机器学习导论作业四

161220097 戚赟

2019 年 4 月 29 日

1 [25pts] Kernel Methods

From Mercer theorem, we know a two variables function $k(\cdot,\cdot)$ is a positive definite kernel function if and only if for any N vectors $x_1, x_2, ..., x_N$, 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 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 $f(X) = [f(x_1), f(x_2), ..., f(x_N)]$;
(3) [10pts] $K_5 = K_1 \cdot K_2$ where '$\cdot$' means Kronecker product.

Solution: (1) 证明: 由于 K_1 和 K_2 皆为半正定矩阵，所以根据定义，对于任意的实非零列向量x，则有 $x^T K_1 x \geq 0$，则对于任意的实非零列向量$\beta \in R^n$，有

$$\beta^T K_3 \beta = \beta^T (a_1K_1 + a_2K_2)\beta = a_1\beta^T K_1\beta + a_2\beta^T K_2\beta \tag{1.1}$$

$$\because a_1 > 0, a_2 > 0, \beta^T K_1\beta \geq 0, \beta^T K_2\beta > 0 \tag{1.2}$$

$$\therefore \beta^T K_3\beta \geq 0 \tag{1.3}$$

所以K_3也为半正定矩阵，所以可以知道根据Mercer定理，与矩阵相关联的函数是positive definite的.

(2) 由于$f(x) = \exp\{-\frac{\|x-\mu\|^2}{2\sigma^2}\}$,所以 $f(x_i)$得到的是实数。
所以,对于任意的实非零列向量$\beta \in R^n$

$$\beta^T f(X)^T f(X)\beta = (f(X)\beta)^T (f(X)\beta) \tag{1.4}$$

$$\because f(X) = [f(x_1), f(x_2), \ldots, f(x_N)] \tag{1.5}$$

$$\therefore f(X)\beta \in R^n \tag{1.6}$$

$$\therefore (f(X)\beta)^T (f(X)\beta) \geq 0 \tag{1.7}$$

所以K_4也为半正定矩阵，根据Mercer定理，与矩阵相关联的函数是positive definite的.

(3)
证明方法一：利用特征值进行证明
设K_1和K_2的维度分别为m和n，假设$\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$为$K_1$的特征值，$\mu_1, \ldots, \mu_n$为$K_2$的特征值，则因为是半正定矩阵，所以可以知道$K_1$和$K_2$的特征值都是非负的.
假设x是K_1的相对于特征值λ_i的一个特征向量，y是K_2的相对于特征值μ_j的一个特征向量, 则设新列向量V，其中V的大小为$m \times n$,$V = \{x_1y_1, x_1y_2, \ldots, x_1y_n, x_2y_1, \ldots, x_my_n\}$.下面证明$V$是$K_5 = K_1 \cdot K_2$的一个特征向量。

$$\because K_5 = K_1 \cdot K_2 \tag{1.8}$$

$$= \begin{bmatrix} k_1(x_1,x_1)K_2 & k_1(x_1,x_2)K_2 & \ldots & k_1(x_1,x_m)K_2 \\ k_1(x_2,x_1)K_2 & k_1(x_2,x_2)K_2 & \ldots & k_1(x_2,x_m)K_2 \\ \cdots & \cdots & \cdots & \cdots \\ k_1(x_m,x_1)K_2 & k_1(x_m,x_1)K_2 & \ldots & k_1(x_m,x_m)K_2 \end{bmatrix} \tag{1.9}$$

$$\therefore K_5 V = \lambda_i \mu_j V \tag{1.10}$$

所以知道$\lambda_i\mu_j (i \in [1,m], j \in [1,n])$为$K_5$的特征值，共有$m \times n$个，而由于$K_1, K_2$是可对角化的，所以不可能存在比$m \times n$还多的特征值，而$\lambda_i \geq 0, \lambda_j \geq 0$，所以$K_5$的所有特征值都大于等于0，所以$K_5$也是半正定矩阵。根据Mercer定理，与矩阵相关联的函数是positive definite的.

证明方法二：利用Kronecker product的乘法性质

$(A_1A_2)\cdot(B_1B_2) = (A_1\cdot B_1)(A_2\cdot B_2)$
由于K_1和K_2是半正定矩阵，所以可以进行对角化$K_1 = Q^T D_1 Q, K_2 = P^T D_2 P$.
所以利用两次性质得

$$K_5 = K_1 \cdot K_2 = (Q^T D_1 Q)\cdot(P^T D_2 P) = (Q\cdot P)^T (D_1\cdot D_2)(Q\cdot P) \quad (1.11)$$

而且$D_1 \cdot D_2$是对角阵，而且每个元素都大于等于0，所以K_5是半正定矩阵。
所以根据Mercer定理，与矩阵相关联的函数是positive definite的.

2 [25pts] SVM with Weighted Penalty

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

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \cdots, m. \end{aligned} \tag{2.1}$$

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)因为我们要对于False positive的例子要施加一个penalty，其值为k ($k > 0$)，所以，我们添加的优化目标为

$$\min_{\mathbf{w},b,\xi_i} \quad \frac{1}{2}\|\mathbf{w}\|^2 + C(\sum_{i\in\mathbf{P}}^{m}\xi_i + \sum_{i\in\mathbf{N}}^{m}k\xi_j) \tag{2.2}$$

$$\text{s.t.} \quad y_i(\mathbf{w}x_i + b) \geq 1 - \xi_i; \tag{2.3}$$

$$\xi_i \geq 0; for\ i\ =\ 1,\ldots,m; \tag{2.4}$$

(2)令$\alpha, \mu(\alpha \geq 0, \mu \geq 0)$为拉格朗日乘子，则

$$L(\mathbf{w},b,\xi,\alpha,\mu) = \frac{1}{2}\|\mathbf{w}\|^2 + C(\sum_{i\in\mathbf{P}}^{m}\xi_i + \sum_{i\in\mathbf{N}}^{m}k\xi_j) + \tag{2.5}$$

$$\sum_{i=1}^{m}\alpha_i(1 - \xi_i - y_i(\mathbf{w}^T\mathbf{x_i} + b)) - \sum_{i=1}^{m}\mu_i\xi_i \tag{2.6}$$

定义一个指示函数$I(y_i)$，当$y_i = 1, I(y_i) = 1$，否则为0
对于$\mathbf{w}, b, \xi$求偏导且令$\Delta = 0$可得到如下的等式:

$$\mathbf{w} = \sum_{i=1}^{m}a_i y_i x_i \tag{2.7}$$

$$\sum_{i=1}^{m}\alpha_i y_i = 0 \tag{2.8}$$

$$C = (\alpha_i + \mu_i)\left[I(y_i \in P) + \frac{1}{k}I(y_i \in N)\right] \tag{2.9}$$

将2.7,2.8,2.9带入原式 (2.5)可以得到对偶问题为

$$\max_{\alpha} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \mathbf{x_i}^T \mathbf{x_i} \tag{2.10}$$

$$s.t. \quad 0 \leq \alpha_i \leq C(I(y_i \in P) + \frac{1}{k} I(y_i \in N)) \tag{2.11}$$

$$\sum_{i=1}^{m} \alpha_i y_i = 0 \tag{2.12}$$

对应的KKT条件为：

$$\begin{cases} \alpha_i \geq 0; \mu_i \geq 0 \\ \xi_i - 1 + y_i f(x_i) \geq 0 \\ \alpha_i (1 - \xi_i - y_i f(x_i)) = 0 \\ \xi_i \geq 0, \mu_i \xi_i = 0 \end{cases} \tag{2.13}$$

3 [25pts] Nearest Neighbor

Let $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of instances sampled completely at random from a p-dimensional unit ball B centered at the origin,

$$B = \left\{ \mathbf{x} : \|\mathbf{x}\|^2 \leq 1 \right\} \subset \mathbb{R}^p. \tag{3.1}$$

Here, $||\mathbf{x}|| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ 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 $\mathcal{D}$ as follows,

$$d^* := \min_{1 \leq i \leq n} \|\mathbf{x}_i\| . \tag{3.2}$$

It can be seen that d^* is a random variable since $\mathbf{x}_i, \forall 1 \leq i \leq n$ are sampled completely at random.

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

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

$$V_p(r) = \frac{(r\sqrt{\pi})^p}{\Gamma(p/2 + 1)}. \tag{3.3}$$

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 $\Pr(d^* \leq t) = 1/2$.

Solution:

(1) 当$p = 2$，$B = \{\mathbf{x} : \|\mathbf{w}\|^2 \geq 1\} \subset R^2$.
所代表的就是以原点为圆心，半径为1的圆内的点.

$$Pr(d^* \leq t) = 1 - Pr(d^* > t) = 1 - Pr(\min_{1 \leq i \leq n} \|X_i\|) \tag{3.4}$$

$$= 1 - \prod_{i=1}^{n} Pr(\|X_i\| > t) \tag{3.5}$$

$$\because Pr(\|X_i\| > t) = 1 - \frac{\pi t^2}{\pi \times 1^2} = 1 - t^2 \tag{3.6}$$

$$\therefore Pr(d^* \leq t) = 1 - (1 - t^2)^n \tag{3.7}$$

(2) 和(1)中过程类似.
当$t < 0$时, $Pr(d^* \leq t) = 0$.
当$t > 1$时, $Pr(d^* \leq t) = 1$.
当$t \in [0, 1]$,

$$Pr(d^* \leq t) = 1 - Pr(d^* > t) = 1 - Pr(\min_{1 \leq i \leq n} \|X_i\|) \tag{3.8}$$

$$= 1 - \prod_{i=1}^{n} Pr(\|X_i\| > t) = 1 - (1 - \frac{V_p(t)}{V_p(1)})^n \tag{3.9}$$

$$= 1 - (1 - t^p)^n \tag{3.10}$$

所以CDF为

$$Pr(d^* \leq t) = \begin{cases} 0, t < 0 \\ 1 - (1 - t^p)^n, t \in [0, 1] \\ 1, t > 1 \end{cases} \tag{3.11}$$

(3)0 可知，$t \in [0, 1]$,所以

$$1 - (1 - t^p)^n = \frac{1}{2} \tag{3.12}$$

$$(1 - t^p)^n = \frac{1}{2} \tag{3.13}$$

$$t = (1 - (\frac{1}{2})^{\frac{1}{n}})^{\frac{1}{p}} \tag{3.14}$$

4 [25pts] 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) 一、PCA与LDA的比较
相同点：
1. 两者均可以对数据进行降维
2. 两者在降维时均使用了矩阵特征分解的思想

3. 两者都假设数据符合高斯分布

不同点：

1. LDA是有监督的降维方法，而PCA是无监督的降维方法

2. LDA降维最多降到类别数k-1的维数，而PCA没有这个限制

3. LDA除了可以用于降维，还可以用于分类

4. LDA选择分类性能最好的投影方向，而PCA选择样本点投影具有最大方差的方向

(2)首先进行归一化，由于一个属性和为0，不需要进行归一化，只需要对第二个属性进行归一化，则归一化的结果为

$$\begin{bmatrix} -1 & 0 & 1 \\ -\frac{1}{3} & \frac{2}{2} & -\frac{1}{3} \end{bmatrix} \tag{4.1}$$

计算协方差矩阵，如下

$$\frac{1}{3}\begin{bmatrix} -1 & 0 & 1 \\ -\frac{1}{3} & \frac{2}{3} & -\frac{1}{3} \end{bmatrix}\begin{bmatrix} -1 & -\frac{1}{3} \\ 0 & \frac{2}{3} \\ 1 & -\frac{1}{3} \end{bmatrix} = \begin{bmatrix} \frac{2}{3} & 0 \\ 0 & \frac{2}{9} \end{bmatrix} \tag{4.2}$$

由于结果是对角阵，明显可以知道第一个属性的特征值最大，所以选择第一个属性作为主成成份.

(3)最大特征值为$\frac{2}{3}$，对应的一个单位特征向量为$[1 \quad 0]$.

所以根据这向量投影的结果为$[-1 \quad 0 \quad 1]$.

投影的直线也就是X轴.