HW 2

Karl Hiner

October 14, 2023

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 $\mathbb{H}$.

1 Minimum enclosing ball (MEB) problem

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

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

Show how to derive the dual optimization problem:

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

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:

1. Make the problem finite dimensional in c using the Kernel trick. Justify this step as done in class.
2. Write down the KKT conditions for the primal problem.

Answer:

1.1 Make the problem finite dimensional in c

Using the representer theorem, we can express the center c of the MEB as a linear combination of the mapped data points:

$$c = \sum_{i=1}^{m} a_i \Phi(x_i)$$

Substituting this expression into the optimization problem gives:

$$\min_{r>0, a\in\mathbb{R}^m} r^2 \text{ subject to } \left\| \sum_{j=1}^{m} a_j \Phi(x_j) - \Phi(x_i) \right\|^2 \leq r^2, \quad i = 1, \ldots, m$$

1.2 KKT conditions for the primal problem

The Lagrangian for this problem is:

$$L(r, a, \alpha) = r^2 + \sum_{i=1}^{m} \alpha_i \left(\left\| \sum_{j=1}^{m} a_j \Phi(x_j) - \Phi(x_i) \right\|^2 - r^2 \right)$$

Where α_i are the Lagrange multipliers for the constraints. Now, we can derive the KKT conditions:

$$\begin{aligned}
\frac{\partial L}{\partial r} &= 0 \\
2r - 2r \sum_{i=1}^{m} \alpha_i &= 0 \\
\sum_{i=1}^{m} \alpha_i &= 1 \quad \text{(Disregarding the trivial case of } r = 0\text{)} \\
\frac{\partial L}{\partial a} &= 0 \\
\alpha_i \left(\left\| \sum_{j=1}^{m} a_j \Phi(x_j) - \Phi(x_i) \right\|^2 - r^2 \right) &= 0, \quad \alpha_i \geq 0 \quad i = 1, \ldots, m \quad \text{(Complementarity)}
\end{aligned}$$

Using the kernel trick and the inner product definition, we can rewrite the squared norm constraint:

$$\begin{aligned}
\left\| \sum_{j=1}^{m} a_j \Phi(x_j) - \Phi(x_i) \right\|^2 &\leq r \\
\langle \sum_{j=1}^{m} a_j \Phi(x_j) - \Phi(x_i), \sum_{k=1}^{m} a_k \Phi(x_k) - \Phi(x_i) \rangle & \qquad \text{Inner product} \\
\sum_{j,k=1}^{m} a_j a_k \langle \Phi(x_j), \Phi(x_k) \rangle - 2 \sum_{j=1}^{m} a_j \langle \Phi(x_j), \Phi(x_i) \rangle + \langle \Phi(x_i), \Phi(x_i) \rangle & \qquad \text{Expand} \\
\sum_{j,k=1}^{m} a_j a_k k(x_j, x_k) - 2 \sum_{j=1}^{m} a_j k(x_j, x_i) + k(x_i, x_i) & \qquad k(x, y) \coloneqq \langle \Phi(x), \Phi(y) \rangle \\
k(x_i, x_i) + \sum_{j,k=1}^{m} a_j a_l k(x_j, x_k) - 2 \sum_{j=1}^{m} a_j k(x_i, x_j) & \qquad \text{Rearrange}
\end{aligned}$$

Substituting this result into the Lagrangian and taking the derivative with respect to a:

$$\begin{aligned}
L(r,a,\alpha) &= r^2 + \sum_{i=1}^{m} \alpha_i \left(k(x_i,x_i) + \sum_{j,k=1}^{m} a_j a_k k(x_j,x_k) - 2\sum_{j=1}^{m} a_j k(x_i,x_j) - r^2 \right) \\
\frac{\partial L}{\partial a_p} &= 0 \\
0 &= \sum_{i=1}^{m} \alpha_i \left(\partial a_p \left(\sum_{j,k=1}^{m} a_j a_k k(x_j,x_k) - 2\sum_{j=1}^{m} a_j k(x_i,x_j) \right)\right) \quad \text{(Drop terms w/o } a_p\text{)} \\
&= \sum_{i=1}^{m} \alpha_i \left(2\sum_{j=1}^{m} a_j k(x_p,x_j) - 2k(x_i,x_p) \right) \quad \forall p \in [m] \quad \text{(Differentiate)} \\
&= \sum_{i=1}^{m} \alpha_i \left(\sum_{j=1}^{m} a_j k(x_p,x_j) - k(x_i,x_p) \right) \quad \forall p \in [m] \quad \text{(Simplify)} \\
&= \sum_{j=1}^{m} a_j k(x_p,x_j) - \sum_{i=1}^{m} \alpha_i k(x_i,x_p), \quad \forall p \in [m] \quad \left(\sum_{i=1}^{m} \alpha_i = 1 \right) \\
&= \sum_{j=1}^{m} a_j k(x_p,x_j) - \sum_{j=1}^{m} \alpha_j k(x_j,x_p), \quad \forall p \in [m] \quad \text{(Change index label)} \\
0 &= \sum_{j=1}^{m} (a_j - \alpha_j) k(x_i,x_j), \quad \forall i \in [m] \quad \text{(Change } p \text{ index to } i\text{. } k(\cdot,\cdot) \text{ is symmetric.)}
\end{aligned}$$

This implies that $a_j = \alpha_j$ for all $j \in [m]$ (ignoring the trivial case where $\forall (i,j), k(x_i,x_j) = 0$).

Now, let's simplify the Lagrangian, $L(r,a,\alpha)$:

$$\begin{aligned}
& r^2 + \sum_{i=1}^{m} \alpha_i \left(k(x_i,x_i) + \sum_{j,k=1}^{m} a_j a_k k(x_j,x_k) - 2\sum_{j=1}^{m} a_j k(x_i,x_j) - r^2 \right) \quad \text{(Start)} \\
& \sum_{i=1}^{m} \alpha_i \left(k(x_i,x_i) + \sum_{j,k=1}^{m} a_j a_k k(x_j,x_k) - 2\sum_{j=1}^{m} a_j k(x_i,x_j) \right) + r^2 - \sum_{i=1}^{m} \alpha_i r^2 \quad \text{(Rearrange)} \\
& \sum_{i=1}^{m} \alpha_i \left(k(x_i,x_i) + \sum_{j,k=1}^{m} a_j a_k k(x_j,x_k) - 2\sum_{j=1}^{m} a_j k(x_i,x_j) \right) \quad \left(\sum_{i=1}^{m} \alpha_i = 1 \right) \\
& \sum_{i=1}^{m} \alpha_i k(x_i,x_i) + \sum_{i=1}^{m} \alpha_i \sum_{j,k=1}^{m} a_j a_k k(x_j,x_k) - 2\sum_{i=1}^{m} \alpha_i \sum_{j=1}^{m} a_j k(x_i,x_j) \quad \text{(Distribute)} \\
& \sum_{i=1}^{m} \alpha_i k(x_i,x_i) + \sum_{i,j=1}^{m} a_i a_j k(x_k,x_j) - 2\sum_{i=1}^{m} \alpha_i \sum_{j=1}^{m} a_j k(x_i,x_j) \quad \left(\sum_{i=1}^{m} \alpha_i = 1 \right) \\
& \sum_{i=1}^{m} \alpha_i k(x_i,x_i) - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(x_i,x_j) \quad (a_j = \alpha_j \forall j \in [m])
\end{aligned}$$

To derive the dual problem, we maximize L with respect to α (which are our dual variables), and apply the complementarity constraints:

$$\max_{\alpha \in \mathbb{R}^m} \left\{ \sum_{i=1}^{m} \alpha_i k(x_i, x_i) - \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \right\}$$

subject to:

$$\alpha_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{m} \alpha_i = 1, \quad i = 1, \ldots, m$$

This is what we wanted to show.

Convex combination proof: The constraints in the dual problem ensure that $\alpha_i \geq 0$ for all i and that their sum is 1. This implies that each α_i is a weight in the convex combination. Therefore, the optimal solution $c = \sum_{i=1}^{m} \alpha_i \Phi(x_i)$ is indeed a convex combination of the features at the training points $x_1, \ldots, x_m$.

2 Anomaly detection hypothesis class

Consider the hypothesis class

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

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.

Answer:

In Part 1, we showed that we can express c as a linear combination of the mapped data points:

$$\begin{aligned}
c &= \sum_{i=1}^{m} \alpha_i \Phi(x_i) \\
\|c\| &= \left\| \sum_{i=1}^{m} \alpha_i \Phi(x_i) \right\| \\
&\leq \sum_{i=1}^{m} |\alpha_i| \|\Phi(x_i)\| \quad \text{By the triangle inequality} \\
&= \sum_{i=1}^{m} \alpha_i \|\Phi(x_i)\| \quad \text{Since } \alpha_i \geq 0, \forall i \in [m] \\
&\leq \sup_x \|\Phi(x)\| \quad \text{Using def. of supremum, and } \sum_{i=1}^{m} \alpha_i = 1 \\
&\leq M \quad \text{by assumption}
\end{aligned}$$

Thus, if c in an optimal hypothesis, $\|c\| \leq M$. Since any c in the hypothesis class $\mathcal{H}$ hs $\|c\| \leq \Lambda$, then any c in the solution to Part 1 is in $\mathcal{H}$ with $\Lambda \leq M$.

Now, we find an expression for the optimal r. From the complementarity conditions in Part 1, we have:

$$\alpha_i \left(\left\| \sum_{j=1}^m \alpha_j \Phi(x_j) - \Phi(x_i) \right\|^2 - r^2 \right) = 0.$$

Ignoring the trivial case of $\alpha_i = 0$, this gives:

$$\begin{aligned} r^2 &= \left\| \sum_{j=1}^m \alpha_j \Phi(x_j) - \Phi(x_i) \right\|^2. \\ &\leq \left(\left\| \sum_{j=1}^m \alpha_j \Phi(x_j) \right\| + \|\Phi(x_i)\| \right)^2 \qquad \text{Triangle inequality} \end{aligned}$$

Given the constraint $\sup_x \|\Phi(x)\| < M$ and since the α's sum to 1 and are non-negative, both terms in the sum can be bounded by M. Thus,

$$\left(\left\| \sum_{j=1}^m \alpha_j \Phi(x_j) \right\| + \|\Phi(x_i)\| \right)^2 \leq (M + M)^2 = 4M^2.$$

And so, $r^2 \leq 4M^2 \implies r \leq 2M$. This, since any r in the hypothesis class $\mathcal{H}$ hs $0 < r \leq R$, then any r in the solution to Part 1 is in $\mathcal{H}$ with $R \leq 2M$.

3 The kernel SVM interpretation

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:

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

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 νm 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.

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 the 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

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

3. Next, derive the dual form of (1) 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, $\sum_{i=1}^{m} \alpha_i \leq C \times$ the number of support vectors. Using this, prove that the two dual forms are equivalent.

Answer:

3.1 Soft SVM Derivation

We start with the maximum margin classification problem for one-class SVM. Here, we maximize the geometric margin while penalizing slack variables.

$$\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\|\xi\|_1 \tag{2}$$

subject to:

$$\begin{aligned} \langle w, \Phi(x_i)\rangle &\geq 1 - \xi_i, \\ \xi &\geq 0, \end{aligned}$$

for $i \in [m]$, where ξ_i are the slack variables, and C is a constant.

3.2 Dual form of Minimum Enclosing Ball

Let's first express the dual form of MEB when $k(x,x) = 1$. The dual form is

$$\min_{\alpha} \sum_{i,j=1}^{m} \alpha_i \alpha_j k(x_i, x_j)$$

subject to:

$$\alpha_i \geq 0, \quad i \in [m] \quad \text{and} \quad \sum_{i=1}^{m} \alpha_i = 1.$$

3.3 KKT Conditions for Equation 2

The Lagrangian of the problem in Equation 2 can be formulated as:

$$L(w, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\|\xi\|_1 - \sum_{i=1}^{m} \alpha_i(\langle w, \Phi(x_i)\rangle - 1 + \xi_i) - \sum_{i=1}^{m} \beta_i \xi_i,$$

with $\alpha_i \geq 0, \beta_i \geq 0$ and $\xi_i \geq 0$ for $i \in [m]$.

Stationarity conditions:

$$\begin{aligned} \frac{\partial L}{\partial w} &= w - \sum_{i=1}^{m} \alpha_i \Phi(x_i) = 0 \\ \implies w &= \sum_{i=1}^{m} \alpha_i \Phi(x_i) \\ \frac{\partial L}{\partial \xi_i} &= C - \alpha_i - \beta_i = 0 \end{aligned}$$

Complementarity conditions:

$$\begin{aligned} \alpha_i(\langle w, \Phi(x_i)\rangle - 1 + \xi_i) &= 0, \\ \beta_i \xi_i &= 0. \end{aligned}$$

Primal feasibility conditions:

$$\langle w, \Phi(x_i) \rangle \geq 1 - \xi_i, \xi_i \geq 0 \text{ for } i \in [m]$$

Dual feasibility conditions:

$$\alpha_i \geq 0, \beta_i \geq 0 \text{ for } i \in [m]$$

3.4 Equivalence of Dual forms

Using the following:

1. The constraint for the dual form of MEB, $\sum_{i=1}^{m} \alpha_i = 1$,
2. The stationarity condition for ξ_i, $C - \alpha_i - \beta_i = 0$, and
3. The dual feasibility condition $\alpha_i \geq 0$ for $i \in [m]$,

we can derive that $\alpha_i \in [0, C]$.

Due to the upper bound ν on the fraction of support vectors, at most νm points can have $\alpha_i \neq 0$. Therefore,

$$\begin{aligned} \sum_{i=1}^{m} \alpha_i &\leq C\nu m \\ &= 1 \quad (\text{substitute } C = \frac{1}{\nu m}), \end{aligned}$$

which is a relaxation of the MEB constraint $\sum_{i=1}^{m} \alpha_i = 1$.

Thus, under $C = \frac{1}{\nu m}$, the dual form of Equation 2 is a relaxed version of the dual form of MEB.