机器学习导论习题课

第四次作业

第一题：核方法

根据Mercer定理，一个函数$k(\cdot,\cdot)$是正定核函数的充要条件是对于N个变量$x_1, x_2, \ldots, x_N$，他们的核矩阵是半正定的。假设$k_1(\cdot,\cdot)$和$k_2(\cdot,\cdot)$是正定核函数，他们对应的核矩阵分别为K_1和K_2，核矩阵的每个元素定义为$K_{ij} = k(x_i, x_j)$，请证明一下对应以下核矩阵的核函数是正定核函数

(1) $K_3 = a_1 K_1 + a_2 K_2$，其中$a_1, a_2 > 0$。

(2) 假设$f(x) = \exp\{-\frac{(x-\mu)^2}{\sigma}\}$，其中$\mu$和$\sigma$都是实常数。 $K_4 = f(X)^T f(X)$，其中$f(x) = [f(x_1), f(x_2), \ldots, f(x_N)]$。

(3) $K_5 = K_1 \otimes K_2$。

第一题：核方法

(1) 解：假设N维列向量β

$$\beta^{\mathrm{T}}K_3\beta=\beta^{\mathrm{T}}(a_1K_1+a_2K_2)\beta=a_1\beta^{\mathrm{T}}K_1\beta+a_2\beta^{\mathrm{T}}K_2\beta\geq 0$$

所以K_3半正定，核函数是正定核函数。

(2) 解：假设N维列向量β

$$\beta^{\mathrm{T}}K_4\beta=[f(X)\beta]^T f(X)\beta=(f(X)\beta)^2\geq 0$$

所以K_4半正定，核函数是正定核函数。

第一题：核方法

(3) 解：假设N^2维列向量$\mathrm{B}=[B_1;B_2;\ldots;B_N]$，其中$B_i=[B_{i1};B_2;\ldots;B_{iN}]$都是$N$维列向量

$$B^T K_5 B = B^T(K_1 \otimes K_2)B = B^T \begin{pmatrix} k_{11}^1 K_2 & \cdots & k_{1N}^1 K_2 \\ \vdots & \ddots & \vdots \\ k_{N1}^1 K_2 & \cdots & k_{NN}^1 K_2 \end{pmatrix} B$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N} k_{ij}^1 B_i^T K_2 B_j = \sum_{i=1}^{N}\sum_{j=1}^{N} k_{ij}^1 B_i^T K_2^{1/2} K_2^{T/2} B_j$$

令$M_j = K_2^{T/2} B_j$，这是一个$N \times 1$的列向量。

第一题：核方法

$$\sum_{i=1}^{N}\sum_{j=1}^{N} k_{ij}^{1} B_i^T K_2^{1/2} K_2^{T/2} B_j = \sum_{i=1}^{N}\sum_{j=1}^{N} k_{ij}^{1} M_i^T M_j = \sum_{i=1}^{N}\sum_{j=1}^{N} k_{ij}^{1} \sum_{t=1}^{N} M_{it} M_{jt}$$

$$= \sum_{t=1}^{N}\sum_{i=1}^{N}\sum_{j=1}^{N} k_{ij}^{1} M_{it} M_{jt}$$

令$M' = \begin{pmatrix} M'_1 = (M_{11}; \cdots; M_{N1}) \\ \vdots \\ M'_N = (M_{1N}; \cdots; M_{NN}) \end{pmatrix}_{N^2\times 1}$，其中$M'_t$是$N$维列向量。

核方法

$$\sum_{t=1}^{N}\sum_{i=1}^{N}\sum_{j=1}^{N} k_{ij}^{1} M_{it}M_{jt} = \sum_{t=1}^{N} M'^{T}_{t} K_1 M'_{t} \geq 0$$

所以K_5半正定，核函数是正定核函数。

第二题：SVM

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，

第二题： SVM

(1) 请给出相应的SVM优化问题.

解：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，可以得到如下的优化目标：

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

P表示所有正例样本的下标集合

N表示负例样本的下标集合

第二题： SVM

(2) 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.

解：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·为真时取值为*1*，否则取值为*0*.

第二题： SVM

Learning And Mining from DatA
http://lamda.nju.edu.cn

我们可以得到对偶问题如下：

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

KKT条件如下：

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

第三题：最近邻

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:

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

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

第三题：最近邻

(1) 当 p = 2 且 t ∈ [0, 1] 时，请计算 Pr($\mathbf{d}^*$ ⩽ t)，即随机变量 $\mathbf{d}^*$ 的累积分布函数 (Cumulative Distribution Function, CDF)

解：当 p = 2 时，单位球退化为单位圆. 那么此时的 CDF 就有如下表示：

$$F_{n,2}(t) = \Pr\left(d^* \leq t\right) = 1 - \Pr\left(d^* > t\right) = 1 - \Pr\left(\|x_i\| > t, \quad \text{for } i = 1, 2, \ldots, n\right)$$

对于一次独立采样，我们有 $Pr\left(d \leq t\right) = \dfrac{\pi t^2}{\pi (1)^2} = t^2$

因为 x_1，. . . ，x_n 相互独立，进而 CDF 可以写成：

$$F_{n,2}(t) = 1 - \prod_{i=1}^{n} \Pr\left(\|\mathbf{x}_i\| > t\right) = 1 - (1 - t^2)^n$$

第三题：最近邻

(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$\Gamma(n+1) = n!$.

解：我们将半径为 t 的球体体积记为 $V_p(t)$，又因为x服从均匀分布，所以有：

$$\begin{aligned} F_{n,p}(t) &= \Pr\left(d^* \le t\right) = 1 - \Pr\left(d^* > t\right) \\ &= 1 - \left(\frac{V_p(1) - V_p(t)}{V_p(1)}\right)^n = 1 - \left(1 - \frac{V_p(t)}{V_p(1)}\right)^n = 1 - (1 - t^p)^n \end{aligned}$$

第三题：最近邻

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

解：要找 d^* 的中位数，我们只需要对 t 求解等式 $\Pr(d^* \leqslant t)$

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

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

第四题：主成分分析

- 1. 请描述PCA和LDA的异同

此题可以描述的点很多，言之有理即可得分。

答：PCA和LDA都是常见的线性降维的算法，从PCA和LDA的求解过程来看其有很大的相似性，但对应的原理其实有所区别。

PCA是一种无监督算法，选择的是投影后数据方差最大的方向，因此，PCA假设方差越大，信息量越多，用主成分来表示原始数据可以去除冗余的维度。而LDA是一种有监督算法，选择的是投影后类内方差小，类间方差大的方向。所以，PCA是保留的最佳描述特征而LDA是保留的分类特征。

举个简单的🌰，在语音识别中，如果想从一段音频中提取出人的语音信号，可以使用PCA进行降维，过滤掉一些固定频率(方差较小)的背景噪声。但是如果任务需求是区分出声音属于那个人，那应该使用LDA，使每个人的信号具有区分性。

求解主成分和新坐标

- 2. 给定三个数据点(-1，1),(0,0),(1,1)，求解第一主成分。

答：首先对数据进行中心化，得到

$$x_1 = \left(-1, \frac{1}{3}\right), x_2 = \left(0, -\frac{2}{3}\right), x_1 = \left(1, \frac{1}{3}\right)$$

计算样本的协方差矩阵， $\mathrm{XX}^T = \begin{pmatrix} 1 & 0 \\ 0 & \frac{2}{3} \end{pmatrix}$。

然后对协方差矩阵做特征值分解 $\mathrm{XX}^T\,W = \lambda W$。$\lambda_1 = 1\,,\lambda_2 = \frac{2}{3}$ 取最大特征值$\lambda_1 = 1$，

特征向量为$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$。所以第一主成分为$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$。

- 3. 在一维空间下的新坐标

答：新坐标：(-1，1)-> -1，(0，0) -> 0，(1，1)->1。