Compsci 571 HW5

Yilin Gao (yg95)

March 21, 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 $X_i, i = 1, \ldots, n$:

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

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

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 $\lambda^* = \frac{4t}{(b-a)^2} (\geq 0)$, 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) Let $X_1, \ldots, X_n$ be independent Bernoulli random variables taking values 1 and -1 each with probability 0.5. For each variable X_i, its mean is $\mu_{X_i} = E[X_i] = 0$.
Its Hoeffding's bound is:

$$Pr(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t) \leq \exp(-\frac{2nt^2}{(1-(-1))^2}) = \exp(-\frac{nt^2}{2}), \text{ for all } t \geq 0$$

And its Bernstein bound is:

$$Pr(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t) \leq \exp(-\frac{nt^2}{2(1+\frac{t}{3})}), \text{ for all } t \geq 0$$

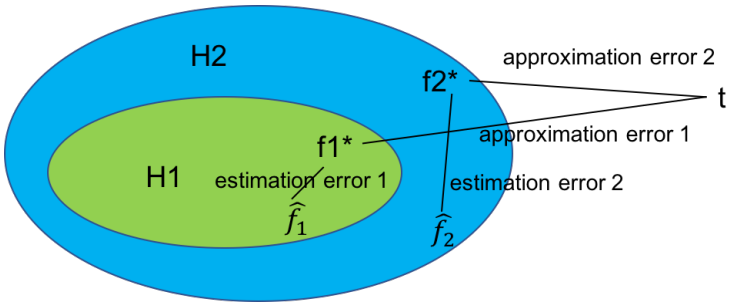
In the two cases, $\exp(-\frac{nt^2}{2(1+\frac{t}{3})}) \geq \exp(-\frac{nt^2}{2})$ for all $t \geq 0$. So the probability of $\frac{1}{n}\sum_{i=1}^{n} X_i \geq t$ is always smaller for Hoeffding's bound, and in other words, the probability of $\frac{1}{n}\sum_{i=1}^{n} X_i < t$ is always larger for Hoeffding's bound, which means the Hoeffding's bound is looser than the Bernstein bound.

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

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

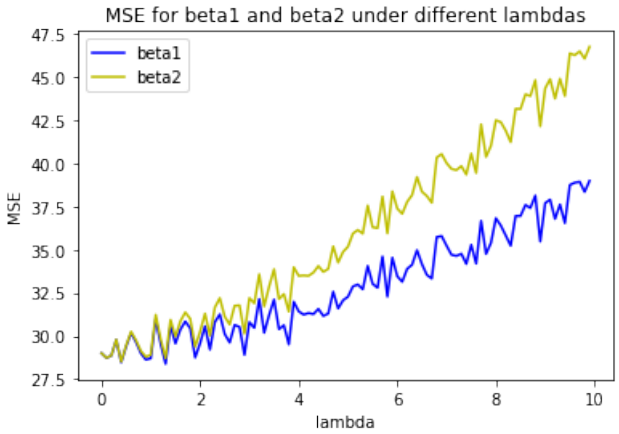
$$\frac{\nabla f(\beta)}{\nabla \beta} = -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\beta) + 2\lambda\beta = 0$$

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) The plot of average ridge MSE with respect to $\lambda \in [0, 10]$ for both β_1 and β_2 is:



From the plot we can see:

(1) As λ increases, the MSEs of both β_1 and β_2 increase. This is because λ will force regression estimated coefficients $\hat{\beta}$ to be not too large to avoid overfitting, thus leading to imprecise prediction. When λ is larger, the bias in estimated coefficient $\hat{\beta}$ will be larger, so will the MSE be.

(2) For the same λ value, the MSE of β_2 is larger than that of β_1 in general. This is because values in β_2^* are larger than those in β_1^*. So their unbiased estimated values will also be larger. With regularization, their estimated values will be shrunk more, so the training MSE will be larger for β_2.

(c) From the problem setting, we can know that $\mathbf{U}_p^T\mathbf{U}_p = \mathbf{I}_p$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}_p$, $\mathbf{L} = \mathbf{L}^T$. And because $\mathbf{X}^T\mathbf{X} = corr(\mathbf{X})$, $\mathbf{U}_p^T$ is orthogonal to $\mathbf{1}_p$, i.e., $\mathbf{U}_p^T\mathbf{1}_p = \mathbf{1}_p^T\mathbf{U}_p = \mathbf{0}_p$.

$$\mathbf{U}^T\mathbf{Y} = \begin{bmatrix} \mathbf{1}_n^T \\ \mathbf{U}_p^T \\ \mathbf{U}_{n-p-1}^T \end{bmatrix} \mathbf{1}_n\alpha + \begin{bmatrix} \mathbf{1}_n^T \\ \mathbf{U}_p^T \\ \mathbf{U}_{n-p-1}^T \end{bmatrix} \mathbf{U}_p\mathbf{L}\mathbf{V}^T\beta + \mathbf{U}^T\epsilon$$

$$= \begin{bmatrix} n\alpha + \mathbf{1}_n^T\mathbf{U}_p\mathbf{L}\mathbf{V}^T\beta \\ \mathbf{U}_p^T\mathbf{1}_n\alpha + \mathbf{U}_p^T\mathbf{U}_p\mathbf{L}\mathbf{V}^T\beta \\ \mathbf{U}_{n-p-1}^T\mathbf{1}_n\alpha + \mathbf{U}_{n-p-1}^T\mathbf{U}_p\mathbf{L}\mathbf{V}^T\beta \end{bmatrix} + \mathbf{U}^T\epsilon$$

$$= \begin{bmatrix} n\alpha \\ \mathbf{L}\mathbf{V}^T\beta \\ \mathbf{U}_{n-p-1}^T\mathbf{1}_n\alpha + \mathbf{U}_{n-p-1}^T\mathbf{U}_p\mathbf{L}\mathbf{V}^T\beta \end{bmatrix} + \mathbf{U}^T\epsilon$$

At the same time, we have the same equation represented by γ:

$$\mathbf{Y}^* = \mathbf{U}^T\mathbf{Y} = \begin{bmatrix} n\alpha \\ \mathbf{L}\gamma \\ \mathbf{0}_{n-p-1} \end{bmatrix} + \mathbf{U}^T\epsilon$$

So we have $\mathbf{L}\mathbf{V}^T\beta = \mathbf{L}\gamma$, or $\gamma = \mathbf{V}^T\beta$.

From the original function $\mathbf{Y} = \mathbf{1}\alpha + \mathbf{U}_p\mathbf{L}\mathbf{V}^T\beta + \epsilon$, we get the objective function of ridge regression:

$$f(\alpha, \beta) = \|\mathbf{Y} - \mathbf{1}\alpha - \mathbf{X}\beta\|_2^2 + \lambda[\alpha^2 + \|\beta\|_2^2]$$

$$\frac{\partial f}{\partial \alpha} = -2\sum_{i=1}^{n}(y_i - \alpha - \mathbf{x}_i\beta) + 2\lambda\alpha = 2(n+\lambda)\alpha - 2\mathbf{1}_n^T(\mathbf{Y} - \mathbf{X}\beta) = 0$$

$$\frac{\nabla f}{\nabla \beta} = -2\mathbf{X}^T(\mathbf{Y} - \mathbf{1}\alpha - \mathbf{X}\beta) + 2\lambda\beta = 0$$

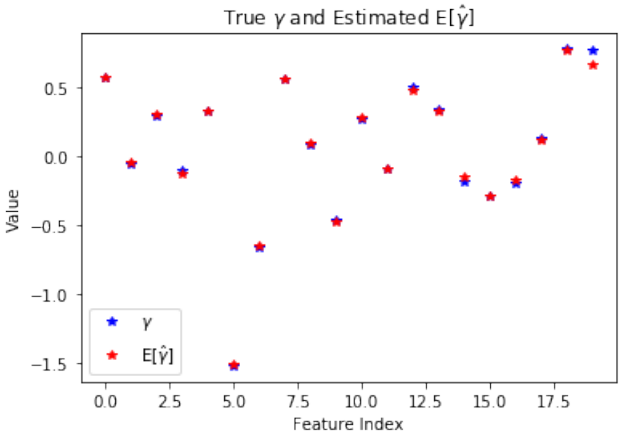
From the 2 equations, we get $\hat{\beta}^{ridge} = [\mathbf{X}^T(\mathbf{I}_n - \frac{1}{n+\lambda}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{X} + \lambda\mathbf{I}_p]^{-1}[\mathbf{X}^T(\mathbf{I}_n - \frac{1}{n+\lambda}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{Y}]$.

Since we have $\mathbf{X} = \mathbf{U}_p\mathbf{L}\mathbf{V}^T$, $\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{L}^2\mathbf{V}^T$, and $\mathbf{X}^T\mathbf{1}_n = \mathbf{V}\mathbf{L}\mathbf{U}_p^T\mathbf{1}_n = \mathbf{0}_p$.

So $\hat{\beta}^{ridge} = [\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_p]^{-1}[\mathbf{X}^T\mathbf{Y}] = \mathbf{V}(\mathbf{L}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{L}\mathbf{U}_p^T\mathbf{Y}$.

And $\hat{\gamma}^{ridge} = \mathbf{V}^T\hat{\beta}^{ridge} = (\mathbf{L}^2 + \lambda\mathbf{I}_p)^{-1}\mathbf{L}\mathbf{U}_p^T\mathbf{Y}$. $\square$

The plot of γ and $E[\hat{\gamma}]$ is:



The plot of $\frac{1}{N}\sum_{i=1}^{N}(\hat{\gamma}_i^{(n)} - \gamma_i)^2$ and $\frac{l_i^2+\gamma_i^2}{(l_i^2+\lambda)^2}$ is: