STA4241 Midterm, Fall 2021

Please turn in your own work, and you may **NOT** discuss your exam with your classmates. Your exam should be uploaded to canvas by 11:59pm on Friday, October 22nd. You should upload two files when submitting your exam: 1) Your written responses to each question and any relevant output, plots, etc. 2) Your .R script used to answer each question. Please do not include any R code in your exam writeup unless otherwise asked to. The .R script should be commented and indented appropriately so that it is easy to follow your logic and understand the steps you took. The cleaner your R code is, the easier it will be for me to see if you did something that was close to correct, and I'll be able to give you more partial points. Please do not send me .Rmd files or code pasted into text documents, such as word or PDF documents.

Note that unless otherwise stated, you can use any built-in R functions to perform these tasks. For instance, you may use built in codes for cross-validation if they exist for the method that you're using.

Note this exam is difficult and long, so get started on it early in the week, and don't get discouraged if you're having trouble. Good luck and feel free to email me with any clarifying questions you have.

(1) For the questions below, give 2-5 sentences explaining your answer and rationale.

(i) (3 points) Describe the curse of dimensionality and explain why it affects nonparametric approaches more than parametric ones.

(ii) (3 points) Logistic regression, LDA, and the linear support vector classifier are all classifiers that are linear in $\boldsymbol{X}$. This means that you classify a binary outcome based on the value of $\beta_0 + \sum_{j=1}^{p} X_j\beta_j$. Because of this, do these classifiers lead to the same predictions? If so, explain why. If not, explain what differs across these algorithms.

(iii) (3 points) Suppose I use an SVM with a radial kernel for a number of γ values, and obtain the error rates on the training data given in the table below. Do you think that overfitting is occurring?

	training data error rates
$\gamma = 0.01$	0.40
$\gamma = 0.1$	0.37
$\gamma = 1$	0.31
$\gamma = 5$	0.25
$\gamma = 10$	0.21

(iv) (3 points) Suppose I want to fit a penalized regression model, but my covariates are on a different scale. For instance, $\text{Var}(X_1) = 1/100$ and $\text{Var}(X_2) = 100$. Is this problematic for penalized regression? Explain your answer.

(v) (3 points) Lasso regression naturally performs variable selection as it forces some coefficients to be exactly zero. Are there any situations when lasso might be expected to perform worse (in terms of prediction error) than a traditional variable selection technique such as best subset selection? If yes, describe such a situation. If not, explain why we would expect lasso to outperform best subset selection in general.

(vi) (3 points) Suppose I do not care about minimizing MSE. Instead my goal is to minimize $P(|Y - \widehat{Y}| > 2)$. How would I choose tuning parameters or select the model that gives me the best performance with respect to this metric.

(2) Suppose we observe independent random variables $X_i \sim \mathcal{N}(\mu, \sigma^2)$, for $i = 1, \dots, n$. We will use $\overline{X}$ as an estimator for the unknown μ. Suppose that we want to use the bootstrap to estimate the variance of $\overline{X}$. Denote each bootstrap sample as $X_i^{(b)}$ for $i = 1, \dots, n$, and each bootstrap replicate of the sample mean by $\overline{X}^{(b)}$.

(i) (2 points) Calculate $\text{Var}(\overline{X})$.

(ii) (4 points) Let $\boldsymbol{X} = (X_1, \dots, X_n)$ denote the original sample of data. Calculate $E(X_i^{(b)}|\boldsymbol{X})$. [Hint: We are only conditioning on $\boldsymbol{X}$ because the bootstrap uses the original data to generate new data sets. Think about what distribution the bootstrapped data sets come from and calculate this expectation directly.]

(iii) (7 points) Calculate $\text{Var}(\overline{X}^{(b)}|\boldsymbol{X})$

(iv) (4 points) Calculate $E_{\boldsymbol{X}}\left[\text{Var}(\overline{X}^{(b)}|\boldsymbol{X})\right]$. The notation $E_{\boldsymbol{X}}$ means the expectation with respect to $\boldsymbol{X}$. In other words, the original data set is now the random variable we are averaging over.

(v) (3 points) Are your answers to (i) and (iv) the same? What does this tell you about the performance of the bootstrap for estimating the uncertainty in $\overline{X}$?

(vi) (7 points) Now use simulation to empirically calculate $E_{\boldsymbol{X}}\left[\text{Var}(\overline{X}^{(b)}|\boldsymbol{X})\right]$ when $\mu = 10, \sigma^2 = 100$, and vary $n \in \{10, 20, 30, 40, 50\}$. Note that you can do this even if you were not able to derive the answers above. Describe in sufficient detail how you will use simulation to calculate this quantity and then run the simulation in R to produce a plot that has the sample size on the x-axis and $E_{\boldsymbol{X}}\left[\text{Var}(\overline{X}^{(b)}|\boldsymbol{X})\right]$ on the y-axis for the five values of n considered.

(3) In this question we are going to use simulation to explore how well certain high-dimensional approaches perform with correlated predictors. Throughout, I want you to set $n = 100$ and $p = 200$. I want you to generate your covariates from a multivariate normal distribution with mean vector given by $\mathbf{0}_p$ and covariance $\boldsymbol{\Sigma}$ defined such that the (i, j) element of $\boldsymbol{\Sigma}$ is given by $\Sigma_{ij} = \rho^{|i-j|}$. You will generate data from the following linear regression model:

$$Y_i = \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1).$$

Further, for your true coefficient vector, I want you to use

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \mathbf{0}_{38}, \beta_{41}, \beta_{42}, \mathbf{0}_{38}, \beta_{81}, \beta_{82}, \mathbf{0}_{38}, \beta_{121}, \beta_{122}, \mathbf{0}_{38}, \beta_{161}, \beta_{162}, \mathbf{0}_{38})$$

(i) (8 points) Explain what values you chose for the nonzero coefficients $(\beta_1, \beta_2, \beta_{41}, \beta_{42}, \dots)$. Then run a simulation study comparing the performance of ridge regression and lasso regression for estimating the unknown coefficients as a function of the correlation, ρ. Our interest will be the mean squared error of the estimates, defined by

$$\text{MSE} = E\left[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right]$$

Your answer should include a plot that has ρ on the x-axis and the MSE on the y-axis, and there should be a line for both ridge regression and lasso regression.

(ii) (3 points) Interpret your findings from part (i) about the effectiveness of ridge regression and lasso regression as the correlation increases among the covariates.

(iii) (8 points) Now we will focus on lasso regression and examine the variable selection properties of this estimator. Remember that a variable is included in the lasso regression model if it's corresponding coefficient estimate is nonzero. For each covariate, keep track of the proportion of simulated data sets that it is chosen by the lasso model as being nonzero. Make a plot that has the covariate index (1-200) on the x-axis, and has the corresponding probability of being nonzero on the y-axis. Do this for $\rho \in \{0.5, 0.75, 0.9, 0.99\}$, so you should have four plots: one for each value of ρ.

(iv) (3 points) Comment on your findings from part (iii). How does the lasso perform as a variable selection tool as the correlation increases in the covariates?

(v) (4 points) Are there any estimators that might perform better than the lasso in terms of variable selection for part(iii)? Explain. [Note that you do not need to run any simulations for this question. You can simply answer in words.]

(4) For this problem, read in the data sets titled problem4training.csv and problem4testing.csv. All models should be fit only on the first of these two data sets (training data) and the testing data should only be used if I ask you to evaluate the predictive performance of a model on testing data. In both data sets there is an outcome Y that is binary and a set of continuous predictors (X_1, X_2, ..., X_p)

(i) (3 points) Fit the following probit regression model with main effects only

$$P(Y = 1|\boldsymbol{X}) = \Phi\bigg(\beta_0 + \sum_{j=1}^{p} X_j\beta_j\bigg)$$

Use this model to classify the outcome on the testing data set and report the out of sample MSE.

(ii) (5 points) Construct a 77% confidence interval for $\beta_1 + e^{\beta_2} + \beta_3^2$. Explain how you got this confidence interval.

(iii) (3 points) Is the confidence interval you constructed above unique in the sense that it is the only 77% confidence interval you could create using your chosen method of constructing a confidence interval? If yes, explain why it is unique. If not, provide a different confidence interval from the one in part (ii) using the same method.

(iv) (4 points) Without using information from the outcome, do you think that principle components analysis would improve predictions for this particular data set. Explain why or why not.

(v) (3 points) Use forward stepwise regression and determine which covariates should be included. Does this model have better out of sample MSE than the full model in (i)?

(vi) (4 points) For this question refer back to the reduced model identified in part (v) above. If you were not able to complete part (v) you can use the full main effects model from part (i) for this question. Do you think there should be any interaction terms added to your model? You need to 1) Explain how you would investigate this, and 2) implement the procedure and determine which, if any, interaction terms should be included.

For the remaining questions we will explore other approaches to classification. Do not consider the probit regression models above or the results in the above questions when answering these questions.

(vii) (4 points) Find the testing data error rate for each of 1) LDA, 2) QDA, 3) a support vector machine with a radial kernel and γ chosen via cross validation, and 4) a support vector machine with a polynomial kernel with the degree chosen via cross validation. Display the results in a table.

(viii) (3 points) Make a plot similar to the plot from problem 2 (i) on homework 2 that illustrates the decision boundary for covariates X_1 and X_2 for both of your support vector machine models (radial kernel and polynomial kernel). To make the plot, fix the remaining covariates (X_3 through X_p) to zero.