# Introduction to Machine Learning
# Homework 4

151250048, 郭浩滨, 151250048@smail.nju.edu.cn

April 30, 2019

## 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 $\mu$ and $\sigma$ 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 '$\cdot$' means Kronecker product.

(1) 由于 $K_1$、$K_2$ 是正定矩阵，故满足对于任意 N 维非零向量 $x$，有

$$x^TK_1x > 0, \quad x^TK_2x > 0$$

成立。故

$$\begin{aligned} x^TK_3x &= x^T(a_1K_1 + a_2K_2)x \\ &= x^Ta_1K_1x + x^Ta_2K_2x \\ &= a_1x^TK_1x + a_2x^TK_2x \\ &> 0 \end{aligned} \tag{1.1}$$

因此 $K_3$ 是正定矩阵。

(2) 由题意可得 $K_{4_{ij}} = f(x_i)f(x_j)$，故对于任意 N 维非零向量 $z$，有

$$\begin{aligned} z^T K_4 z &= \sum_{i=1}^{N}\sum_{j=1}^{N} z_i f(x_i) f(x_j) z_j \\ &= \left(\sum_{i=1}^{N} z_i f(x_i)\right)^2 \\ &\geq 0 \end{aligned} \tag{1.2}$$

可得 $K_4$ 是半正定矩阵。

(3) 由于 $K_1$ 和 $K_2$ 均为正定矩阵，故满足：$K_1 = P_1^T D_1 P_1$，$K_2 = P_2^T D_2 P_2$，其中 $D_1$、$D_2$ 为实对角矩阵，且两者特征值均大于 0。

根据克罗内克积的混合乘积性质，可得：

$$\begin{aligned} K_5 &= K_1 \cdot K_2 \\ &= (P_1^T D_1 P_1) \cdot (P_2^T D_2 P_2) \\ &= (P_1 \cdot P_2)^T (D_1 \cdot D_2)(P_1 \cdot P_2) \end{aligned} \tag{1.3}$$

易得 $D_1 \cdot D_2$ 仍为实对角矩阵，且特征值均大于 0。因此对于任意 $N^2$ 维非零向量 $z$，有：

$$\begin{aligned} z^T K_5 z &= z^T (K_1 \cdot K_2) z \\ &= z^T (P_1 \cdot P_2)^T (D_1 \cdot D_2)(P_1 \cdot P_2) z \\ &= ((P_1 \cdot P_2) z)^T (D_1 \cdot D_2)((P_1 \cdot P_2) z) \\ &> 0 \end{aligned} \tag{1.4}$$

因此 $K_5$ 为正定矩阵。

## 2 [25pts] SVM with Weighted Penalty

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

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

Note that in (2.1), 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.

(1) 由题意可知我们需要对正例样本中被分错的样本对应的 $\xi$ 乘以权值 $k$，设所有正例样本的下标组成的集合为 $P$，所有负例样本的下标组成的集合为 $N$，则对应的 SVM 优化问题可以表述为：

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

(2) 由上述 SVM 问题，通过拉格朗日乘子法可得到以下拉格朗日函数：

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

令 $L(w,b,\alpha,\xi,\mu)$ 对 $w,b,\xi_i$ 的偏导数为零，可得：

$$\begin{aligned}&w = \sum_{i=1}^{m}\alpha_i y_i x_i, \\ &0 = \sum_{i=1}^{m}\alpha_i y_i, \\ &C = (\alpha + \mu_i)(\mathbb{I}(i\in P)\frac{1}{k} + \mathbb{I}(i\in N))\end{aligned} \tag{2.4}$$

将以上结果代入拉格朗日函数，可得到对偶问题：

$$\begin{aligned}\max_{\alpha} \quad & \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j \\ \text{s.t.} \quad & \sum_{i=1}^{m}\alpha_i y_i = 0 \\ & 0 \le \alpha_i \le C(\mathbb{I}(i\in P)k + \mathbb{I}(i\in N))\end{aligned} \tag{2.5}$$

故可得 KKT 条件如下：

$$\begin{cases}\alpha_i \ge 0, \quad \mu_i \ge 0 \\ y_i(w^T x_i + b) - 1 + \xi_i \ge 0 \\ \alpha_i(y_i(w^T x_i + b) - 1 + \xi_i) = 0 \\ \xi_i \ge 0, \quad \mu_i\xi_i = 0\end{cases} \tag{2.6}$$

## 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,

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

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,

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

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 $r$,

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

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 $\Pr\left(d^* \leq t\right) = 1/2$.

(1) 由于 $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$ 相互独立，因此

$$Pr(d^* \leq t) = 1 - Pr(d^* > t) = 1 - (1 - t^2)^n \tag{3.4}$$

即 CDF 为 $F(t) = 1 - (1 - t^2)^n$

(2) 由题意可得：$\frac{V_p(t)}{V_p(1)} = t^p$，与 (1) 同理，可得：

$$F(t) = Pr(d^* \leq t) = 1 - Pr(d^* > t) = 1 - \left(1 - \frac{V_p(t)}{V_p(1)}\right)^n = 1 - (1 - t^p)^n \quad (3.5)$$

(3) 因为 $Pr(d^* \leq t) = 1 - (1 - t^p)^n$，结合题意可得：

$$1 - (1 - t^p)^n = \frac{1}{2} \quad (3.6)$$

易解得 $t = (1 - (\frac{1}{2})^{\frac{1}{n}})^{\frac{1}{p}}$

## 4 [25pts] Principal Component Analysis

(1) [5 pts] Please 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 cordinates?

(1) 相似点：PCA 与 LDA 都是样本数据进行降维的手段，一定程度上实现对样本数据的分离。

不同点：

1. PCA 不需要了解样本的类别信息，本质上是一种无监督学习；而 LDA 则相反，是一种监督式学习。
2. PCA 的优化目标是使得所有样本点方差最大化；而 LDA 优化目标是所有样本中同类样例投影点尽可能接近，异类样本投影点尽可能远离。
3. PCA 往往需要先对数据进行中心化处理。
4. LDA 不仅用于降维，还可以用于数据分类任务中。

(2) 首先我们对数据点进行中心化处理，三个数据点的坐标平均值为 $(0, \frac{2}{3})$，故可得中心化后各数据点的坐标为：$(-1, \frac{1}{3})$，$(0, -\frac{2}{3})$，$(1, \frac{1}{3})$ 。

设投影变换后得到的新坐标系为：$(e_1, e_2)$，则有 $e_1^2 + e_2^2 = 1$，可算出数据点投影后方差为（这里用 n 而不是 n-1 作为分母，不影响最终优化结果）：

$$\begin{aligned} S &= \frac{1}{3}[(-e_1 - \frac{1}{3}e_2)^2 + (\frac{2}{3}e^2) + (e_1 - \frac{1}{3}e_2)^2] \\ &= \frac{1}{3}(2e_1^2 + \frac{2}{3}e_2^2) \\ &= \frac{1}{3}(\frac{4}{3}e_1^2 + \frac{2}{3}) \end{aligned}$$

故当 $e_1 = 1$ 时，方差 S 取得最大值。因此第一个主元为 (1, 0)。

(3) 由 (2) 中结果我们可知第一个主元为 (1, 0)，根据题意可得投影后的坐标分别为：

$$\begin{aligned} (-1*1, 1*0) &= (-1, 0), \\ (0*1, 0*0) &= (0, 0), \\ (1*1, 1*0) &= (1, 0) \end{aligned} \tag{4.1}$$

即投影后的坐标分别为：(-1, 0), (0, 0), (1, 0)