CSE 6740 A/ISyE 6740, Fall 2023, Georgia Tech

CSE 6740: Homework 3: Kernel methods

Due Oct 14th, '23 (11:59 pm ET) on Gradescope

Cite any sources and collaborators; do not copy. See syllabus for policy.

Textbook exercise 6.22 from Mohri et al; Parts 1 and 2 solved in class

The question with hints is given below. Some solutions are in the lecture notes, but please submit the solutions you have reworked on your own. Hints are given below as well.

Let the input domain be $\mathcal{X}$. Consider a Hilbert space $\mathbb{H}$, a feature map $\Phi : \mathcal{X} \to \mathbb{H}$ and a kernel $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product on $\mathbb{H}$.

Part 1: Minimum enclosing ball (MEB) problem (20 pts)

Consider the following optimization problem for finding the minimum enclosing ball (MEB) of a set of points $S = \{x_1, \ldots, x_m\} \subset \mathcal{X}$:

$$\min_{r>0, c\in\mathbb{H}} \quad r^2 \quad \text{subject to} \quad \|c - \Phi(x_i)\|^2 \le r^2, \quad i = 1, \ldots, m. \tag{1}$$

Show how to derive the dual optimization problem:

$$\max_{\alpha\in\mathbb{R}^m} \quad \sum_{i=1}^m \alpha_i k(x_i, x_i) - \frac{1}{2} \sum_{i,j=1}^m \alpha_i \alpha_j k(x_i, x_j) \quad \text{subject to} \quad \alpha_i \ge 0 \text{ and } \sum_{i=1}^m \alpha_i = 1 \quad i = 1, \ldots, m. \tag{2}$$

Prove that the optimal solution $c = \sum_{i=1}^m \alpha_i \Phi(x_i)$ is a convex combination of the features at the training points $x_1, \ldots, x_m$.

Hints: i) Make problem finite dimensional in c using Kernel trick. Justify this step as done in class. ii) write down the KKT conditions for the primal problem.

Part 2: Anomaly detection hyothesis class (20 pts)

Consider the hypothesis class

$$\mathcal{H} = \left\{ h_{c,r}(x) = r^2 - \|c - \Phi(x)\|^2 : \|c\| \le \Lambda, 0 < r \le R \right\}, \tag{3}$$

where $\|\cdot\|$ is the norm induced by the inner product on $\mathbb{H}$, i.e., $\|c\| = \sqrt{\langle c, c \rangle}$. A hypothesis $h_{c,r}$ is an anomaly detector that flags an input x as an anomaly if $h_{c,r}(x) < 0$. Show that if

$\sup_x \|\Phi(x)\| < M$, then the solution to the MEB problem in Part 1 is in $\mathcal{H}$ with $\Lambda \leq M$ and $R \leq 2M$.

Hint: Use the complementarity conditions in Part 1 to get an expression for an optimal r in terms of α and $\Phi(x_i)$. Now that you have expressions for optimal c and r, prove that their norms are upper bounded by M and $2M$ respectively.

Part 3: the kernel SVM interpretation (30 pts)

Let $k(x, x) = 1$, a constant independent of x (this is, e.g., true for the Gaussian kernel). Derive the following margin-maximization and minimization of the slack penalty $\sum_i \xi$ for finding a hyperplane for this 1-class classification problem:

$$\min_{w,\xi} (1/2)\|w\|^2 + C\|\xi\|_1 \quad \text{subject to} \quad \langle w, \Phi(x_i)\rangle \geq 1 - \xi_i, \xi \geq 0, i \in [m]. \tag{4}$$

Here, all the training points have true labels 1. Suppose ν is an upper bound on the fraction of support vectors out of m training points. Equivalently, a maximum of $\nu\, m$ points are allowed to have $\alpha_i \neq 0$: they could be misclassified as anomalies ($\xi > 1$) or classified with a nonzero penalty ($1 > \xi_i \geq 0$) as non-anomalies.

Show that when $C = 1/(\nu m)$, the above problem is equivalent to MEB in Part 1. This means that one can equivalently find a hyperplane instead of a minimal enclosing hypersphere in feature space (see [1] for more on this).

Hints: 1) follow the derivation done in class of maximum (geometric) margin classification leading to the soft SVM problem; now there is only one label class and the domain space is feature space, i.e., $x_i \to \Phi(x_i)$. 2) show that the dual form of MEB in Part 1 reduces, when $k(x, x) = 1$ to

$$\min_{\alpha} \sum_{i,j}^{m} \alpha_i \alpha_j k(x_i, x_j) \quad \text{subject to} \quad \alpha_i \geq 0, \sum_i \alpha_i = 1, i \in [m]. \tag{5}$$

3) Next derive the dual form of (4) by first writing down the KKT conditions. Your results should be very similar to the soft-SVM KKT conditions (5.26-5.30 in Mohri et al). 4) Now, $\alpha_i = 0$ or $0 < \alpha \leq C$. Thus, $\sum_{i=1}^{m} \alpha_i \leq C\times$ number of support vectors. Using this, prove that the two dual forms are equivalent.

References

[1] Bernhard Schölkopf, John C Platt, John Shawe-Taylor, Alex J Smola, and Robert C Williamson. Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471, 2001.