Introduction to Machine Learning
Homework 4

181860066 牛铭杨

2020 年 12 月 10 日

1 [30pts] SVM with Weighted Penalty

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

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

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;

将样本集合分为正例集合 P 和反例集合 N，满足

$$\begin{aligned} P &= \{\boldsymbol{x_i}|y_i = +1\} \\ N &= \{\boldsymbol{x_i}|y_i = -1\} \end{aligned} \tag{1.2}$$

则相应的 SVM 优化问题变为

$$\begin{aligned} \min_{\boldsymbol{w},b,\xi_i} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i\in P}\xi_i + kC\sum_{i\in N}\xi_i \\ \text{s.t.} \quad & y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, \cdots, m. \end{aligned} \tag{1.3}$$

(2) [15pts] Please give the corresponding dual problem and detailed derivation steps, especially such as KKT conditions.
通过拉格朗日乘子法，可得到上式的拉格朗日函数

$$\begin{aligned} L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\xi}, \boldsymbol{\mu}) = & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i\in P}\xi_i + kC\sum_{i\in N}\xi_i \\ & + \sum_{i=1}^{m}\alpha_i(1 - \xi_i - y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b)) - \sum_{i=1}^{m}\mu_i\xi_i \end{aligned} \tag{1.4}$$

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, $\boldsymbol{\xi}$ 的偏导为零可得

$$\boldsymbol{w} = \sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i, \tag{1.5}$$

$$0 = \sum_{i=1}^{m}\alpha_i y_i, \tag{1.6}$$

$$C = \alpha_i + \mu_i, i \in P, \tag{1.7}$$

$$kC = \alpha_i + \mu_i, i \in N. \tag{1.8}$$

将 (1.5)-(1.8) 带入拉格朗日函数就得到了对偶问题

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j \\ \text{s.t.} \quad & \sum_{i=1}^{m}\alpha_i y_i = 0, \\ & 0 \leq \alpha_i \leq C, i \in P, \\ & 0 \leq \alpha_i \leq kC, i \in N, \end{aligned} \tag{1.9}$$

而 KKT 条件要求

$$\begin{cases} \alpha_i \geq 0, \quad \mu_i \geq 0, \\ y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) - 1 + \xi_i \geq 0, \\ \alpha_i(y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b) - 1 + \xi_i) = 0, \\ \xi_i \geq 0, \quad \mu_i\xi_i = 0. \end{cases} \tag{1.10}$$

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^*.

$$\begin{aligned} \Pr(d^* \leq t) &= 1 - \Pr(d^* > t) \\ &= 1 - (\Pr(\|\mathbf{x}_i\| > t))^n \\ &= 1 - (\frac{\frac{4}{3}\pi 1^3 - \frac{4}{3}\pi t^3}{\frac{4}{3}\pi 1^3})^n \\ &= 1 - (1 - t^3)^n. \end{aligned} \tag{2.3}$$

(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,

$$V_p(r) = \frac{(r\sqrt{\pi})^p}{\Gamma(p/2 + 1)}. \tag{2.4}$$

Here, $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(1) = 1$, and $\Gamma(x + 1) = x\Gamma(x), \forall x > 0$. For

$n \in \mathbb{N}^*$, $\Gamma(n+1) = n!$.

$$\begin{aligned} \Pr(d^* \leq t) &= 1 - \Pr(d^* > t) \\ &= 1 - (\Pr(\|\mathbf{x}_i\| > t))^n \\ &= 1 - (\frac{V_p(1) - V_p(t)}{V_p(1)})^n \\ &= 1 - (1 - t^p)^n. \end{aligned} \tag{2.5}$$

(3) [10pts] Calculate the median of the value of random variable d^*, i.e., calculate the value of t that satisfies $\Pr(d^* \leq t) = \frac{1}{2}$.

$$\begin{aligned} 1 - (1 - t^p)^n &= \frac{1}{2} \\ \Rightarrow (1 - t^p)^n &= \frac{1}{2} \\ \Rightarrow 1 - t^p &= \frac{1}{2^{\frac{1}{n}}} \\ \Rightarrow t &= (1 - \frac{1}{2^{\frac{1}{n}}})^{\frac{1}{p}}. \end{aligned} \tag{2.6}$$

3 [30pts] Principal Component Analysis

(1) [10 pts] Please describe the similarities and differences between PCA and LDA.

PCA 和 LDA 两种降维方法的求解过程有着很大的相似性，都用到了特征值分解进行降维，但对应的原理却有所区别。首先从目标出发，PCA 选择的是投影后数据方差最大的方向。由于它是无监督的，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，达到降维。而 LDA 选择的是投影后类内方差小、类间方差大的方向，用到了类别标签信息，是有监督的。目的是找到数据中具有判别性的维度，使得原始数据在这些方向上投影后，不同类别尽可能区分开。

(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.)

第一主成分是 $\boldsymbol{e}_1 = (1, 0)$。

$\boldsymbol{X}$ 向量中心化后为 $\begin{pmatrix} -2 \\ 0 \\ 2 \end{pmatrix}$，

$\boldsymbol{Y}$ 向量中心化后为 $\begin{pmatrix} \frac{2}{3} \\ -\frac{4}{3} \\ \frac{2}{3} \end{pmatrix}$，

协方差矩阵 $Q = \begin{pmatrix} \frac{8}{3} & 0 \\ 0 & \frac{8}{9} \end{pmatrix}$，

已经是对角矩阵，所以较大特征值对应的特征向量为第一主成分，即第一主成分是 $\boldsymbol{e}_1 = (1, 0)$。

(3) [10 pts] If we projected the data into 1-d subspace, what are their new corrdinates?

降为一维变量，只取主成分对应的值，新的坐标为 $-2, 0, 2$