Introduction to Machine Learning
Homework 4

181860155 朱晓晴
heloize@126.com

2020 年 12 月

1 [30pts] SVM with Weighted Penalty

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

$$\begin{aligned} \min_{\mathbf{w},b,\xi_i} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \ldots, m. \end{aligned} \tag{1.1}$$

Note that in (1.1), 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) [15pts] Please give the corresponding SVM optimization problem;
(2) [15pts] Please give the corresponding dual problem and detailed derivation steps, especially such as KKT conditions.

解：(1) 定义如下两个集合：

$$P\text{：正例的下标集合}$$
$$N\text{：反例的下标集合}$$

SVM 优化问题为：

$$\begin{aligned} \min_{\boldsymbol{w},b,\xi_i} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C(\sum_{i\in P}^{m}\xi_i + k\sum_{i\in N}^{m}\xi_i) \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \ldots, m. \end{aligned} \tag{1.2}$$

(2) 通过拉格朗日乘子法可得到式 (1.2) 的拉格朗日函数

$$\begin{aligned} L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}) =& \frac{1}{2}\|\boldsymbol{w}\|^2 + C(\sum_{i\in P}^{m}\xi_i + k\sum_{i\in N}^{m}\xi_i) \\ & + \sum_{i=1}^{m}\alpha_i(1 - \xi_i - y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b)) - \sum_{i=1}^{m}\mu_i\xi_i \end{aligned} \tag{1.3}$$

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

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

$$\boldsymbol{w} = \sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i \tag{1.4}$$

$$0 = \sum_{i=1}^{m}\alpha_i y_i \tag{1.5}$$

$$C(I(i \in P) + kI(i \in N)) = \alpha_i + \mu_i \tag{1.6}$$

其中，$I(\cdot)$ 为示性函数，其定义如下

$$I(x) = \begin{cases} 1 & x\text{为真} \\ 0 & x\text{为假} \end{cases}$$

将式 (1.4)-(1.6) 代入式 (1.3) 得到

$$
\begin{aligned}
&L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}) \\
&= \frac{1}{2}\|\boldsymbol{w}\|^2 + C(\sum_{i\in P}^{m} \xi_i + k\sum_{i\in N}^{m} \xi_i) + \sum_{i=1}^{m} \alpha_i(1 - \xi_i - y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b)) - \sum_{i=1}^{m} \mu_i\xi_i \\
&= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j + C(\sum_{i\in P}^{m} \xi_i + k\sum_{i\in N}^{m} \xi_i) \\
&+ \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i\xi_i - \sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j - b\sum_{i=1}^{m} \alpha_i y_i - \sum_{i=1}^{m} \mu_i\xi_i \\
&= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j + \sum_{i\in P}^{m} (\alpha_i + \mu_i)\xi_i + k\sum_{i\in N} \frac{1}{k}(\alpha_i + \mu_i)\xi_i \\
&+ \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} (\alpha_i + \mu_i)\xi_i \\
&= \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j
\end{aligned}
\tag{1.7}
$$

KKT 条件要求

$$
\alpha_i \geq 0, \mu_i \geq 0
$$

从而有

$$
0 \leq \alpha_i \leq C(I(i \in P) + kI(i \in N))
$$

因此，相应的对偶问题为

$$
\begin{aligned}
\max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j \\
\text{s.t.} \quad & \sum_{i=1}^{m} \alpha_i y_j = 0, \\
& 0 \leq \alpha_i \leq C(I(i \in P) + kI(i \in N)), i = 1, 2, \ldots, m.
\end{aligned}
\tag{1.8}
$$

KKT 条件为

$$
\begin{cases}
\alpha_i \geq 0, \quad \mu_i \geq 0 \\
y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) - 1 + \xi_i \geq 0 \\
\alpha_i[y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) - 1 + \xi_i] = 0 \\
\xi_i \geq 0, \quad \mu_i\xi_i = 0
\end{cases}
\tag{1.9}
$$

2 [35pts] 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, i.e.,

$$B = \left\{\mathbf{x} : \|\mathbf{x}\|^2 \leq 1\right\} \subset \mathbb{R}^p. \tag{2.1}$$

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,

$$d^* := \min_{1 \leq i \leq n} \|\mathbf{x}_i\|. \tag{2.2}$$

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) [10pts] Assume $p = 3$ and $t \in [0, 1]$, calculate $\Pr(d^* \leq t)$, i.e., the cumulative distribution function (CDF) of random variable d^*.

(2) [15pts] 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 r,

$$V_p(r) = \frac{(r\sqrt{\pi})^p}{\Gamma(p/2 + 1)}. \tag{2.3}$$

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 $\Pr(d^* \leq t) = \frac{1}{2}$.

解：(1) 对于任意一次随机取样，有

$$\Pr(\|\mathbf{x}_i\| > t) = 1 - \Pr(\|\mathbf{x}_i\| \leq t) = 1 - \frac{\frac{4}{3}\pi t^3}{\frac{4}{3}\pi 1^3} = 1 - t^3 \tag{2.4}$$

因此，有

$$\begin{aligned} \Pr(d^* \leq t) &= 1 - \Pr(d^* > t) \\ &= 1 - \Pi_{i=1}^n \Pr(\|\mathbf{x}_i\| > t) \\ &= 1 - (1 - t^3)^n \end{aligned} \tag{2.5}$$

(2) 对于任意一次随机取样，有

$$\begin{aligned}\Pr(\|\mathbf{x}_i\| > t) &= 1 - \Pr(\|\mathbf{x}_i\| \leq t) \\ &= 1 - \frac{V_p(t)}{V_p(1)} \\ &= 1 - \frac{\frac{(t\sqrt{\pi})^p}{\Gamma(p/2+1)}.}{\frac{(\sqrt{\pi})^p}{\Gamma(p/2+1)}.} \\ &= 1 - t^p\end{aligned} \tag{2.6}$$

因此，有

$$\begin{aligned}\Pr(d^* \leq t) &= 1 - \Pr(d^* > t) \\ &= 1 - \Pi_{i=1}^n \Pr(\|\mathbf{x}_i\| > t) \\ &= 1 - (1 - t^p)^n\end{aligned} \tag{2.7}$$

(3) 根据题意，解出以下关于 t 的方程即可

$$\begin{aligned}&\Pr(d^* \leq t) = \frac{1}{2} \\ \Leftrightarrow\ &1 - (1 - t^p)^n = \frac{1}{2} \\ \Leftrightarrow\ &1 - t^p = \sqrt[n]{\frac{1}{2}} \\ \Leftrightarrow\ &t = \sqrt[p]{1 - \frac{1}{\sqrt[n]{2}}}\end{aligned} \tag{2.8}$$

因此，满足 $\Pr(d^* \leq t) = \frac{1}{2}$ 的 t 的取值为 $\sqrt[p]{1 - \frac{1}{\sqrt[n]{2}}}$。

3 [30pts] Principal Component Analysis

(1) [10 pts] Please describe describe the similarities and differences between PCA and LDA.
(2) [10 pts] Consider 3 data points in the 2-d space: (-2, 2), (0, 0), (2, 2), What is the first principal component? (Maybe you don't really need to solve any SVD or eigenproblem to see this.)
(3) [10 pts] If we projected the data into 1-d subspace, what are their new corrdinates?

解：(1) 同：PCA 和 LDA 都为线性降维算法；就过程而言，PCA 和 LDA 实际都是求某一矩阵的特征值，投影矩阵由特征值对应的特征向量构成。
异：PCA 为无监督算法，LDA 为有监督算法；PCA 假设方差越大，包含的信息越多，因此选择使得投影后的数据方差最大的方向作为主成分。而 LDA 则选择使得投影后类内方差小、类间方差大的方向，能合理运用标签信息，使得投影后的维度具有判别性。
(2) 首先，对所有样本进行中心化，即令

$$\boldsymbol{x}_i \leftarrow \boldsymbol{x}_i - \frac{1}{3}\sum_{i=1}^{3}\boldsymbol{x}_i \tag{3.1}$$

得到以下三个数据点

$$(-2, \frac{2}{3}), (0, -\frac{4}{3}), (2, \frac{2}{3})$$

计算样本的协方差矩阵

$$X = \begin{bmatrix} -2 & 0 & 2 \\ \frac{2}{3} & -\frac{4}{3} & \frac{2}{3} \end{bmatrix}, XX^{\mathrm{T}} = \begin{bmatrix} \frac{8}{3} & 0 \\ 0 & \frac{8}{9} \end{bmatrix}$$

对协方差矩阵做特征值分解，得到以下特征值和相应的特征向量

$$\lambda_1 = \frac{8}{3}, \quad \boldsymbol{w}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\lambda_2 = \frac{8}{9}, \quad \boldsymbol{w}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

因此，最大特征值 $\frac{8}{3}$ 相应的特征向量 $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$ 为第一主成分。
(3) 令

$$\boldsymbol{x}_i \leftarrow \boldsymbol{w}_1^{\mathrm{T}}\boldsymbol{x}_i \tag{3.2}$$

得到 $(-2, 2), (0, 0), (2, 2)$ 的新坐标分别为-2，0，2。

参考文献

[1] 周志华. 机器学习 [M]. 清华大学出版社, 2016.

[2] PCA 主成分分析
https://www.zhihu.com/question/41120789

[3] PCA 与 LDA 的比较
https://www.jianshu.com/p/982c8f6760de
https://www.zhihu.com/question/35666712