

Homework 5

2018 Spring STA 561

March 1, 2018

1 Hoeffding's Inequality (20 pts)

a. (15 pts) Chernoff Bounds: Let X be a random variable, for any $t \geq 0$

$$Pr(X \geq \mu_X + t) \leq \min_{\lambda \geq 0} M_{X-\mu_X}(\lambda) e^{-\lambda t},$$

where $\mu_X = \mathbb{E}[X]$ is the mean and $M_X(\lambda) = \mathbb{E}[e^{\lambda X}]$ is the moment generating function.

Hoeffding's Lemma: Let X be a bounded random variable with $X \in [a, b]$. Then

$$\mathbb{E}[e^{\lambda(X-\mu_X)}] \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right), \text{ for all } \lambda \in \mathbb{R}.$$

Use Chernoff bounds and Hoeffding's lemma to prove Hoeffding's inequality

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

where X_1, \dots, X_n are independent random variables with $X_i \in [a, b]$ for all i .

b. (5 pts) Hoeffding's inequality is very loose in certain cases. Please give a simple distribution of X_i where the bound can be much sharper than Hoeffding's bound.

2 VC Dimension (40 pts)

Given data $(x_i, y_i)_i^n$ drawn from a complicated binary classification function. We have the following two kernel functions k_1, k_2 , two hypothesis spaces $\mathcal{H}_1, \mathcal{H}_2$, and two estimators \hat{f}_1, \hat{f}_2 :

The linear kernel: $k_1(\mathbf{u}, \mathbf{v}) = \mathbf{u}^T \mathbf{v}$.

The second order polynomial kernel: $k_2(\mathbf{u}, \mathbf{v}) = (\mathbf{u}^T \mathbf{v} + 1)^2$.

$$\mathcal{H}_1 = (f : f(\mathbf{x}) = \text{Sign}[\sum_{i=1}^N \alpha_i \mathbf{x}_i^T \mathbf{x}])$$

$$\mathcal{H}_2 = (f : f(\mathbf{x}) = \text{Sign}[\sum_{i=1}^N \alpha_i (\mathbf{x}_i^T \mathbf{x} + 1)^2])$$

$$\hat{f}_1 = \arg \min_{f \in \mathcal{H}_1} \frac{1}{n} \sum_{i=1}^n \mathbf{I}(y_i \neq f(\mathbf{x}_i))$$

$$\hat{f}_2 = \arg \min_{f \in \mathcal{H}_2} \frac{1}{n} \sum_{i=1}^n \mathbf{I}(y_i \neq f(\mathbf{x}_i))$$

where $\mathbf{u}, \mathbf{v} \in \mathbb{R}^p, \alpha_i \in \mathbb{R}, \mathbf{x}_i \in \mathbb{R}^p, y_i \in \{0, 1\}, N \in \mathbb{Z}_+$.

a. (10 pts) What is the VC-dimension of \mathcal{H}_1 and \mathcal{H}_2 .

b. (20 pts) Draw a picture for the approximation and estimation error for $\mathcal{H}_1, \mathcal{H}_2$ and \hat{f}_1, \hat{f}_2 and write them down. Explain how the two errors change as n increases. (Hint: you may find the picture and notations in the notes helpful.)

c. (10 pts) Please find at least one function class F where the VC dimension is not equal to the number of parameters of the function class. This will demonstrate that complexity of a function class is not always measured by the number of parameters. (Hint: If you have trouble you can look it up on the Internet. Hint 2: Prof. Rudin will provide an example of this in the lecture that you can use.)

3 Ridge Regression (40 pts)

Given a response vector $\mathbf{y} \in \mathbb{R}^n$ and a predictor matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, the ridge regression coefficients are defined as

$$\hat{\beta}^{ridge} = \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2$$

Here λ is a tuning parameter which controls the strength of the penalty term. When $\lambda = 0$, we get the linear regression estimate.

a. (5 pts) Derive the closed form solution of $\hat{\beta}^{ridge}$.

b. (15 pts) Assume $n = 50$ and $p = 20$ and use the provided \mathbf{X} as input. The response $\mathbf{y} \in \mathbb{R}^{50}$ is drawn from the model $\mathbf{y} = \mathbf{X}\beta^* + \epsilon$, where the entries of $\epsilon \in \mathbb{R}^{50}$ are *i.i.d.* $N(0, 1)$. The true regression coefficients are $\beta_1^* = (0.1, 0.3, 0.2, 0.2, 0.9, 0.8, 0.9, 0.1, 0.4, 0.2, 0.7, 0.3, 0.1, 0.7, 0.8, 0.3, 0.2, 0.8, 0.1, 0.7)^T$, $\beta_2^* = (0.5, 0.6, 0.7, 0.9, 0.9, 0.8, 0.9, 0.8, 0.6, 0.5, 0.7, 0.6, 0.7, 0.7, 0.8, 0.8, 0.9, 0.8, 0.5, 0.7)^T$. Repeat the following $N = 100$ times: 1. Generate a response vector $\mathbf{y}^{(n)}$ for $n = 1, \dots, N$; 2. Compute the estimated coefficients $\hat{\beta}^{(n)}$ use ridge regression; 3. record the error $1/N \sum_{n=1}^N \|\mathbf{y}^{(n)} - \mathbf{X}^{(n)}\hat{\beta}^{(n)}\|^2$. We average the observed error to get the estimated MSE.

Compute and compare the linear MSE for both β_1 and β_2 . Plot the ridge MSE with respect to λ for both β_1 and β_2 . What do you find? Try to explain what you find.

c. (20 pts) This question aims to deal with the matrix inverse problem encountered in ridge regression. \mathbf{X} is the centered and standardized version of the previous question, i.e. $\mathbf{X}^T \mathbf{X} = \text{corr}(\mathbf{X})$. Use $\beta = \beta_1^*$. Suppose $\mathbf{Y} = \mathbf{1}\alpha + \mathbf{U}_p \mathbf{L} \mathbf{V}^T \beta + \epsilon$, $\epsilon \sim N(\mathbf{0}, \mathbf{I}_n)$, where the data $\mathbf{X} \in \mathbb{R}^{n \times p}$ is decomposed as $\mathbf{X} = \mathbf{U}_p \mathbf{L} \mathbf{V}^T$ by singular value decomposition, where $\mathbf{U}_p \in \mathbb{R}^{n \times p}$, $\mathbf{L} \in \mathbb{R}^{p \times p}$, $\mathbf{V} \in \mathbb{R}^{p \times p}$ and $\mathbf{U}_p^T \mathbf{U}_p = \mathbf{I}_p$. \mathbf{L} is diagonal matrix. Let $\mathbf{U} = [\mathbf{1}_n, \mathbf{U}_p, \mathbf{U}_{n-p-1}]$ be an $n \times n$ orthogonal matrix. Then we have $\mathbf{U}^T \mathbf{Y} = \mathbf{U}^T \mathbf{1}_n \alpha + \mathbf{U}^T \mathbf{U}_p \mathbf{L} \mathbf{V}^T \beta + \mathbf{U}^T \epsilon$. If we further define $\mathbf{Y}^* = \mathbf{U}^T \mathbf{Y}$ and $\epsilon^* = \mathbf{U}^T \epsilon$, then

$$\mathbf{Y}^* = \begin{pmatrix} n & \mathbf{0}_p^T \\ \mathbf{0}_p & \mathbf{L} \\ \mathbf{0}_{n-p-1} & \mathbf{0}_{(n-p-1) \times p} \end{pmatrix} \begin{pmatrix} \alpha \\ \gamma \end{pmatrix} + \epsilon^*$$

$\mathbf{0}_p$ is a vector with all zero of length p ($\mathbf{1}_n$ is all one vector of length n). Calculate and write down the estimation of γ using ridge regression in closed form, denote as $\hat{\gamma}$. $\lambda = 1$ and $\alpha = 0.1$. Run $N = 100$ times for different ϵ . Plot γ and $\mathbb{E}[\hat{\gamma}]$ together, where $\mathbb{E}[\hat{\gamma}] = \frac{1}{N} \sum_{n=1}^N \hat{\gamma}^{(n)}$. Then on a different figure, plot $\frac{1}{N} \sum_{n=1}^N (\hat{\gamma}_i^{(n)} - \gamma_i)^2$ and $\frac{l_i^2 + \gamma_i^2}{(l_i^2 + \lambda)^2}$ for $i = 1, \dots, p$ together. Note that $\gamma = [\gamma_1, \gamma_2, \dots, \gamma_p]$ and $\mathbf{L} = \text{diag}(l_1, l_2, \dots, l_p)$.