CS4269/6362 Machine Learning, Spring 2016: Homework 3

Zejian Zhan

Instructions Please fill in your name above.

Question 1:

(a) Because V_c is the volume of the hypercube which has the sphere touching its each side, we can get the formula: $V_c = (2r)^d$ so the ratios of V_s and V_c will be:

$$\lim_{d\to\infty} \frac{V_s}{V_c} = \lim_{d\to\infty} \frac{r^d \pi^{d/2}}{\Gamma(d/2+1)} \Bigg/ (2r)^d = \lim_{d\to\infty} (\pi/4)^{d/2} \frac{1}{\Gamma(d/2+1)} \tag{1}$$

Let $z = d/2$:

$$\lim_{d\to\infty} (\pi/4)^{d/2} \frac{1}{\Gamma(d/2+1)} = \lim_{z\to\infty} (\pi/4)^z \frac{1}{\Gamma(z+1)} \tag{2}$$

From the given equation:

$$\lim_{z\to\infty} \frac{\Gamma(z+1)}{\sqrt{2\pi z} e^{-z} z^z} = 1 \tag{3}$$

we can derive:

$$\lim_{z\to\infty} (\pi/4)^z \frac{1}{\Gamma(z+1)} = \lim_{z\to\infty} (\pi/4)^z \frac{1}{\sqrt{2\pi z} e^{-z} z^z} = \lim_{z\to\infty} (\pi e/4z)^z \frac{1}{\sqrt{2\pi z}} \tag{4}$$

We can see that when z is approaching infinite, the value of $(\pi e/4z)^z$ will approach 0 and $\frac{1}{\sqrt{2\pi z}}$ will also approach 0, thus $\lim_{z\to\infty} (\pi e/4z)^z \frac{1}{\sqrt{2\pi z}}$ is 0.

(b) The curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings. And $Distance\ functions$ is one of the domains curse of dimensionality suggests. The distance between the center and the corners is $r\sqrt{d}$, which increases without bound for fixed r. In other words, the high-dimensional unit hypercube can be said to consist almost entirely of the "corners" of the hypercube, with almost no "middle". And $Distance\ functions$ loses their usefulness because the minimum and the maximum distance between a random reference point Q and a list of n random data points P1,...,Pn become indiscernible compared to the minimum distance: 0.

Question 2: We should choose (c_1, γ_1). The reason is that the both of two parameters achieve the same cross validation error, but (c_1, γ_1) has fewer support vectors, which means the model with parameter (c_2, γ_2) tries to fit more with cross validation data set. So this model is more complex than another one. In this case, it may incur overfitting. Choosing (c_1, γ_1) in the final model can obtain better generalization performance for unknown data set.

Question 3:

(a) We can prove Hinge Loss function is convex by its definition that For all x_1, x_2, and $\lambda \in (0, 1)$, if

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2) \tag{5}$$

then $f(x)$ is a convex function.

We can classify the proof into four scenarios:

Scenario 1: Let $x_1 < x_2 < 1$

$$\begin{aligned} & f(\lambda x_1 + (1-\lambda)x_2) - (\lambda f(x_1) + (1-\lambda)f(x_2)) \\ & = 1 - \lambda x_1 - (1-\lambda)x_2 - \lambda(1-x_1) - (1-\lambda)(1-x_2) \\ & = 0 \end{aligned} \tag{6}$$

Scenario 2: Let $1 \leq x_1 < x_2$

$$f(\lambda x_1 + (1-\lambda)x_2) - (\lambda f(x_1) + (1-\lambda)f(x_2)) = 0 \tag{7}$$

Scenario 3: Let $x_1 < 1 \leq x_2$ and if $\lambda x_1 + (1-\lambda)x_2 < 1$

$$\begin{aligned} & f(\lambda x_1 + (1-\lambda)x_2) - (\lambda f(x_1) + (1-\lambda)f(x_2)) \\ & = 1 - \lambda x_1 - (1-\lambda)x_2 - \lambda f(x_1) \\ & = (1-\lambda)(1-x_2) \leq 0 \end{aligned} \tag{8}$$

Scenario 4: Let $x_1 < 1 \leq x_2$ and if $\lambda x_1 + (1-\lambda)x_2 \geq 1$

$$\begin{aligned} & f(\lambda x_1 + (1-\lambda)x_2) - (\lambda f(x_1) + (1-\lambda)f(x_2)) \\ & = 0 - \lambda(1-x_1) = \lambda(x_1 - 1) < 0 \end{aligned} \tag{9}$$

Based on the above scenarios, we can see that Hinge Loss function $H(a)$ is a convex function of a.

(b) When $\lambda' = 2\lambda$, the two formulars are equivalent. With the new Hinge Loss function, we have:

$$\min_{w,b} \sum_{i=1}^{n} H'(y_i(w^T x_i)) + \lambda' \|w\|_2^2 \tag{10}$$

Then we can rewrite the formula as follows:

$$\begin{aligned}
&\min_{w_\beta,b}\sum_{i=1}^{n}\max(0.5-y_i{w_\beta}^T x_i,0)+\lambda'||w_\beta||_2^2\\
&=\min_{w,b}\sum_{i=1}^{n}\max(0.5-y_i w^T x_i/2,0)+\lambda'||w/2||_2^2\\
&=\min_{w,b}\sum_{i=1}^{n}\frac{1}{2}*2\max(0.5-y_i w^T x_i/2,0)+\lambda'||w/2||_2^2\\
&=\min_{w,b}\sum_{i=1}^{n}\frac{1}{2}\max(1-y_i w^T x_i,0)+\frac{\lambda'}{4}||w||_2^2
\end{aligned} \tag{11}$$

If they are equal, then the condition $\lambda'/4=\lambda/2$ must be satisfied. So $\lambda'=2\lambda$

Question 4:

(a) Increasing d make over-fitting more likely, because as d increases, the training instances will have stronger influence on predicting jobs. Those new unknown instances which are near the each instance of the training set will have very much larger kernel values. However, those new instances which are not that near with any one in training set will has low values. This results in overfitting very likely.

(b) Increasing σ make over-fitting less likely, because σ servers as an amplifier of the distance between x and x'. If the distance between x and x' is much larger than σ, the kernel function tends to be zero. Smaller σ tends to make a local classifier while larger σ tends to make a much more general classifier.

(c) **Prove:**

Assume we have the descriptions about kernel functions:

$$K(x_i,x_{i'})=K_1(x_i,x_{i'})+K_2(x_i,x_{i'})=<\phi^1(x_i),\phi^1(x_{i'})>+<\phi^2(x_i),\phi^2(x_{i'})> \tag{12}$$

$$\phi^1(x)=(\varphi_1^1(x),\varphi_2^1(x),...,\varphi_a^1(x)) \tag{13}$$

$$\phi^2(x)=(\varphi_1^2(x),\varphi_2^2(x),...,\varphi_b^2(x)) \tag{14}$$

Then equation $K(x_i,x_{i'})=K_1(x_i,x_{i'})+K_2(x_i,x_{i'})$ can be derived as:
$\varphi_1^1(x_i)*\varphi_1^1(x_{i'})+\varphi_2^1(x_i)*\varphi_2^1(x_{i'})+...+\varphi_a^1(x_i)*\varphi_a^1(x_{i'})+\varphi_1^2(x_i)*\varphi_1^2(x_{i'})+\varphi_2^2(x_i)*\varphi_2^2(x_{i'})+...+\varphi_b^2(x_i)*\varphi_b^2(x_{i'})$

So that we can rewrite a kernel function as follows:

$$\phi^0(x)=(\varphi_1^1(x),\varphi_2^1(x),...,\varphi_a^1(x),\varphi_1^2(x),\varphi_2^2(x),...,\varphi_b^2(x)) \tag{15}$$

Finally, we can get the form for $K(x_i,x_{i'})$:

$$K(x_i,x_{i'})=<\phi^0(x_i),\phi^0(x_{i'})>. \tag{16}$$

So $K(x_i,x_{i'})=K_1(x_i,x_{i'})+K_2(x_i,x_{i'})$

Question 5:

(a) The computational complexity of prediction of a linear SVM is $O(mn)$.

(b) The computational complexity of prediction of a non-linear SVM is $O(ms)$.