CSE512 Machine Learning

Chumki Acharya (112683478)

25 June, 2020

Homework-2

Question 1

Solution:

Let us consider two set of observations $S = (x_1, \ldots, x_n)$ and $T = (z_1, \ldots, z_m)$. If we prove that if convex hulls of S and T intersects, they are not linearly separable then we can say S and T are linearly separable if and only if their convex hulls do not intersect.

Now if the convex hulls intersect, there must be at least one point in common between **x** and **z**. Let's call that point **xz**. Since **xz** belongs to both convex hulls, there must be a set of $\{\alpha_n\}$ and $\{\beta_m\}$ that give rise to **xz**. We know the linear discriminant for **x** and **z** can be written as

$$y(\mathbf{x}) = \sum_n \alpha_n \left(\hat{\mathbf{w}}^T \mathbf{x}^n + w_0\right) \tag{1}$$

and

$$y(\mathbf{z}) = \sum_m \beta_m \left(\hat{\mathbf{w}}^T \mathbf{z}^m + w_0\right) \tag{2}$$

Thus The linear discriminant for **xz** can now be written as

$$y(\mathbf{xz}) = \sum_n \alpha_n \left(\hat{\mathbf{w}}^T \mathbf{x}^n + w_0\right) = \sum_m \beta_m \left(\hat{\mathbf{w}}^T \mathbf{z}^m + w_0\right) \tag{3}$$

If the sets are linearly separable, we must have

$$\begin{aligned} y\left(\mathbf{x}^n\right) &= \hat{\mathbf{w}}^T \mathbf{x}^n + w_0 > 0 \\ \text{and } y\left(\mathbf{z}^m\right) &= \hat{\mathbf{w}}^T \mathbf{z}^m + w_0 < 0 \end{aligned} \tag{4}$$

Now from the non-negativity and simplex constraints on α and β, (3) and (4), we can say that the linear discriminant **y(xz)** has to be simultaneously greater than and less than zero which is impossible. This arise a contradiction.

Again, if the sets are linearly separable, we know that

$$\begin{aligned} y\left(\mathbf{x}^n\right) &= \hat{\mathbf{w}}^T \mathbf{x}^n + w_0 > 0 \\ \text{and } y\left(\mathbf{z}^m\right) &= \hat{\mathbf{w}}^T \mathbf{z}^m + w_0 < 0 \end{aligned} \tag{5}$$

and considering that there is a point **xz** lying in the intersection of the convex hulls we have,

$$y(\mathbf{xz}) = \sum_n \alpha_n \left(\hat{\mathbf{w}}^T \mathbf{x}^n + w_0\right) = \sum_m \beta_m \left(\hat{\mathbf{w}}^T \mathbf{z}^m + w_0\right) \tag{6}$$

Thus the equality in above is not possible given the inequality (5) that the sets are linearly separable. Hence we can say that the sets are linearly separable is their convex hull do not intersect.

Question 2

Solution:

We are given a linear model of the form

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$$

with a training set $S = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$,

Thus the empirical risk is measured by

$$L_S(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^{m} \left(f\left(\mathbf{x}_i\right) - y_i\right)^2$$

Now the expected empirical risk averaged over the noise distribution can be given as

$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\left(\hat{f}_n\left(x_i\right) - f^*\left(x_i\right)\right)^2\right]$

By Tikhonov regularization,

$A(S) = \underset{\mathbf{w}}{\operatorname{argmin}} \left(L_S(\mathbf{w}) + \lambda \|\mathbf{w}\|^2\right)$

Thus comparing these two we can say that these are equivalent.

Question 3

Solution:

We can say for every training set with at least two labeled observations, which is separable by a homogeneous hyperplane, the hard- and soft-SVM learning algorithms will not always return the exact same hypothesis. To prove this

Let us consider a $\lambda > 0$ for every sample S of m > 1 examples is separable with a class of homogeneous halfspace.

Now lets consider, $x_0 = (0, \alpha) \in \mathbb{R}^2$, where $\alpha \in (0, 1)$, $x_k = (0, k)$ where $k = 1, \ldots, m-1$, $y_0 = \ldots = y_{m-1} = 1$ and $S = \{(x_i, y_i) : i \in \{1, \ldots, m-1\}\}$.

The hard-SVM problem is given as

$$\min_{\mathbf{w}} \|\mathbf{w}\|^2 \text{ s.t. } \forall i, y_i \langle \mathbf{w}, \mathbf{x}_i \rangle \geq 1$$

Solving this with parameter λ with value $\frac{1}{\alpha^2}$ we get w = $(0, \frac{1}{\alpha})$.

The Soft-SVM problem is given by

$$\min_{\mathbf{w}, b, \xi} \left(\lambda \|\mathbf{w}\|^2 + \frac{1}{m} \sum_{i=1}^{m} \xi_i\right)$$

$$\text{s.t. } \forall i, \quad y_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right) \geq 1 - \xi_i \text{ and } \xi_i \geq 0$$

Now considering, $\lambda \cdot 1 + \frac{1}{m}(1-\alpha) \leq \frac{1}{\alpha^2}$ we get the solution of soft-SVM as w = $(0, 1)$.

We know $\alpha \in (0, 1)$, therefore we can write $\frac{1}{\alpha^2} > \lambda + \frac{1}{m}$. Thus, there exists $\alpha_0 > 0$ s.t. for every $\alpha < \alpha_0$, the above inequality becomes true. Now if α becomes very small, then soft-SVM tends to not considering x_0 and hence disproves the above statement.

Question 4

Solution:

Let us take K_1 and K_2 are valid kernels on a domain X , i.e., their Gram matrix is symmetric and positive-definite.

(a) We are given, $K(u, v) = \alpha K_1(u, v) + \beta K_2(u, v)$ for any scalars $\alpha, \beta \geq 0$ we need to prove this is a valid kernel function.

We can write,

$\alpha K_1(u, v) = \langle \sqrt{\alpha}\phi_1(u), \sqrt{\alpha}\phi_1(v) \rangle$ and

$\beta K_2(u, v) = \left\langle \sqrt{\beta}\phi_2(u), \sqrt{\alpha}\phi_2(v) \right\rangle$

Again, it is given their gram matrix, i.e inner product is also symmetric and thus K_1 and K_2 are also symmetric . Thus we can say

$$
\begin{aligned}
& K(u, v) \\
&= \alpha K_1(u, v) + \beta K_2(u, v) \\
&= \langle \sqrt{\alpha}\phi_1(u), \sqrt{\alpha}\phi_1(v) \rangle + \left\langle \sqrt{\beta}\phi_2(u), \sqrt{\alpha}\phi_2(v) \right\rangle \\
&= \left\langle [\sqrt{\alpha}\phi_1(u), \sqrt{\alpha}\phi_1(v)], \left[\sqrt{\beta}\phi_2(u), \sqrt{\alpha}\phi_2(v)\right] \right\rangle
\end{aligned}
$$

Since $K(u, v)$ can be shown in terms of an inner product, we can say it is symmetric and positive semi-definite by definition and hence a valid kernel function.

(b) We are given, $K(u, v) = K_1(u, v)K_2(u, v)$

Also, it is given that K_1K_2 are valid kernels on domain X with symmetric and positive-definite gram matrix and we know gram matrix is the element-wise, entrywise product of K_1 and K_2.

Now if we assume K_1 and K_2 to be covariance matrices $(X_1, X_2, \ldots X_n)$ and $(Y_1, Y_2, \ldots Y_n)$ respectively and take their element-wise, entrywise product, then we get $(X_1Y_1, X_2Y_2, \ldots X_nY_n)$ which is also a covariance matrix.

Now we know convariance matrices are symmetric and positive-definite and thus we can write $K(u, v)$ in terms of one and hence we can conclude that $K(u, v)$ is a valid kernel function.

References

http://www.stat.cmu.edu/ larry/=sml2008/hw5_solution.pdf
http://www.stat.cmu.edu/Alon Gonen_Dana Rubinstein