BIS555 final

December 17, 2021

- *The exam takes place between December 17 6:00 pm ET to December 18 6pm ET. You may use the class text, notes, calculators, and computers, but you may not discuss with other people.*
- *There are five questions for this exam. Problem 1 requires more calculations and mathematical derivations that consist of many intermediate steps. However, if you ever find yourself overwhelmed by the tedious calculation in sub-questions from Problem 1, you may not be on the right track and may want to check if you understand the question correctly.*
- *If you have questions regarding the problem descriptions, I will be available to answer questions via email anytime between December 17 6:00 pm - 11 pm or December 18 10:00 am - 6 pm. Please do not distribute this final to other people.*

1 Problem 1 (30 points)

You have a data set $(\mathbf{X}, \mathbf{y})$ where $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the feature matrix and $\mathbf{y}$ is the length N response. Let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}$ represent the feature variable and response variable. You are interested to predict a feature response y_{n+1} its associated feature x_{n+1}. As a first attempt, you fit a linear regression model and obtain the least square estimate $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$. Your prediction for the test point x_{n+1} is $\hat{y}_{n+1} = x_{n+1}^\top \hat{\beta}$. Although least square fit minimizing the training MSE, you are not sure that it gives you good performance in the test sample. Hence, you decide to investigate its performance compared to other linear predictors where you use a linear model $X\theta$ for prediction and $\theta = M\mathbf{y}$ for some linear transformation matrix M that does not depend on $\mathbf{y}$. You assume that the linear model assumptions hold: $Y = X\beta + \varepsilon$ for some noise i.i.d generated ε (independent of X) with $E[\varepsilon] = 0$ and $E[\varepsilon^2] = 1$.

(1) Consider any construction of θ such that $E[x_{n+1}^\top \theta] = x_{n+1}^\top \beta$ is unbiased for the underlying signal for all β, how do you compare the expected square error loss at x_{n+1} using $\hat{\beta}$ and θ?

(1.1) (3 points) Show that $E[\hat{\beta}] = \beta$ and derive the formula for $E[(y_{n+1} - x_{n+1}^\top \hat{\beta})^2]$.

(1.2) (2 points) Using bias-variance decomposition, argue that

$$E[(y_{n+1} - x_{n+1}^\top \hat{\beta})^2] - E[(y_{n+1} - x_{n+1}^\top \theta)^2] = \mathrm{Var}(\mathrm{x}_{\mathrm{n+1}}^\top \hat{\beta}) - \mathrm{Var}(\mathrm{x}_{\mathrm{n+1}}^\top \theta).$$

(1.3) (5 points) You can write $\theta = \hat{\beta} + h = \hat{\beta} + M_h \mathbf{y}$, where $M_h \in \mathbb{R}^{p \times n}$ is some $\mathbf{y}$-independent linear transformation that defines h. Show that you can decompose the variance $\mathrm{Var}(\mathrm{x}_{\mathrm{n+1}}^\top \theta)$ as

$$\mathrm{Var}(\mathrm{x}_{\mathrm{n+1}}^\top \theta) = \mathrm{Var}(\mathrm{x}_{\mathrm{n+1}}^\top \hat{\beta}) + \mathrm{Var}(\mathrm{x}_{\mathrm{n+1}}^\top \mathrm{h}) + 2\mathrm{x}_{\mathrm{n+1}}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X} \mathrm{M}_{\mathrm{h}}^\top \mathrm{x}_{\mathrm{n+1}},$$

(1.4) (5 points) Show that $x_{n+1}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top M_h^\top x_{n+1} = 0$ when $E[x_{n+1}^\top \theta] = E[x_{n+1}^\top \beta]$. Hence, you always have $\mathrm{Var}(\mathrm{x}_{\mathrm{n+1}}^\top \theta) \geq \mathrm{Var}(\mathrm{x}_{\mathrm{n+1}}^\top \hat{\beta})$ and the OLS solution thus offers a best linear unbiased estimate for predictions. Here, you can use the relationship that if $\zeta^\top \beta = 0$ for all β, we must have $\zeta = 0$:

$$\zeta^\top \beta = 0 \text{ for all } \beta \Leftrightarrow \zeta = 0.$$

(2) Consider a linear regression with ridge penalty:

$$\frac{1}{2} \|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda \|\beta\|_2^2,$$

Let $\mathbb{I}_p$, $\mathbb{I}_n$ denote the $p \times p$ and $n \times n$ identity matrices, and the solution to the ridge regression is $\hat{\beta}_{ridge} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbb{I}_p)^{-1} \mathbf{X}^\top \mathbf{y}$. Consider the case where $\mathbf{X}^\top \mathbf{X} = n\mathbb{I}_p$ and $p = \frac{n}{2}$ (n is even).

(2.1) (3 points) Show that we can write $\hat{\beta}_{ridge}$ as $\hat{\beta}_{ridge} = a_1\beta + a_2\mathbf{X}^\top\varepsilon$ for some a_1, a_2 that depends only on λ and n, and $\varepsilon = \mathbf{y} - \mathbf{X}\beta$ is the noise in the training samples.

(2.2) (5 points) Calculate the average expected loss on the training $\mathbf{X}$: let $\mathbf{y}_{new} = \mathbf{X}\beta + \varepsilon_{new}$ be some new realizations generated at $\mathbf{X}$, derive the expression for $\mathrm{Err}(\hat{\beta}_{ridge})$ defined below.

$$\mathrm{Err}(\hat{\beta}_{ridge}) = \frac{1}{N}E[\|\mathbf{y}_{new} - \mathbf{X}\hat{\beta}_{ridge}\|_2^2] \tag{1}$$

(2.3) (2 points) Calculate $\mathrm{Err}(\hat{\beta})$, the average expected loss using the least square estimate $\hat{\beta}$ as in equation (1).

(2.4) (5 points) Compare $\mathrm{Err}(\hat{\beta}_{ridge})$ with $\mathrm{Err}(\hat{\beta})$: Suppose that $\|\beta\|_2^2 = 1$, show that when $0 < \lambda \leq \lambda_0$ for some value $\lambda_0 > 0$, $\mathrm{Err}(\hat{\beta}_{ridge}) < \mathrm{Err}(\hat{\beta})$. Does contradict with what you derived in part 1?

2 Problem 2 (10 points)

An economist has a collection of data on startup companies in the software sector in Silicon Valley. She has several sets of measurements for each company: 11 variables aimed at measuring the growth potential of the company, 5 variables measuring educational levels and quality of their research team, and 7 variables measuring the saturation of the relevant markets in which the companies compete. The outcome is binary–whether or not the company was acquired/IPO-ed within 10 years. She uses a logistic regression model, and finds (via cross-validation) that it is strongly predictive. However, to her horror none of the variables has a significant coefficient. What has happened and can you give her any suggestion for identifying sources that are predictive of the response?
[Explanation of the phenomenon - 5 points; Suggestions - 5 points]

3 Problem 3 (20 points)

You fit a big logistic regression model to a GWAS dataset. Your binary outcome is `diabetes` (present or absence), and you have variables: `ethnicity` (4 categories), and binary vector of `genotypes` at 274 SNPs (wildtype or not). Since there are so many SNPs, you decide to impose a ridge penalty on their coefficients, but leave the coefficients for ethnicity unpenalized. You hence fit the model

$$\text{logit}(p_i) = \text{logit}[\Pr(Y_i = 1|E_i, X_i)] = E_i'\beta_E + X_i'\beta_X.$$

Here E_i is a 4-element binary vector coding `ethnicity` for subject i, and X_i a 724 element binary vector representing `genotype`. You fit the model by maximum penalized likelihood:

$$\max_{\beta_E,\beta_X} \sum_{i=1}^{N} [y_i \log p_i + (1 - y_i) \log(1 - p_i)] - \lambda \|\beta_X\|_2^2.$$

Let $\hat{p}^\lambda(x)$ be the fitted function for probability estimation on any observation x, $\hat{p}_i^\lambda$ be the fitted probability for observation i, and denote by r_j the proportion of ethnic-class j in the training data having diabetes ($j = 1 \ldots 4$): $r_j = \frac{m_j}{n_j}$ where m_j is the number of people having diabetes in ethnic group j and n_j is the total number of people in group j.

(1) Show that the average fitted probability in ethnic group j is equal to the observed proportion r_j, irrespective of λ. [10 points]

(2) You want to apply this model on a new population cohort that contain only the first ethnic group. However, due to some reasons (e.g., this group of people have have different occupations compared with the old group and thus different living style), has a lower diabetes prevalence with $P(Y = 1) = 0.1$ for this new cohort where $Y = 1$ representing diabetes, while $P(Y = 1) = 0.2$ for the old cohort. Can you account for this while making predictions? Plase explain how and why. (Using the given diabetes prevalences and the trained prediction function $\hat{p}^\lambda(x)$.) [10 points]

4 Problem 4 (20 points)

(1) Your friend is analyzing a data set for the hospital that are collected from patients with skin cancer. The data contains mass spectrometry imaging measurements measured at human specimens collected from different patients. There are in total 8000 training samples from 12 patients (each patient have many specimens collected), and a independent test data set of 2000 samples from another five patients. He wants to to predict if a given specimen is normal or having cancer. He uses standard 10 fold cross-validation and obtains a CV error rate of 2%. But when he applies the model to the test set, the error rate is 15%. He is bothered and comes to you for help. Please explain to him what has happened and offer a fix, if possible. [10 points]

(2) Your friend comes for help again. This times, he also has a problem with the discrepancy between CV and test errors, but for some different reasons. He carries out a forward stepwise algorithm for best-subset linear regression model as described in the notes, and select the subset size by 10-fold cross-validation. This procedure has a minimum CV error of 0.27, and on the left-out test set you report a RMSE of 0.32. However, he thinks this model might be a bit deficient, and is troubled by the fact that in each of the 9/10th folds, the best subsets of each size do not necessarily contain the same variables. Hence for every one of the 2^p subset models (p = number of features), he estimates the RMSE by cross-validation (using the same folds as before), and finds a model with CV error of 0.19. He is really excited by the improvement. However, he is aghast to find that his model has test error of 0.39. Can you offer an explanation for the discrepancy? [10 points]

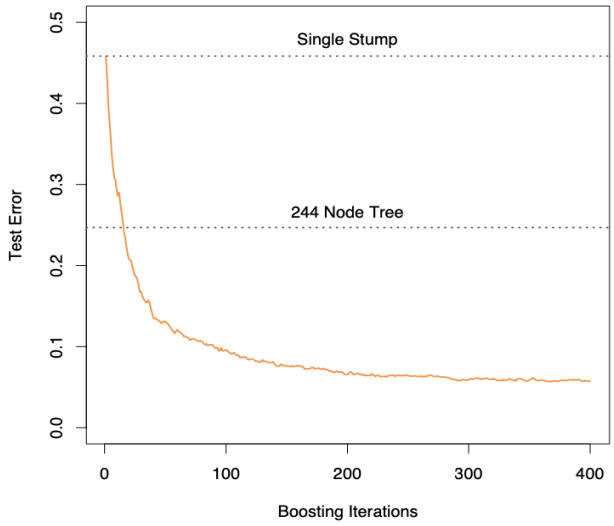
5 Problem 5 (20 points)

(1) Bagging usually does not work well with "weak" learners. In contrast, we often want to use "weak" learners when boosting. Why? (12 points)

(2) Consider the example we discussed in class: The features $X_1, \ldots, X_{10}$ are standard independent Gaussian, and the deterministic target is defined by

$$Y = \begin{cases} 1 & \text{if } \sum_j X_j^2 > \chi^2_{10}(.5) \\ -1 & \text{otherwise} \end{cases}$$

Consider a stump classifier (a tree with two terminal-node), a large tree classifier with 244 Node and Adaboost with stump individual learner. Figure below shows their error rate on the test data.



Why is Adaboost better than the single stump? (3 points) Why is it better than a large classification tree? (5 points)