机器学习导论
作业四

151220097, 孙旭东, 248381185@qq.com

2018 年 5 月 19 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$K_5$.

Solution.

(1) 要证明K_3对应的核函数正定，即证明K_3半正定。
由于K_1和K_2对应的核函数是正定核函数，所以K_1和K_2都是半正定矩阵。所以有：对任意非零向量$\mathbf{x}$，都有：

$$\mathbf{x}^{\mathrm{T}}K_1\mathbf{x}\geq 0 \tag{1.1}$$

$$\mathbf{x}^{\mathrm{T}}K_2\mathbf{x}\geq 0 \tag{1.2}$$

又因为$a_1,a_2\geq 0$，所以:

$$\mathbf{x}^{\mathrm{T}}a_1K_1\mathbf{x}\geq 0 \tag{1.3}$$

$$\mathbf{x}^{\mathrm{T}}a_2K_2\mathbf{x}\geq 0 \tag{1.4}$$

那么可以得到:

$$\mathbf{x}^{\mathrm{T}}K_3\mathbf{x}=\mathbf{x}^{\mathrm{T}}(a_1K_1+a_2K_2)\mathbf{x}=\mathbf{x}^{\mathrm{T}}a_1K_1\mathbf{x}+\mathbf{x}^{\mathrm{T}}a_2K_2\mathbf{x}\geq 0 \tag{1.5}$$

所以可得K_3是半正定矩阵，所以对应的核函数正定。

(2) 设K_{ij}表示K_4的i行，第j列的项。对于任意非零向量$\mathbf{x}$，用x_i表示$\mathbf{x}$的第i个项，可以有：

$$\mathbf{x}^{\mathrm{T}}K_4\mathbf{x}=\sum_{i=1}^{N}\sum_{j=1}^{N}x_iK_{ij}x_j \tag{1.6}$$

又因为$k_4(\mathbf{x}_i,\mathbf{x}_j)=f(\mathbf{x}_i)f(\mathbf{x}_j)$，所以为了方便证明令$t_i=f(\mathbf{x}_i)$，可得$K_{ij}=k_4(\mathbf{x}_i,\mathbf{x}_j)=f(\mathbf{x}_i)f(\mathbf{x}_j)=t_it_j$，结合上式可得：

$$\begin{aligned}\mathbf{x}^{\mathrm{T}}K_4\mathbf{x}=&\sum_{i=1}^{N}\sum_{j=1}^{N}x_iK_{ij}x_j=\sum_{i=1}^{N}\sum_{j=1}^{N}x_it_it_jx_j\\=&\sum_{i=1}^{N}(x_it_i)^2+2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}x_it_ix_jt_j\\=&(t_1x_1+t_2x_2+...+t_nx_n)^2\geq 0\end{aligned} \tag{1.7}$$

所以可得K_4是半正定矩阵，所以对应的核函数正定。

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) 定义$\mathbf{P}$表示正例的下标的集合，$\mathbf{N}$表示负例的下标的集合，相应的SVM优化问题为：

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

(2) 根据上式(2.2)，使用拉格朗日乘子法可得拉格朗日函数如下：

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

其中$\alpha_i \geq 0$，$\mu_i \geq 0$是拉格朗日乘子。

令$L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu})$对$\mathbf{w}$，$b$，$\xi_i$的偏导为零可得

$$
\mathbf{w} = \sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i \tag{2.4}
$$

$$
0 = \sum_{i=1}^{m}\alpha_i y_i \tag{2.5}
$$

$$
C = (\alpha_i + \mu_i)(\mathbb{I}(i \in \mathbf{P}) + \frac{1}{k}\mathbb{I}(i \in \mathbf{N})) \tag{2.6}
$$

为了方便解答，这里用到了指示器函数，定义如下：

$$\mathbb{I}(\cdot)=\begin{cases}1 & \cdot \textit{ is true}\\ 0 & \cdot \textit{ is false}\end{cases} \tag{2.7}$$

现在将(2.4) − (2.6)代入式(2.3)即可得到式(2.2)的对偶问题

$$\begin{aligned}\max_{\alpha} \quad & \sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^{\mathrm{T}}\mathbf{x}_j\\ s.t. \quad & \sum_{i=1}^{m}\alpha_i y_i=0\\ & 0\le\alpha_i\le, i=1,2,\cdots,m.\end{aligned} \tag{2.8}$$

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：

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

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:

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

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

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

(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)*

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. 求最小重构平方误差;

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$U_k,\mathbf{x}$以及$\mathbf{w}$的最小二乘形式 (方法A).

(2) **[10pts]** 证明方法A的解就是$U_k^{\mathrm{T}}\mathbf{x}$，也就是$\mathbf{x}$在$U_k$列向量空间中的投影 (方法B).

Solution. *此处用于写解答(中英文均可)*