Homework 5

2018 Spring STA 561

March 22, 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 $X_1, ..., X_n$ 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: $k_1(\boldsymbol{u},\boldsymbol{v}) = \boldsymbol{u}^T\boldsymbol{v}$.
The second order polynomial kernel: $k_2(\boldsymbol{u},\boldsymbol{v}) = (\boldsymbol{u}^T\boldsymbol{v}+1)^2$.

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

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

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

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

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

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

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

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 $\hat{\boldsymbol{\beta}}^{ridge}$.

b. (15 pts) Assume $n = 50$ and $p = 20$ and use the provided $\boldsymbol{X}$ as input. The response $\boldsymbol{y} \in \mathbb{R}^{50}$ is drawn from the model $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$, where the entries of $\boldsymbol{\epsilon} \in \mathbb{R}^{50}$ are *i.i.d.* $N(0,1)$. The true regression coefficients are
$\boldsymbol{\beta}_1^* = (0.1, 0.3, 0.2, 0.2, 0.9, 0.8, 0.9, 0.1, 0.4, 0.2, 0.7, 0.3, 0.1, 0.7, 0.8, 0.3, 0.2, 0.8, 0.1, 0.7)^T$,
$\boldsymbol{\beta}_2^* = (0.5, 0.6, 0.7, 0.9, 0.9, 0.8, 0.9, 0.8, 0.6, 0.5, 0.7, 0.6, 0.7, 0.7, 0.8, 0.8, 0.9, 0.8, 0.5, 0.7)^T$.
Repeat the following $T = 100$ times: 1. Generate a response vector $\boldsymbol{y}^{(t)}$ for $t = 1, \cdots, T$; 2. Compute the estimated coefficients $\hat{\boldsymbol{\beta}}^{(t)}$ using ridge regression; 3. record the error $\frac{1}{T}\sum_{t=1}^{T} ||\boldsymbol{y}^{(t)} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(t)}||^2$, where $\boldsymbol{y}^{(t)}$ and $\hat{\boldsymbol{\beta}}^{(t)}$ are vectors at the t-th iteration.

Compute and compare the linear regression error for both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Plot the ridge regression error with respect to λ for both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. What do you find? Try to explain what you find. (Note: The noise $\boldsymbol{\epsilon}$ should not be fixed during 100 iterations, however it should be the same for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ at each iteration so that we can compare.)

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 $\boldsymbol{\beta} = \boldsymbol{\beta}_1^*$. Suppose $\mathbf{Y} = \mathbf{1}\alpha + \mathbf{U}_p\mathbf{L}\mathbf{V}^T\boldsymbol{\beta} + \boldsymbol{\epsilon}$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}_n)$, 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 $\mathbf{U}_p^T\boldsymbol{U}_p = \mathbf{I}_p$. $\mathbf{L}$ 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 $\mathbf{U}^T\mathbf{Y} = \mathbf{U}^T\mathbf{1}_n\alpha + \mathbf{U}^T\mathbf{U}_p\mathbf{L}\mathbf{V}^T\boldsymbol{\beta} + \mathbf{U}^T\boldsymbol{\epsilon}$. If we further define $\mathbf{Y}^* = \mathbf{U}^T\mathbf{Y}$ and $\boldsymbol{\epsilon}^* = \mathbf{U}^T\boldsymbol{\epsilon}$, then

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

$\mathbf{0}_p$ 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 $\boldsymbol{\gamma}$ using ridge regression in closed form, denote as $\hat{\boldsymbol{\gamma}}$. $\lambda = 1$ and $\alpha = 0.1$. Run $T = 100$ times for different $\boldsymbol{\epsilon}$. Plot $\boldsymbol{\gamma}$ and $\mathbb{E}[\hat{\boldsymbol{\gamma}}]$ together, where $\mathbb{E}[\hat{\boldsymbol{\gamma}}] = \frac{1}{T}\sum_{t=1}^{T}\hat{\boldsymbol{\gamma}}^{(t)}$. Then on a different figure, plot $\frac{1}{T}\sum_{t=1}^{T}(\hat{\gamma}_i^{(t)} - \gamma_i)^2$ and $\frac{l_i^2+\gamma_i^2}{(l_i^2+\lambda)^2}$ for $i = 1, \cdots, p$ together. Note that $\boldsymbol{\gamma} = [\gamma_1, \gamma_2, \cdots, \gamma_p]$ and $\boldsymbol{L} = diag(l_1, l_2, \cdots, l_p)$. (Hint: $\boldsymbol{\gamma} = V^T\boldsymbol{\beta}$)