1 机器学习导论

姓名: 殷天润 学号:171240565

题目 (ML problem 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 '·' means Kronecker product.

解答:

1. 对任意的非零列向量 X:$X^TK_3X = a_1(X^TK_1X) + a_2(X^TK_2X)$;

   因为 K_1, K_2 都是 positive definite, 所以 $X^TK_1X, X^TK_2X > 0$, 而 $a_1, a_2 > 0$; 所以 $X^TK_3X > 0$

2. 对于任意的非零列向量 X, 假设 $X_T = \{a_1, a_2, ..., a_n\}$, 所以

   $X^TK_4X = (X^Tf(x)^T)(Xf(x))$=$a_1^2f(x_1)^2 + a_2^2f(x_2)^2 + ... + a_n^2f(x_n)^2$;

   因为 $f(x_n)^2 = \exp\{-\frac{\|x-\mu\|^2}{2\sigma^2}\}^2 > 0$, 所以 $X^TK_4X > 0$, 是正定的;

3. 关于 Kronecker 积的引理:

   (1)$(A \cdot B)$ 每一个特征值可表示为 A 与 B 的特征值之积即:

   $\lambda(A) = \{\lambda_1, ..., \lambda_n\}$
   $\lambda(B) = \{\mu_1, ..., \mu_m\}$
   $\lambda(A \cdot B) = \{\lambda_i\mu_j, i = 1, ..., n; j = 1, ..., m\}$[1]

   因为 K_1, K_2 都是 positive definite, 所以它们的特征值 λ_1, λ_2 都是正的;

   所以:$\lambda(K_1 \cdot K_2) = \lambda(K_1) \cdot \lambda(K_2)$, 因此 K_5 的特征值是正的, 所以 K_5 是正定矩阵;

注: [1]: 宋乾坤. 复正定矩阵的一些性质 [J]. 四川师范大学学报: 自然科学版,1997,20(3):44-48

题目 (ML problem 2)

[25pts] SVM with Weighted Penalty

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

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

Note that in (7), 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.

解答:

1. 注意到 $y_i = \{-1, 1\}$, 所以:

$$f(y_i) = \frac{(k+1) + y_i(k-1)}{2} = \begin{cases} 1 & y_i = -1 \\ k & y_i = +1 \end{cases}$$

因此问题可以转化为:

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

2. 使用拉格朗日乘子法:

$$\begin{aligned} L(w, b, \alpha, \xi, \mu) = & \frac{1}{2}||w||^2 + C * \frac{(k+1) + y_i(k-1)}{2} * \sum_{i=1}^{m}\xi_i \\ & + \sum_{i=1}^{m}\alpha_i(1 - \xi_i - y_i(w^T w_i + b)) - \sum_{i=1}^{m}\mu_i\xi_i \end{aligned} \tag{3}$$

其中 α_i, μ_i 是拉格朗日乘子;

令 $L(w, b, \alpha, \xi, \mu)$ 对 Lw, α, ξ_i 偏导置为 0 可得:

$$w = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{4}$$

$$0 = \sum_{i=1}^{m} \alpha_i y_i \tag{5}$$

$$C * \frac{(k+1) + y_i(k-1)}{2} = \alpha_i + \mu_i \tag{6}$$

代入可以得到对偶问题:

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j \\ &\text{s.t.} \sum_{i=1}^{m} \alpha_i y_i \\ &\quad 0 \leq \alpha_i \leq C * \frac{(k+1) + y_i(k-1)}{2}, i = 1, 2, \cdots, m. \end{aligned} \tag{7}$$

题目 (ML problem 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,

$$B = \left\{\mathbf{x} : \|\mathbf{x}\|^2 \leq 1\right\} \subset \mathbb{R}^p. \tag{8}$$

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,

$$d^* := \min_{1 \leq i \leq n} \|\mathbf{x}_i\|. \tag{9}$$

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 $p \in \{1, 2, 3, \ldots\}$. You may need to use the volume formula of sphere with radius equals to r,

$$V_p(r) = \frac{(r\sqrt{\pi})^p}{\Gamma(p/2 + 1)}. \tag{10}$$

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

解答：

(1)

$$\begin{aligned} Pr(d^* \le t) &= P(min_{1\le i\le n}(||x_i|| \le t) \\ &= 1 - P(min_{1\le i\le n}(||x_i|| > t) \\ &= 1 - \Pi_{i=1}^{n} P(||x_i|| > t) \end{aligned} \tag{11}$$

而因为是二维,$P(||x_i|| > t) = 1 - P(||x_i|| < t) = 1 - t^2$

$$\begin{aligned} Pr(d^* \le t) &= 1 - \Pi_{i=1}^{n}(1 - t^2) \\ &= 1 - (1 - t^2)^n \end{aligned} \tag{12}$$

(2) 由第一问:$Pr(d^* \le t) = 1 - \Pi_{i=1}^{n} P(||x_i|| > t)$

而 $P(||x_i|| > t) = 1 - P(||x_i|| < t)$

对于 p 维向量 x,$P(||x_i|| < t) = V_p(t)/V_p(1) = \frac{(t\sqrt{\pi})^p}{(\sqrt{\pi})^p} = t^p$

所以类似的

$$\begin{aligned} Pr(d^* \le t) &= 1 - \Pi_{i=1}^{n}(1 - t^p) \\ &= 1 - (1 - t^p)^n \end{aligned} \tag{13}$$

(3) 代入计算：

$$\begin{aligned} 1 - (1 - t^p)^n &= \frac{1}{2} \\ t^p &= 1 - (\frac{1}{2})^{\frac{1}{n}} \\ t &= (1 - (\frac{1}{2})^{\frac{1}{n}})^{\frac{1}{p}} \end{aligned} \tag{14}$$

题目 (ML problem 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.)

(2) [10 pts] If we projected the data into 1-d subspace, what are their new corrdinates?

解答：

(1) :

(a) 相同: 都是降维的方法, 可以把原来的 N 维数据降成 K 维;

(b) 不同:

i. 出发思想不同。PCA 主要是从特征的协方差角度，去找到比较好的投影方式，即选择样本点投影具有最大方差的方向; 而 LDA 则更多的是考虑了分类标签信息，寻求投影后不同类别之间数据点距离更大化以及同一类别数据点距离最小化，即选择分类性能最好的方向。

ii. 降维后可用维度数量不同。LDA 降维后最多可生成 C-1 维子空间（分类标签数-1），因此 LDA 与原始维度 N 数量无关，只有数据标签分类数量有关；而 PCA 最多有 n 维度可用，即最大可以选择全部可用维度。

iii. 习模式不同。PCA 属于无监督式学习，因此大多场景下只作为数据处理过程的一部分，需要与其他算法结合使用，例如将 PCA 与聚类、判别分析、回归分析等组合使用；LDA 是一种监督式学习方法，本身除了可以降维外，还可以进行预测应用，因此既可以组合其他模型一起使用，也可以独立使用。

(2) 每一个点的第一个量, 也就是 $\{-1, 0, 1\}$, 因为它们“拉”的比较开

(3)

$$DATA = \begin{bmatrix} -1 & 1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}, \tag{15}$$

$$DATAadjust = \begin{bmatrix} -1 & \frac{1}{3} \\ 0 & -\frac{-2}{3} \\ 1 & \frac{1}{3} \end{bmatrix}, \tag{16}$$

求解特征协方差矩阵：

协方差矩阵:

$$C = \begin{bmatrix} cov(x,x) = 1 & con(x,y) = 0 \\ cov(y,x) = 0 & cov(y,y) = \frac{1}{3} \end{bmatrix}, \tag{17}$$

特征值:$1, \frac{1}{3}$

特征向量:$(1,0)^T$,$(0,1)^T$

因此最大的特征值是 1, 相应的特征矩阵是:$(1,0)^T$

因此最后的结果:

$$FinalData = DATAadjust(3 \times 2) * (1,0)^T = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \tag{18}$$

注： 第一题参考了 https://blog.csdn.net/dongyanwen6036/article/details/78311071
PCA:https://zhuanlan.zhihu.com/p/21580949
https://blog.csdn.net/zhongkelee/article/details/44064401