HW5 Handwritten Assignment

December 2022

Problem 1 and Problem 2 will not be graded.

Problem 1 (Kernel) ~~(1%)~~

Consider the following data points:

- $c_1 = \{(3, 3), (3, -3), (-3, 3), (-3, -3)\}$
- $c_2 = \{(6, 6), (6, -6), (-6, 6), (-6, -6)\}$

The data are not linearly separable in this case. Write down a feature map and kernel function to transform the data into a new space, in which the data are linearly separable. Note that you do not just give me a feature map; please explain why.

Problem 2 (SVM with Gaussian kernel) ~~(1%)~~

Consider the task of training a support vector machine using the Gaussian kernel $K(x, z) = \exp(-\frac{\|x-z\|^2}{\tau^2})$. We will show that as long as there are no two identical points in the training set, we can always find a value for the bandwidth parameter τ such that the SVM achieves zero training error.

Recall from class that the decision function learned by the support vector machine can be written as

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i k(x_i, x) + b$$

Assume that the training data $\{(x_1, y_1), \cdots, (x_N, y_N)\}$ consists of points which are separated by at least distance of ϵ; that is, $\|x_j - x_i\| \geq \epsilon$, for any $i \neq j$. For simplicity, we assume $\alpha_i = 1$ for all $i = 1, \cdots, m$ and $b = 0$. Find values for the Gaussian kernel width τ such that x_i is correctly classified, for all $i = 1, \cdots, N$

Hint: Notice that for $y \in \{-1, +1\}$ the prediction on x_i will be correct if $|f(x_i) - y_i| < 1$, so find a value of τ that satisfies this inequality for all i.

Problem 3 (Support Vector Regression)(2%)

Suppose we are given a training set $\{(x_1, y_1), \cdots, (x_m, y_m)\}$, where $x_i \in \mathbb{R}^{(n+1)}$ and $y_i \in \mathbb{R}$. We would like to find a hypothesis of the form $f(x) = w^T x + b$. It is possible that no such function $f(x)$ exists to satisfy these constraints for all points. To deal with otherwise infeasible constraints, we introduce slack variables ξ_i for each point. The (convex) optimization problem is

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i \tag{1}$$

$$\text{s.t. } y_i - w^T x_i - b \le \epsilon + \xi_i \qquad i = 1, \ldots, m \tag{2}$$

$$w^T x_i + b - y_i \le \epsilon + \xi_i \qquad i = 1, \ldots, m \tag{3}$$

$$\xi_i \ge 0 \qquad i = 1, \ldots, m \tag{4}$$

where $\epsilon > 0$ is a given, fixed value and $C > 0$. Denote that $\xi = (\xi_1, \cdots, \xi_m)$.

(a) ~~(0.2%)~~ (0.3%) Write down the Lagrangian for the optimization problem above. Consider the sets of Lagrange multiplier α_i, α_i^*, β_i corresponding to the (2), (3), and (4), so that the Lagrangian would be written as $\mathcal{L}(w, b, \xi, \alpha, \alpha^*, \beta)$, where $\alpha = (\alpha_1, \cdots, \alpha_m)$, $\alpha^* = (\alpha_1^*, \cdots, \alpha_m^*)$, and $\beta = (\beta_1, \cdots, \beta_m)$.

(b) ~~(0.3%)~~ (0.5%) Derive the dual optimization problem. You will have to take derivatives of the Lagrangian with respect to w, b, and ξ

(c) Suppose that $(\bar{w}, \bar{b}, \bar{\xi})$ and $(\bar{\alpha}, \bar{\alpha^*}, \bar{\beta})$ are the optimal solutions to a primal and dual optimization problem, respectively.

Denote $\bar{w} = \sum_{i=1}^{m}(\bar{\alpha}_i - \bar{\alpha_i^*})x_i$

(1) (0.2%) Prove that

$$\bar{b} = \arg\min_{b \in \mathbb{R}} C\sum_{i=1}^{m} \max(|y_i - (\bar{w}^T x_i + b)| - \epsilon, 0) \tag{5}$$

(2) (1%) Define $e = y_i - \left(\overline{w}^T x_i + \bar{b}\right)$ Prove that

$$\begin{cases} \bar{\alpha}_i = \bar{\alpha}_i^* = 0, & \bar{\xi_i} = 0, & \text{if } |e| < \epsilon \\ 0 \le \bar{\alpha}_i \le C, & \bar{\xi_i} = 0, & \text{if } e = \epsilon \\ 0 \le \bar{\alpha}_i^* \le C, & \bar{\xi_i} = 0, & \text{if } e = -\epsilon \\ \bar{\alpha}_i = C, & \bar{\xi_i} = e - \epsilon & \text{if } e > \epsilon \\ \bar{\alpha}_i^* = C, & \bar{\xi_i} = -(e + \epsilon) & \text{if } e < -\epsilon \end{cases} \tag{6}$$

(d) ~~(0.3%)~~ Show that the algorithm can be kernelized and write down the kernel form of the decision function. For this, you have to show that

(1) The dual optimization objective can be written in terms of inner products or training examples

(2) At test time, given a new x the hypothesis $f(x)$ can also be computed in terms of inner produce.