^n = \frac{1}{2} \Rightarrow 1-t^p = 2^{-\frac{1}{n}} \Rightarrow t = (1-2^{-\frac{1}{n}})^{\frac{1}{p}}$

Problem 4 (Principal Component Analysis)

(1) [5 pts] Please describe describe the similarities and differences between PCA and LDA.

(2) [10 pts] Consider 3 data points in the 2-d space: $(-1,1),(0,0),(1,1)$, What is the first principal component? (Maybe you don't really need to solve any SVD or eigenproblem to see this.)

(3) [10 pts] If we projected the data into 1-d subspace, what are their new corrdinates?

Solution

(1) Similatiries:

a. Both PCA and LDA are classical dimensionality reduction algorithms.

b. Both PCA and LDA assume that the data conforms to the gaussian distribution.

c. Both PCA and LDA make use of the idea of matrix eigendecomposition.

Differences:

a. PCA is unsupervised and LDA is supervised.

b. PCA is to remove the redundant dimension of the original data, while LDA is to select an optimal projection direction, so that the data of the same category are distributed compactness after projection, and the data of different categories are as far away from each other as possible.

c. At most, LDA can be reduced to k-1 dimension (k is the number of training samples, k-1 is because the mean of the last dimension can be represented by the mean of the previous k-1 dimension).

d. LDA may overfit the data.

(2) As we can see $(1,0)$ is the first principal component.

(3) Their new corrdinates are as follows:

$$(-1,1) \rightarrow -1$$

$$(0,0) \rightarrow 0$$

$$(1,1) \rightarrow 1$$

第二部分　订正

第三部分　反馈