机器学习导论
作业四

学号, 作者姓名, 邮箱

2018 年 5 月 29 日

1 [30pts] Kernel Methods

Mercer定理告诉我们对于一个二元函数$k(\cdot,\cdot)$，它是正定核函数当且仅当对任意N和$\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N$，它对应的核矩阵是半正定的. 假设$k_1(\cdot,\cdot)$和$k_2(\cdot,\cdot)$分别是关于核矩阵$K_1$和$K_2$的正定核函数. 另外，核矩阵$K$中的元素为$K_{ij}=k(\mathbf{x}_i,\mathbf{x}_j)$. 请根据Mercer定理证明对应于以下核矩阵的核函数正定.

(1) **[10pts]** $K_3=a_1K_1+a_2K_2$, 其中$a_1,a_2\geq 0$.

(2) **[10pts]** $f(\cdot)$是任意实值函数，由$k_4(\mathbf{x},\mathbf{x}')=f(\mathbf{x})f(\mathbf{x}')$定义的$K_4$.

(3) **[10pts]** 由$k_5(\mathbf{x},\mathbf{x}')=k_1(\mathbf{x},\mathbf{x}')k_2(\mathbf{x},\mathbf{x}')$定义的$K_5$.

Proof. 此处用于写证明*(中英文均可)*

(1) K_3的第ij个元素为$a_1k_1(\mathrm{x}_i,\mathrm{x}_j)+a_2k_2(\mathbf{x}_i,\mathbf{x}_j)$. 因为$K_1$,$K_2$正定，对任意的$\eta_1$,有${\eta_1}^{\mathrm{T}}K_1\eta_1>0$,${\eta_1}^{\mathrm{T}}K_2\eta_1>0$,所以对任意的$\eta_1$,且$a_1>0$,$a_2>0$,有${\eta_1}^{\mathrm{T}}K_3=a_1{\eta_1}^{\mathrm{T}}K_1\eta_1+a_2{\eta_1}^{\mathrm{T}}K_2\eta_1>0$,得证.

(2) K_4的第ij个元素为$f(x_i)(x_j)$,所以

$$K_4=(f(\mathbf{x}_1),f(\mathbf{x}_2),\ldots,f(\mathbf{x}_n))^{\mathrm{T}}(f(\mathbf{x}_1),f(\mathbf{x}_2),\ldots,f(\mathbf{x}_n))$$

对于任意的η_1，

$$\begin{aligned}{\eta_1}^{\mathrm{T}}K_4\eta_1&={\eta_1}^{\mathrm{T}}(f(\mathbf{x}_1),f(\mathbf{x}_2),\ldots,f(\mathbf{x}_n))^{\mathrm{T}}(f(\mathbf{x}_1),f(\mathbf{x}_2),\ldots,f(\mathbf{x}_n))\eta_1\\&=b^{\mathrm{T}}b,\end{aligned}$$

其中$b=(f(\mathbf{x}_1),f(\mathbf{x}_2),\ldots,f(\mathbf{x}_n))\eta_1$.上式显然不小于*0*，得证.

(3) 这个题目等价于证明A，B两个矩阵半正定，则A，B对应元素相乘所得矩阵也半正定。具体证明过程参见$Schur$乘积定理。

2 [25pts] SVM with Weighted Penalty

考虑标准的SVM优化问题如下(即课本公式(6.35))，

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

注意到，在(2.1)中，对于正例和负例，其在目标函数中分类错误的“惩罚”是相同的. 在实际场景中，很多时候正例和负例错分的“惩罚”代价是不同的. 比如考虑癌症诊断问题，将一个确实患有癌症的人误分类为健康人，以及将健康人误分类为患有癌症，产生的错误影响以及代价不应该认为是等同的.

现在，我们希望对负例分类错误的样本(即false positive)施加$k > 0$倍于正例中被分错的样本的“惩罚”. 对于此类场景下，

(1) **[10pts]** 请给出相应的SVM优化问题.

(2) **[15pts]** 请给出相应的对偶问题，要求详细的推导步骤，尤其是如KKT条件等.

Solution. 此处用于写解答(中英文均可)

(1) 考虑所有正例样本的下标集合为$\mathcal{P}$以及负例样本的下标集合为$\mathcal{N}$，根据题干中的要求，我们只需要对负例分类错误的样本施加$k > 0$倍于正例样本被分错得到的“惩罚”即可。因此，我们可以得到如下的优化目标

$$\begin{aligned}&\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + C\left(\sum_{i\in\mathcal{P}}^{m}\xi_i + k\cdot\sum_{i\in\mathcal{N}}^{m}\xi_i\right) \\ &s.t.\ y_i(\mathbf{w}\mathbf{x}_i + b) \geq 1 - \xi_i, \\ &\qquad \xi_i \geq 0, \quad for\ i = 1, \cdots, m.\end{aligned} \tag{2.2}$$

(2) 记α, μ表示拉格朗日乘子，则

$$\begin{aligned}L(\mathbf{w}, b, \xi, \alpha, \mu) = & \frac{1}{2}\|\mathbf{w}\|^2 + +C\left(\sum_{i\in\mathcal{P}}^{m}\xi_i + k\cdot\sum_{i\in\mathcal{N}}^{m}\xi_i\right) \\ & + \sum_{i=1}^{m}\alpha_i(1 - \xi_i - y_i(\mathbf{w}\mathbf{x}_i + b)) - \sum_{i=1}^{m}\mu_i\xi_i.\end{aligned} \tag{2.3}$$

令$\nabla_{\mathbf{w}}L = \nabla_b L = \nabla_{\xi_i}L = 0$，则有

$$\begin{cases}\mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i \\ 0 = \sum_{i=1}^{m}\alpha_i y_i \\ C = (\alpha_i + \mu_i)\cdot\left(\frac{1}{k}\mathbb{I}(i\in\mathcal{P}) + \mathbb{I}(i\in\mathcal{N})\right)\end{cases} \tag{2.4}$$

其中，$\mathbb{I}(\cdot)$为示性函数*(indicator function)*，当$\cdot$为真时取值为*1*，否则取值为*0.*

于是，我们可以得到对偶问题如下：

$$
\begin{aligned}
&\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i\alpha_j y_i y_j \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j) \\
&s.t.\ \sum_{i=1}^{m} y_i\alpha_i = 0 \\
&\qquad 0 \le \alpha_i \le C\cdot(k\mathbb{I}(i\in\mathcal{P}) + \mathbb{I}(i\in\mathcal{N}))
\end{aligned}
\tag{2.5}
$$

因此，可以得到KKT条件如下：

$$
\begin{cases}
\alpha_i, \mu_i, \xi_i \ge 0 \\
\xi_i - 1 + y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \ge 0 \\
\alpha_i(1 - \xi_i - y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b)) = 0 \\
\mu_i\xi_i = 0.
\end{cases}
\tag{2.6}
$$

3 [30pts+10*pts] Nearest Neighbor

假设数据集$\{\mathbf{x}_1, ..., \mathbf{x}_n\}$是从一个以$\mathbf{0}$为中心的$p$维单位球中独立均匀采样而得到的$n$个样本点. 这个球可以表示为：

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

其中，$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$，$\langle \mathbf{x}, \mathbf{x} \rangle$是$\mathbb{R}^p$空间中向量的内积. 在本题中，我们将探究原点$O$与其最近邻(1-NN)的距离$d^*$，以及这个距离$d^*$与$p$之间的关系. 在这里，我们将原点$O$以及其1-NN之间的距离定义为:

$$d^* \coloneqq \min_{1 \leq i \leq n} \|\mathbf{x}_i\|, \tag{3.2}$$

不难发现d^*是一个随机变量，因为$\mathbf{x}_i$是随机产生的.

(1) **[5pts]** 当$p = 1$且$t \in [0, 1]$时，请计算$\Pr(d^* \leq t)$，即随机变量d^*的累积分布函数(Cumulative Distribution Function, **CDF**).

(2) **[10pts]** 请写出d^*的**CDF**的一般公式，即当$p \in \{1, 2, 3, ...\}$时d^*对应的取值. 提示：半径为r的p维球体积是：

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

其中，$\Gamma(1/2) = \sqrt{\pi}$，$\Gamma(1) = 1$，且有$\Gamma(x + 1) = x\Gamma(x)$对所有的$x > 0$成立；并且对于$n \in \mathbb{N}^*$，有$\Gamma(n + 1) = n!$.

(3) **[10pts]** 请求解随机变量d^*的中位数，即使得$\Pr(d^* \leq t) = 1/2$成立时的t值. 答案是与n和p相关的函数.

(4) **[附加题10pts]** 请通过**CDF**计算使得原点O距其最近邻的距离d^*小于1/2的概率至少0.9的样本数n的大小. 提示：答案仅与p相关. 你可能会用到$\ln(1 - x)$的泰勒展开式：

$$\ln(1 - x) = -\sum_{i=1}^{\infty} \frac{x^i}{i}, \quad \text{for } -1 \leq x < 1. \tag{3.4}$$

(5) **[5pts]** 在解决了以上问题后，你关于n和p以及它们对1-NN的性能影响有什么理解.

Solution. *此处用于写解答(中英文均可)*

(1) 当$p = 1$时，单位球退化为区间$[-1, 1]$. 那么此时的$\mathcal{CDF}$就有如下表示：

$$F_{n,1}(t) = \Pr(d^* \leq t) = 1 - \Pr(d^* > t) = 1 - \Pr(\|x_i\| > t, \quad for\ i = 1, 2, \dots, n). \tag{3.5}$$

因为$\mathbf{x}_1, \dots, \mathbf{x}_n$是独立的，进而$\mathcal{CDF}$就可以写成：

$$F_{n,1}(t) = 1 - \prod_{i=1}^{n} \Pr(\|\mathbf{x}_i\| > t) = 1 - (1 - t)^n. \tag{3.6}$$

(2) 在一般情况下，我们不妨假设$p \in \mathbb{N}^*$. 那么很明显，$\mathcal{CDF}$也会有类似的表达形式：

$$\begin{aligned} F_{n,p}(t) = \Pr(d^* \leq t) &= 1 - \Pr(d^* > t) \\ &= 1 - \Pr(\|\mathbf{x}_i\| > t, i = 1, 2, \dots, n) \\ &= 1 - \prod_{i=1}^{n} \Pr(\|\mathbf{x}_i\| > t). \end{aligned} \tag{3.7}$$

我们将半径为t的球体体积记为$V_p(t)$，又因为$\mathbf{x}_i$服从均匀分布，我们便可以把上式(3.7)改写为：

$$F_{n,p}(t) = 1 - (\frac{V_p(1) - V_p(t)}{V_p(1)})^n = 1 - (1 - \frac{V_p(t)}{V_p(1)})^n. \tag{3.8}$$

显然，最终可以得到$F_{n,p} = 1 - (1 - t^p)^n$.

(3) 要找d^*的中间值，我们只需要对t求解等式$\Pr(d^* \leq t) = 1/2$:

$$\begin{aligned} P(d^* \leq t) = \frac{1}{2} &\Leftrightarrow F_{n,p}(t) = \frac{1}{2} \\ &\Leftrightarrow 1 - (1 - t^p)^n = \frac{1}{2} \Leftrightarrow (1 - t^p)^n = \frac{1}{2} \\ &\Leftrightarrow 1 - t^p = \frac{1}{2^{1/n}} \Leftrightarrow t^p = 1 - \frac{1}{2^{1/n}}. \end{aligned} \tag{3.9}$$

因此，$t_{med}(n,p) = (1 - \frac{1}{2^{1/n}})^{1/p}$.

*(4)[**附加题10pts**]* 基于上面的结果，我们有：

$$\begin{aligned} \Pr(d^* \leq 0.5) \geq 0.9 &\Leftrightarrow F_{n,p}(0.5) \geq 0.9 \\ &\Leftrightarrow 1 - (1 - \frac{1}{2^p})^n \geq 0.9 \\ &\Leftrightarrow 1 - (1 - \frac{1}{2^p})^n \leq 0.1 \\ &\Leftrightarrow n \cdot \ln(1 - \frac{1}{2^p}) \leq -\ln 10 \\ &\Leftrightarrow \frac{\ln 10}{-\ln(1 - \frac{1}{2^p})}. \end{aligned} \tag{3.10}$$

用泰勒公式展开$\ln(1 - 1/2^p)$，此时$x = 1/2^p$:

$$\begin{aligned} &\Pr(d^* \leq 0.5) \geq 0.9 \Rightarrow \\ &n \geq (\ln 10) 2^p \frac{1}{1 + \frac{1}{2} \cdot \frac{1}{2^p} + \frac{1}{3} \cdot \frac{1}{2^2 p} + \cdots + \frac{1}{n} \cdot \frac{1}{2^(n-1)p} + \cdots} \\ &\quad \Rightarrow n \geq 2^p - 1 \ln 10. \end{aligned} \tag{3.11}$$

进一步，我们知道：对任意的$p \geq 1$，都有$\frac{1}{3 \cdot 2^p} < \frac{1}{4}$成立；并且，$\frac{1}{n \cdot 2^(n-1)p} \leq \frac{1}{2^n}$与$2^n \leq n \cdot 2^(n-1)p$在$p \geq 1$以及$n \geq 2$时成立. 因此，可以得到：

$$\begin{aligned} 1 + \frac{1}{2} \cdot \frac{1}{2^p} + \frac{1}{3} \cdot \frac{1}{2^2 p} + \cdots + \frac{1}{n} \cdot \frac{1}{2^(n-1)p} + \cdots < \\ 1 + \frac{1}{2} + \frac{1}{4} + \cdots + \frac{1}{2^n} + \cdots \to \frac{1}{1 - 1/2} = 2. \end{aligned} \tag{3.12}$$

因此，$\Pr(d^* \leq 0.5) \geq 0.9 \Rightarrow n \geq 2^p - 1 \ln 10$.

(5) 酌情给分.

4 [15pts] Principal Component Analysis

一些经典的降维方法，例如PCA，可以将均值为**0**的高维数据通过对其协方差矩阵的特征值计算，取较高特征值对应的特征向量的操作而后转化为维数较低的数据. 在这里，我们记U_k为$d \times k$的矩阵，这个矩阵是由原数据协方差矩阵最高的k个特征值对应的特征向量组成的.

在这里我们有两种方法来求出低维的对应于$\mathbf{x} \in \mathbb{R}^d$的重构向量$\mathbf{w} \in \mathbb{R}^k$:

A. 利用$U_k\mathbf{w}$重构出对应的$\mathbf{x}$时, 最小化重构平方误差;

B. 将$\mathbf{x}$投影在由U_k的列向量张成的空间中.

在这里，我们将探究这两种方法的关系.

(1) **[5pts]** 写出方法A中最小化重构平方误差的目标函数的表示形式.

(2) **[10pts]** 证明通过方法A得到的重构向量就是$U_k^{\mathrm{T}}\mathbf{x}$，也就是$\mathbf{x}$在$U_k$列向量空间中的投影(通过方法B得到的重构向量). 这里, 有$U_k^{\mathrm{T}}U_k = I_k$成立, 其中的$I_k$是$k \times k$的单位矩阵.

Solution. *此处用于写解答(中英文均可)*

(1) 目标方程是$\|U_k\mathbf{w} - \mathbf{x}_i\|^2$, 因此表示形式是$\min\limits_{\mathbf{w}}\|U_k\mathbf{w} - \mathbf{x}_i\|^2$.

(2) 对于一般情况, 我们求解$\min\limits_{\mathbf{x}}\|A\mathbf{x} - \mathbf{y}\|^2$时, 所得到的最优解是$\mathbf{x}^* = (A^{\mathrm{T}}A)^{-1}A^{\mathrm{T}}\mathbf{y}$, 当且仅当$A$是满秩的.

在本题中U_k是满秩的, 带入一般最优解的表达式中可以得到本题中$\mathbf{w}$的最优解是$\mathbf{w}^* = (U_k^{\mathrm{T}}U_k)^{-1}U_k^{\mathrm{T}}\mathbf{x}_i$.

又因为$U_k^{\mathrm{T}}U_k = I_k$, 因此$\mathbf{w}^* = U_k\mathbf{x}_i$.