Introduction to Machine Learning
Homework 4

April 18, 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] $K_3 = a_1 K_1 + a_2 K_2$ where $a_1, a_2 > 0$;

(2) [10pts] Assume $f(x) = \exp\{-\frac{\|x-\mu\|^2}{2\sigma^2}\}$ where μ and σ 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.

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.

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

4 [25pts] Principal Component Analysis

(1) [5 pts] Please describe 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 corrdinates?