Compsci 571 HW5

Yilin Gao (yg95)

March 16, 2018

1 Hoeffding's Inequality

(a) For all $t \geq 0$,

$$Pr(\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{X_i}) \geq t) = Pr(\sum_{i=1}^{n}(X_i - \mu_{X_i}) \geq nt)$$

Because $X_1, \ldots, X_n$ are independent random variables, $\sum_{i=1}^{n} X_i$ is also a random variable, and its mean is $E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \mu_{X_i}$. Denote $X' = \sum_{i=1}^{n} X_i$, and $\mu_{X'} = E[X']$.

So the above probability is:

$$= Pr(X' - \mu_{X'} \geq nt)$$

According to Chernoff Bounds:

$$\leq \min_{\lambda \geq 0} M_{X'-\mu_{X'}}(\lambda)e^{-\lambda nt} = \min_{\lambda \geq 0} E[e^{\lambda(X'-\mu_{X'})}]e^{-\lambda nt} = \min_{\lambda \geq 0} E[e^{\lambda \sum_{i=1}^{n}(X_i-\mu_{X_i})}]e^{-\lambda nt}$$

$$= \min_{\lambda \geq 0} E[\prod_{i=1}^{n} e^{\lambda(X_i-\mu_{X_i})}]e^{-\lambda nt} = \min_{\lambda \geq 0}(\prod_{i=1}^{n} E[e^{\lambda(X_i-\mu_{X_i})}])e^{-\lambda nt}$$

Apply Hoeffding's Lemma to each X_i:

$$\leq \min_{\lambda \geq 0}(\prod_{i=1}^{n} \exp(\frac{\lambda^2(b-a)^2}{8}))\exp(-\lambda nt) = \min_{\lambda \geq 0} \exp(\sum_{i=1}^{n} \frac{\lambda^2(b-a)^2}{8})\exp(-\lambda nt)$$

$$= \min_{\lambda \geq 0} \exp(\frac{n\lambda^2(b-a)^2}{8} - \lambda nt)$$

Because the exponential function is increasing on its argument, so its minimal value occurs at its minimal argument value.

For $\frac{n\lambda^2(b-a)^2}{8} - \lambda nt$, when $\lambda^* = \frac{4t}{(b-a)^2}$, its minimal value is $-\frac{2nt^2}{(b-a)^2}$.

So we got $\min_{\lambda \geq 0} \exp(\frac{n\lambda^2(b-a)^2}{8} - \lambda nt) = \exp(-\frac{2nt^2}{(b-a)^2})$. $\square$

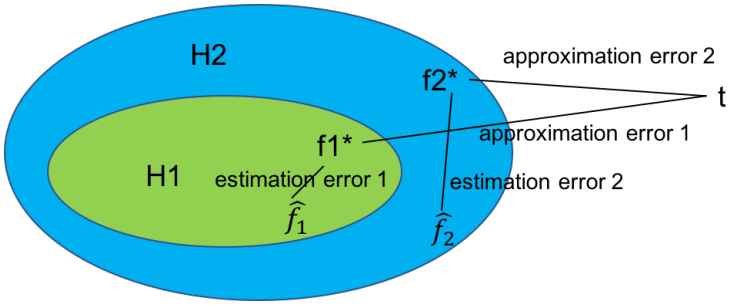
(b)

2 VC Dimension

(a) The VC dimension of $\mathcal{H}_1$ is p, because its feature space is p-dimensional, and all separating hyperplanes are linear in the feature space and goes though the origin point.

The VC dimension of $\mathcal{H}_2$ is $\frac{(p+2)(p+1)}{2}$, because its feature space is $\frac{(p+2)(p+1)}{2}$-dimensional, and all separating hyperplanes are linear in the feature space and goes though the origin point.

(b) The approximation and estimation errors for $\mathcal{H}_1$, $\mathcal{H}_2$ and $\hat{f}_1$, $\hat{f}_2$ are depicted as following:



In the picture, t stands for the global optimal function based on true risk, f_1^* and f_2^* stand for the optimal functions based on true risk in function classes $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively, and $\hat{f}_1$ and $\hat{f}_2$ are optimal functions based on empirical risk in function classes $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively.

As n increases, the estimation error increases because with more data we are able to estimate the model more accurately, and decrease the discrepancy between true risk and empirical risk, i.e., the estimation error. And the approximation error won't be affected significantly because the approximation error is dependent on the function class $\mathcal{H}$, but not the data.

(c) One function class $\mathcal{F}$ such that its VC dimension is not equal to its number of parameters is the set of functions generated by the K-nearest nerighbors algorithm with $K = 1$. It only has 1 parameter K, while its VC dimension is infinity (because with $K = 1$ it classifies every data point into the class same as its true class, thus being able to shatter any number of data points).

3 Ridge Regression

(a) Denote $f(\beta) = \sum_{i=1}^{n} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_2^2$

$$\frac{df(\beta)}{d\beta_j} = -2\boldsymbol{x}_{\cdot j}^T(\boldsymbol{y} - \boldsymbol{X}\beta) + 2\lambda\beta_j$$

$$\frac{\nabla f(\beta)}{\nabla \beta} = -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta) + 2\lambda\beta = 0$$

We get:

$$\hat{\beta}^{ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

in which $\boldsymbol{I}$ is a $p \times p$ diagonal matrix with 1's as its diagonal values.

(b)

(c)