Problem Set 4 for Machine Learning 15 Fall

Jingyuan Liu
AndrewId: jingyual
jingyual@andrew.cmu.edu

1 VC dimension

1.1 Show the VC dimension of linear classifier

To prove that the linear classifier h with x in R^n has the VC dimension of n + 1, we need to prove that $VCdim(h_n) >= n + 1$, and then prove that $VCdim(h_n) <= n + 1$.

(a). Prove $VCdim(h_n) >= n + 1$

First of all, we could use the Mathematical Induction to prove that $VCdim(H) >= n + 1$:

For n = 1, it is easy to get $VCdim(h_1) >= 2$

For n = 2, we could also get $VCdim(h_2) >= 3$, which could be proved using following figures:

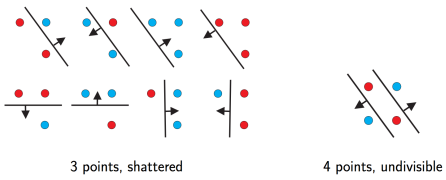

Figure 1: The representation of $VCdim(h_2)$

For n = i, we assume that the $VCdim(h_i) >= i + 1$. We could form a matrix of (i+1)*(i+1), with the rank of i+1, since the VCdim is i+1.

For n = i+1, we could form a matrix of (i+2)*(i+2), since the variable x have one more independent dimension. Therefore, $VCdim(h_{i+1}) >= VCdim(h_i) + 1 >= i + 2$.

Therefore, we have $VCdim(h_n) >= n + 1$.

(b). Prove $VCdim(h_n) <= n + 1$

Then we should prove that $VCdim(h_n) <= n + 1$:

Suppose we could shatter n+2 points, then using the convex combinations of points in C, we could seperate points into S1 and S2, that:

$$conv(s1) \cap conv(s2) \neq \emptyset \tag{1}$$

However, if $VCdim(h_n) = n + 2$, then we could spilt points to half space contains s1 and the complement of the half space contains s2. This implies that the both half space contains the convex hull of s1 and the complement of the half space the contains the convex hull of s2. Thus, we get:

$$conv(s1) \cap conv(s2) = \emptyset \tag{2}$$

These two observations are contradictory to each other. So we know that for the n+2 points here, we could not use the linear classifier to shatter all of them.

In conclusion, we prove that $VCdim(h_n) >= n+1$ and $VCdim(h_n) <= n+1$ for the linear classifier, then we have $VCdim(h_n) = n+1$

1.2 Show the VC dimension of axis-aligned boxes

Similarily, for this case, to prove that the axis-aligned boxes classifier h with x in R^n has the VC dimension of 2n, we need to prove that $VCdim(h_n) >= 2n$, and then prove that $VCdim(h_n) <= 2n$.

(a). Prove $VCdim(h_n) >= 2n$

First of all, we need to prove that for any case, $VCdim(h_n) >= 2n$, which means if the x has n dimensions, we can use the axis-aligned boxes classifiers to shatter 2n points.

Suppose we have n dimensions, then we can map all the x to a dimension i. In the dimension i, we can have x^i_{max} and x^i_{min}. Therefore, in this dimension i, we could always shatter at least 2 points via a split between the x^i_{max} and x^i_{min}.

Considering we have n dimensions, and the mapping of each dimension of the data is independent to the mapping of other dimension of the data. Therefore, we could at least shatter 2*n points via the axias-alighed boxes. Therefore, we have $VCdim(h_n) >= 2n$.

(b). Prove $VCdim(h_n) <= 2n$

Then we need to prove that $VCdim(h_n) <= 2n$. Suppose we have 2n + 1 points, we could always find the 2n "boundries", which is the min value and max value for each dimension, and form an "area". Then the 2n+1th points will be guaranteed to appear within the selected "area".

For any 2n+1 points, we could always transfer to the above mentioned scenario. In this scenario, we could not classify the 2n+1th point, because it is contained within the "area", and no rules are guaranteed to rightly classify it.

Therefore, by mapping the data to each dimension and form an "area", we could prove that $VCdim(h_n) <= 2n$.

In conclusion, we prove that $VCdim(h_n) >= 2n$ and $VCdim(h_n) <= 2n$ for the axis-alighed boxes, then we have $VCdim(h_n) = 2n$

2 Support Vector Machines

2.1 Support Vector Regression

2.1.1. write the dual problem

Suppose we have ξ_i ofr each x_i:

$$|f(x) - y| < \epsilon \tag{3}$$

$$-\epsilon < w^T x_i - y_i < \epsilon \tag{4}$$

Let's set:

$$\left|w^T x_i - y_i\right| - \epsilon = \xi_i \tag{5}$$

$$L = \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}\xi_i - \sum_i \alpha_i(\epsilon + \xi_i - y_i + wx_i) - \sum_i \alpha_i^\star(\epsilon + \xi_i + y_i - wx_i) - \sum_i \beta_i\xi_i \tag{6}$$

$$\alpha_i >= 0, \alpha_i^\star >= 0, \beta_i >= 0 \tag{7}$$

We have to satisfy saddle constraints condition, so

$$\partial_w L = w - \sum_i (\alpha_i - \alpha_i^\star) x_i = 0 \tag{8}$$

$$\partial_{\xi_i} L = C - \alpha_i - \alpha_i^\star - \beta_i = 0 \tag{9}$$

Take the the conditions back and we can get the form of dual problem of SVR:

$$maximize : -\frac{1}{2} \sum_{i,j=1} (\alpha_i - \alpha_i^\star)(\alpha_j - \alpha_j^\star) x_i^T x_j - \epsilon \sum_i (\alpha_i + \alpha_i^\star) + \sum_i y_i (\alpha_i - \alpha_i^\star) \tag{10}$$

$$subject \quad to : \sum_i (\alpha_i - \alpha_i^\star) = 0, \qquad \alpha_i, \alpha_i^\star \in [0, C] \tag{11}$$

2.1.2. KKT Conditions

The KKT condition for this problem is:

$$\partial_w L = w - \sum_i (\alpha_i - \alpha_i^\star) x_i = 0 \tag{12}$$

$$\partial_{\xi_i} L = C - \alpha_i - \alpha_i^\star - \beta_i = 0 \tag{13}$$

$$\epsilon + \xi_i - y_i + w^T x_i >= 0 \tag{14}$$

$$\epsilon + \xi_i^\star + y_i - w^T x_i >= 0 \tag{15}$$

$$\alpha_i >= 0, \alpha_i^\star >= 0, \beta_i >= 0 \tag{16}$$

$$\alpha_i (\epsilon + \xi_i - y_i + w^T x_i) = 0 \tag{17}$$

$$\alpha_i^\star (\epsilon + \xi_i^\star + y_i - w^T x_i) = 0 \tag{18}$$

2.1.3. Kernelized SVR

The prediction rule is:

$$w = \sum_i (\alpha_i - \alpha_i^\star) x_i \tag{19}$$

$$f(x) = \sum_i (\alpha_i - \alpha_i^\star) x_i^T x \tag{20}$$

Add we can use kernel to transfer the x, therefore:

$$f(x_j) = \sum_i (\alpha_i - \alpha_i^\star) K(x_i, x_j) \tag{21}$$

2.1.4. Why using SVR primal

We are using the primal of SVM and the dual of SVR, this is because we want to compose the term:

$$\sum_i x_i x \tag{22}$$

Therefore, we could use the kernel trick to implicitly transfer the feature of the data to other spaces.

2.2 Support Kernel Machines

2.2.1.(a) The minimum of the Lagrangian function

Given L, the Lagrangian function, we could derive to optimize the w:

$$\frac{\partial L}{\partial w_j} = \frac{1}{2}\frac{\partial(\sum_j d_j \left\|w_i\right\|_2)^2}{\partial w_j} - \frac{\partial \alpha_i y_i \sum_j w_j^T x_{ji}}{\partial w_j} \tag{23}$$

We can have $\frac{\partial L}{\partial w_j} = 0$. Then we can both times the w_j^T:

$$\frac{w_j d_j \sum_j \left\|w_j\right\|_2 d_j}{\left\|w_j\right\|_2} = \sum_i \alpha_i y_i x_{ji} \tag{24}$$

$$\left\|w_j\right\|_2 d_j \gamma = w_j^T \sum_i \alpha_i y_i x_{ji} \tag{25}$$

2.2.1.(b) Show that $\left\|\sum_i \alpha_i y_i x_{ji}\right\|_2 < d_j \gamma$

In the last section, we have:

$$\frac{w_j d_j \sum_j \left\|w_j\right\|_2 d_j}{\left\|w_j\right\|_2} = \sum_i \alpha_i y_i x_{ji} \tag{26}$$

Therefore, when the w is not 0, we can take the l2-norm of the this two vectors and get:

$$\left\|\frac{w_j d_j \sum_j \left\|w_j\right\|_2 d_j}{\left\|w_j\right\|_2}\right\|_2 = \left\|\sum_i \alpha_i y_i x_{ji}\right\|_2 \tag{27}$$

Because w is a vector, we know that:

$$\left\|\frac{w_j}{\left\|w_j\right\|_2}\right\|_2 = 1 \tag{28}$$

So we have:

$$\left\|w_j\right\|_2 d_j \gamma = w_j^T \sum_i \alpha_i y_i x_{ji} \tag{29}$$

2.2.1.(c) Show that

For the first point, from the previous problem, we would know that:

$$\left\|\sum_i \alpha_i y_i x_{ji}\right\|_2 <= d_j \gamma \tag{30}$$

And we have also proven that, when w is not zero:

$$\left\| \sum_i \alpha_i y_i x_{ji} \right\|_2 = d_j \gamma \tag{31}$$

Then with the $w_j = 0$, we should have:

$$\left\| \sum_i \alpha_i y_i x_{ji} \right\|_2 < d_j \gamma \tag{32}$$

For the second point, we would know that:

$$\frac{w_j d_j \sum_j \|w_j\|_2 d_j}{\|w_j\|_2} = \sum_i \alpha_i y_i x_{ji} \tag{33}$$

$$w_j = \frac{\|w_j\|_2}{d_j \gamma} \sum_i \alpha_i y_i x_j i \tag{34}$$

We could know that $\|w_j\|_2 > 0, d_j > 0, \gamma > 0$, therefore, we could always have a η_i:

$$w_j = \eta_j \sum_i \alpha_i y_i x_{ji}, \eta_j > 0 \tag{35}$$

When $w_j = 0$, we could set $\eta_j = 0$. Therefore, we always have:

$$\eta_j >= 0 \tag{36}$$

2.2.2. Encourage Sparsity From the 2.2.1.(c), we would know that:

$$if \left\| \sum_i \alpha_i y_i x_{ji} \right\|_2 < d_j \gamma, \quad then \quad w_j = 0 \tag{37}$$

We can see that if the result of $y_i x_i \alpha_i$ is smaller than a certain number, which is given by the regularization terms, then the parameter w_j would tend to vanish to minimize the total loss.

2.2.3. Kernelized version For the kernelized version, We could derive the graient for the minimal in the same way:

$$\frac{\partial L}{\partial w_j} = \frac{1}{2} \frac{\partial (\sum_j d_j \|w_i\|_2)^2}{\partial w_j} - \frac{\partial \alpha_i y_i \sum_j w_j^T \phi(x_{ji})}{\partial w_j} \tag{38}$$

$$\frac{w_j d_j \sum_j \|w_j\|_2 d_j}{\|w_j\|_2} = \sum_i \alpha_i y_i \phi(x_{ji)} \tag{39}$$

$$w_j = \frac{\|w_j\|_2}{d_j \gamma} \sum_i \alpha_i y_i \phi(x_j i) \tag{40}$$

Similar with previous problem, we could have:

$$w_j = \eta_j \sum_i \alpha_i y_i \phi(x_{ji}), \eta_j > 0 \tag{41}$$

When $w_j = 0$, we could set $\eta_j = 0$. Therefore, we always have:

$$\eta_j >= 0 \tag{42}$$

3 Collaboration

I dicussed with Zheng Chen with problem 1.2 on how to calculate the parameters, and he also informed me about 2.2.1(b).