机器学习导论
作业四

151250189, 翟道京, zhaidj@smail.nju.edu.cn

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.

Solution.

(1) 对于正定核函数 $k_1(\cdot,\cdot)$ 和 $k_2(\cdot,\cdot)$,首先根据 *Mercer* 定理,考虑对任意数据 $\mathbf{D} = \{\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}\}$, 核矩阵 K_1 和 K_2 都是半正定的。

接着根据非负实数与半正定矩阵的数乘矩阵是半正定的，对 $a_1, a_2 \geq 0$，a_1K_1 与 a_2K_2 都是半正定的。

然后根据两个半正定矩阵的和是半正定的，得知 $K_3 = a_1K_1 + a_2K_2$ 是半正定的。

最后使用 *Mercer* 定理，K_3 的核函数正定。

(2) 这里 K_4 可以表示为

$$K_4 = \begin{bmatrix} f(\mathbf{x}_1)f(\mathbf{x}_1) & f(\mathbf{x}_1)f(\mathbf{x}_2) & f(\mathbf{x}_1)f(\mathbf{x}_3) & \ldots & f(\mathbf{x}_1)f(\mathbf{x}_n) \\ f(\mathbf{x}_2)f(\mathbf{x}_1) & f(\mathbf{x}_2)f(\mathbf{x}_2) & f(\mathbf{x}_2)f(\mathbf{x}_3) & \ldots & f(\mathbf{x}_2)f(\mathbf{x}_n) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ f(\mathbf{x}_n)f(\mathbf{x}_1) & f(\mathbf{x}_n)f(\mathbf{x}_2) & f(\mathbf{x}_n)f(\mathbf{x}_3) & \ldots & f(\mathbf{x}_n)f(\mathbf{x}_n) \end{bmatrix}$$

对该实矩阵 K_4 进行分解可以得到

$$K_4 = \begin{bmatrix} f(\mathbf{x}_1) & 0 & 0 & \ldots & 0 \\ f(\mathbf{x}_2) & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f(\mathbf{x}_n) & 0 & 0 & \ldots & 0 \end{bmatrix} * \begin{bmatrix} f(\mathbf{x}_1) & f(\mathbf{x}_2) & f(\mathbf{x}_3) & \ldots & f(\mathbf{x}_n) \\ 0 & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 0 \end{bmatrix} := AA^T$$

因此对任意 $\mathbf{x} \in \mathbb{R}^n$，有

$$\mathbf{x}^T K_4 \mathbf{x} = \mathbf{x}^T A A^T \mathbf{x} = (A^T \mathbf{x})^T (A^T \mathbf{x}) \geq 0$$

故 K_4 为半正定矩阵，根据 *Mercer* 定理，K_4 的核函数正定。
(3) 这里 K_5 可以表示为

$$K_5 = \begin{bmatrix} k_1(\mathbf{x_1}, \mathbf{x_1}) k_2(\mathbf{x_1}, \mathbf{x_1}) & k_1(\mathbf{x_1}, \mathbf{x_2}) k_2(\mathbf{x_1}, \mathbf{x_2}) & k_1(\mathbf{x_1}, \mathbf{x_3}) k_2(\mathbf{x_1}, \mathbf{x_3}) & \dots & k_1(\mathbf{x_1}, \mathbf{x_n}) k_2(\mathbf{x_1}, \mathbf{x_n}) \\ k_1(\mathbf{x_2}, \mathbf{x_1}) k_2(\mathbf{x_2}, \mathbf{x_1}) & k_1(\mathbf{x_2}, \mathbf{x_2}) k_2(\mathbf{x_2}, \mathbf{x_2}) & k_1(\mathbf{x_2}, \mathbf{x_3}) k_2(\mathbf{x_2}, \mathbf{x_3}) & \dots & k_1(\mathbf{x_2}, \mathbf{x_n}) k_2(\mathbf{x_2}, \mathbf{x_n}) \\ \dots\dots & \dots\dots & \dots\dots & \dots & \dots\dots \\ k_1(\mathbf{x_n}, \mathbf{x_1}) k_2(\mathbf{x_n}, \mathbf{x_1}) & k_1(\mathbf{x_n}, \mathbf{x_2}) k_2(\mathbf{x_n}, \mathbf{x_2}) & k_1(\mathbf{x_n}, \mathbf{x_3}) k_2(\mathbf{x_n}, \mathbf{x_3}) & \dots & k_1(\mathbf{x_n}, \mathbf{x_n}) k_2(\mathbf{x_n}, \mathbf{x_n}) \end{bmatrix}$$

其中 $K_5 = K_1 \circ K_2$，对对称半正定矩阵 K_1 和 K_2 进行特征值分解得

$$K_1 = \sum_{i=1}^{n} \lambda_i \mathbf{u_i} \mathbf{u_i}^T$$

$$K_2 = \sum_{j=1}^{n} \mu_j \mathbf{v_j} \mathbf{v_j}^T$$

其中特征值 $\lambda_i \geq 0$ 且 $\mu_j \geq 0$，因此

$$\begin{aligned} K_5 &= \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \mu_j (\mathbf{u_i} \mathbf{u_i}^T) \circ (\mathbf{v_j} \mathbf{v_j}^T) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \mu_j (\mathbf{u_i} \mathbf{v_i}) \circ (\mathbf{u_i} \mathbf{v_i})^T \\ &= \sum_{k=1}^{n^2} \gamma_k \mathbf{w_k} \mathbf{w_k}^T \end{aligned}$$

其中 $\gamma_k = \lambda_{\lfloor k/n \rfloor} \mu_{k\ mod\ n} \geq 0$ 且 $\mathbf{w_k} = \mathbf{u}_{\lfloor k/n \rfloor} \mathbf{v}_{k\ mod\ n}$，因此对任意 $\mathbf{x} \in \mathbb{R}^n$，有

$$\mathbf{x}^T K_5 \mathbf{x} = \sum_{k=1}^{n^2} \gamma_k \mathbf{x}^T \mathbf{w_k} \mathbf{w_k}^T \mathbf{x} = \sum_{k=1}^{n^2} \gamma_k (\mathbf{w_k}^T \mathbf{x})^T (\mathbf{w_k}^T \mathbf{x}) \geq 0$$

故 K_5 为半正定矩阵，根据 *Mercer* 定理，K_5 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(**??**)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.

(1) 在此种情况下，SVM 优化问题为

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

(2) 使用拉格朗日乘子法得到拉格朗日函数为

$$
\begin{aligned}
L(\mathbf{w},b,\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\mu}) = & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{\{i|y_i=1\}}^{m_+}\xi_i + kC\sum_{\{j|y_j=-1\}}^{m_-}\xi_j \\
& + \sum_{i=1}^{m}\alpha_i(1-\xi_i-y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i+b)) - \sum_{i=1}^{m}\mu_i\xi_i.
\end{aligned}
$$

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

KKT 条件要求

$$
\begin{cases}
\alpha_i \geq 0, \mu_i \geq 0, \\
y_i f(\mathbf{x}_i) - 1 + \xi_i \geq 0, \\
\alpha_i(y_i f(\mathbf{x}_i) - 1 + \xi_i) = 0, \\
\xi_i \geq 0, \mu_i\xi_i = 0.
\end{cases} \tag{2.3}
$$

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

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

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

$$C = \alpha_i + \mu_i, i \in m_+ \tag{2.6}$$

$$kC = \alpha_j + \mu_j, j \in m_- \tag{2.7}$$

我们将 *(2.3)-(2.6)* 代入 $L(\mathbf{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu})$，得到对偶问题

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \mathbf{x_i}^T \mathbf{x_j} \\
&s.t. \sum_{i=1}^{m} \alpha_i y_i = 0 \\
&0 \le \alpha_i \le C, i \in m_+ \\
&0 \le \alpha_j \le kC, j \in m_-
\end{aligned} \tag{2.8}$$

可以采用核方法求解此对偶问题。对于该软间隔支持向量机，引入核函数后得到和书上 *(6.24)* 式子相同的支持向量展式

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i y_i \kappa(\mathbf{x}, \mathbf{x_i}) + b \tag{2.9}$$

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.
(1) 当 $p = 1$ 时, 在 $\mathbb{R}^1$ 空间中, $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = |x|$, 为求解随机变量 d^* 的累积分布函数, 我们转化为下列问题: 在 $[0, 1]$ 区间内独立随机撒 n 个样本点, 由于数据集 $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ 是独立均匀采样得到的, 每个样本点等概率出现在样本空间 $[0, 1]$ 中, 记所有点都出现在 $[t, 1]$ 区间内的概率为 $p(t)$, 则

$$\Pr(d^* \leq t) = 1 - p(t) = 1 - (1 - t)^n$$

(2) 考虑 **CDF** 的一般公式, 根据 *(1)* 中的想法, 用 t 表示距原点距离, 则

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

$$V_p(t) = \frac{(t\sqrt{\pi})^p}{\Gamma(p/2+1)}$$

根据 *(1)* 的想法，记在半径为 *1* 的区域内独立均匀撒 n 个样本点，均落在 $[t,1]$ 半径范围内的概率为 $p(t)$，则

$$\Pr(d^* \le t) = 1 - p(t) = 1 - \left(\frac{V_p(1) - V_p(t)}{V_p(1)}\right)^n$$

代入公式得

$$\Pr(d^* \le t) = 1 - (1 - t^p)^n$$

(3) 令

$$\Pr(d^* \le t) = 1 - (1 - t^p)^n = \frac{1}{2}$$

得

$$t = \sqrt[p]{1 - \sqrt[n]{\frac{1}{2}}}$$

(4) 要求

$$\Pr(d^* \le \frac{1}{2}) = 1 - (1 - (\frac{1}{2})^p)^n \ge 0.9$$

即

$$\begin{aligned}(1 - (\frac{1}{2})^p)^n \le 0.1 &\Rightarrow n \ln(1 - (\frac{1}{2})^p) \le \ln\frac{1}{10} \\ &\Rightarrow n \ge -\frac{\ln 10}{\ln(1 - (\frac{1}{2})^p)} \\ Taylor\ Series &\Rightarrow n \ge \frac{\ln 10}{\sum_{i=1}^{\infty} \frac{1}{i \cdot 2^{pi}}}\end{aligned}$$

其中我们对一定维度进行数值分析：

$p = 10$ 时，$n \ge 2357$

$p = 15$ 时，$n \ge 75450$

$p = 20$ 时，$n \ge 2414434$

(5) 思考与总结

(i) 在 p 维单位球下进行独立均匀采样，为保证密采样，需要确定较高的采样数。

(ii) 1-NN 采样会收到维数灾难的影响，为保持一定的密采样，采样数 n 随 p 的增长近乎成指数增长。*(下证)*

证明. 以 (4) 题中情形为例，在 p 维下，为保证题目中所示的密采样，选择采样数值为

$$n(p) = \frac{\ln 10}{\sum_{i=1}^{\infty} \frac{1}{i \cdot 2^{pi}}}$$

在 p+1 维下

$$
\begin{aligned}
n(p+1) &= \frac{\ln 10}{\sum_{i=1}^{\infty} \frac{1}{i \cdot 2^{(p+1)i}}} \\
&= \frac{\ln 10}{\sum_{i=1}^{\infty} \frac{1}{i \cdot 2^{pi} \cdot 2^{i}}} \\
&\geq \frac{\ln 10}{\sum_{i=1}^{\infty} \frac{1}{i \cdot 2^{pi} \cdot 2}} \\
&= \frac{2 \ln 10}{\sum_{i=1}^{\infty} \frac{1}{i \cdot 2^{pi}}} \\
&= 2n(p)
\end{aligned}
$$

因此采样数 n 随 p 的增长近乎成指数增长，即

$$n(p+1) \geq 2n(p)$$

$\square$

4 [15pts] Principal Component Analysis

一些经典的降维方法，例如 PCA，可以将均值为 $\mathbf{0}$ 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.

(1) 最小化重构平方误差的目标函数的表示形式为

$$\min_{\mathbf{w}} \|U_k\mathbf{w} - \mathbf{x}\|^2 = (U_k\mathbf{w} - \mathbf{x})^T(U_k\mathbf{w} - \mathbf{x}) \tag{4.1}$$

(2) 通过求解 (4.1) 式，证明

$$\mathbf{w} = U_k^T\mathbf{x} \tag{4.2}$$

证明. 根据 (4.1) 式，对范数平方进行展开，原问题转化为

$$\min_{\mathbf{w}} \|U_k\mathbf{w} - \mathbf{x}\|^2 = (U_k\mathbf{w} - \mathbf{x})^T(U_k\mathbf{w} - \mathbf{x}) \tag{4.3}$$

$$\Rightarrow \min_{\mathbf{w}} \mathbf{w}^T U_k^T U_k \mathbf{w} - 2U_k^T\mathbf{x}\mathbf{w} \tag{4.4}$$

其中考虑 $U_k^{\mathrm{T}}U_k = I_k$，(4.4) 式转化为

$$\min_{\mathbf{w}} \mathbf{w}^T\mathbf{w} - 2U_k^T\mathbf{x}\mathbf{w} \tag{4.5}$$

对 (4.5) 式关于 $\mathbf{w}$ 求偏导并置为零

$$2\mathbf{w} - 2U_k^T\mathbf{x} = 0$$

因此

$$\mathbf{w} = U_k^T\mathbf{x} \tag{4.6}$$

□