93 Well done!

ST 563 601 – SPRING 2025 – POST
Exam #1

Student's Name: VIDYUL JAIN

Date of Exam: Thursday, February 6, 2025 - Friday, February 7, 2025
Time Limit: 75 minutes
Allowed Materials: None (closed book & closed notes)

Student – NC State University Pack Pledge

I, VIDYUL JAIN have neither given nor received unauthorized aid on this exam or assignment. I have read the instructions and acknowledge that this is the correct exam.

Student's Printed Name

02/07/2025

Student Signature *Date*

Exam must be turned in by: 3:32 PM NJ.

Exam End Time *Student's Initial Agreement*

NOTE: Failure to turn in exam on time may result in penalties at the instructor's discretion.

Exam 1

Please write your answers below each question. You should not have access nor use any materials during this exam.

A reminder that, by taking this exam, you are required to uphold the NC State honor pledge:

"I have neither given nor received unauthorized aid on this test or assignment."

1. In the statistical learning paradigm, we discussed three major goals: statistical inference, predictive modeling, and pattern finding.

   Give a brief real world example for each of these goals. Specify a possible model or method we discussed in class that would help answer the question from each real world example.

- Statistical Inference (4 pts)

The goal of statistical inference is to evaluate / measure the significance of the data in the computation of an output. A simple example could be understanding what features among possible ones like location, neighbourhood, etc contribute most towards predicting the price of a house. Eg. hypothesis testing, R^2, MSE, etc ok

- Predictive Modeling (4 pts)

In predictive modeling, our goal is to be able to accurately predict a future value based on historical data. Here, the goal is not much about understanding feature importance but to accurately predict future values based on model's learning on training data. Eg. Linear Regression model or a KNN Regression to predict future home prices in a particular city.

- Pattern Finding (4 pts)

Pattern finding is a subset of unsupervised learning wherein we don't have any output/response variable. The goal here is to be able to extract pattern in our data. Some methods in class include clustering, PCA, etc. Eg. We want to detect patterns in the stocks in the S&P 500 index and group them based on their historic price performance.

↳ Sounds like a response variable..

-1

2. Consider having models characterized by flexibility with the scale going from not very flexible to very flexible.

   a. What type of relationship between flexibilty and squared bias would we expect? Why? (4 pts)

As flexibility of a model increases, its bias reduces since as we increase flexibility, we allow the model to become more complex by either adding more predictors or higher degree terms or interactions, thereby reducing its bias and increasing its variance.

   b. What type of relationship between flexibilty and variance would we expect? Why? (4 pts)

As we increase model flexibility, we allow the model to become more complex and fitting to the patterns in the data more closely. This increases the variance as the model becomes more complex with either added features, higher degree terms or interactions.

   c. What type of relationship between flexibilty and training error would we expect? Why? (4 pts)

A model with less flexibility has a higher bias leading to a very simple model that isn't able to capture any pattern in the data. As we increase model flexibility, training error reduces as we are able to capture patterns in the data by allowing more complexity. In general we can expect low training errors as we increase model flexibility.

   d. What type of relationship between flexibilty and test error would we expect? Why? (4 pts)

For low model flexibility, we have a high bias leading to underfitting and hence high test error. As we allow model flexibility to increase and test error to reduce. For very high flexibility in model, our model starts to overfit thereby again increasing test error since the model starts to overfit on training data thereby fitting on the noise, leading to poor test errors.

3. What is a tuning parameter or hyperparameter? How does this differ from a 'regular' parameter in a parametric model? (4 pts)

A tuning parameter is a parameter in model that has to be chosen externally to implement the model and capture the relationship between the response & predictors. Eg. Number of neighbors (K) in the KNN model. A parametric model assumes a structure to the data and hence has predefined parameters to fit that assumed data. Hyper-parameters exist for non-parametric models that allow more flexibility in terms of determining structure in the data.

5. In the multiple linear regression setting, we discussed a number of model selection methods. State four model selection methods that can be used in the $p > n$ situation. (4 pts)

1. Ridge Regression
2. Lasso Regression
3. Principal component ~~Analysis~~ regression ok
4. Elastic Net Regression.

6. State true or false (no need to explain). (3 pts each)

   a. Ordinary least squares performs variable selection. FALSE

   b. Ordinary least squares performs shrinkage of coefficient estimates. FALSE

   c. Best subset selection performs variable selection. TRUE

   d. Best subset selection performs shrinkage of coefficient estimates. FALSE

   e. Ridge Regression performs variable selection. FALSE

   f. Ridge Regression performs shrinkage of coefficient estimates. TRUE

   g. LASSO performs variable selection. TRUE

   h. LASSO performs shrinkage of coefficient estimates. TRUE

7. Suppose we have a large data set where we want to perform a regression task. We want to determine the best overall model between a LASSO model and a kNN regression model. We want to use a train test split and compare the best kNN and LASSO model on the test set. We wish to determine the appropriate tuning parameters on the training set only using cross-validation. Fully outline the process for splitting the data, tuning, comparing, and fitting a final overall best model. (10 pts)

1. Split the data using train-test split in ~~70:30~~ / 80:20 / 60:40 ratio.

2. Use the training set to further split into internal training set and validation set. This split depends on the kind of cross validation selected. Eg. K-fold CV / LOOCV. Assume, we use a K-fold CV. Here we divide the training set into K-folds, training our hyperparameters on the K-1 folds and testing on the K^{th} fold. We repeat this step multiple times to compute best performance for different values of our hyper-parameter. (repeat using each fold as test set once) −1

3. Once we select the best hyperparameter using the internal training and validation sets, we select that hyper-parameter to fit our model on the entire training data.

4. For KNN, we find the optimal value of K and for lass we find optimal value of λ

5. We fit both models on the entire training set with optimal hyper-parameters.

6. Test both models on the unseen test set and compare their performance.

7. Select the model that gives the lowest test error & fit on the entire dataset & use it for predicting future values. −1

8. Consider the Ridge Regression procedure for fitting a multiple linear regression model. With this model we minimize the following criterion (recall $\lambda \geq 0$):

$$\sum_i \left(Y_i - \beta_0 - X_{i1}\beta_1 - \cdots - X_{ip}\beta_p\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

a. What are the benefits of fitting a Ridge Regression model as compared to an ordinary least squares model? (4 pts)

- Ridge regression allows us to shrink parameter coefficients towards zero thereby reducing the effect of not-so important features.
- less computation power needed to train the model.
- Increasing accuracy of model performance by selecting important features and reducing effect of redundant features.
- Allows for greater transparency in model to inference. ?

-3

b. What happens to our coefficient estimates for a 'large' value of the tuning parameter? What happens for a tuning parameter value near 0? (4 pts)

- For a larger value of λ, we increase the penalty on our coefficient estimates and hence shrink these coefficients to values near zero.
- For $\lambda = 0$, our ridge regression problem converges to an ordinary least squares regression problem, thus reducing the penalty value to 0.

-3

9. Suppose we fit a multiple linear regression model to data about how much people earn. Our response variable is the wage (in 1000's of dollars) and our predictors are `marital_status` (`married`, `never_married`, or `divorced`), and age.

   We fit a linear and quadratic term for age and include an interaction between `marital_status` and age and an interaction between `marital_status` and age squared in the model. Output for the model is given below.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept) λ_0	25.293	38.116	0.664	0.507
marital_statusmarried λ_1	-19.780	40.405	-0.490	0.624
marital_statusnever_married λ_2	-31.760	40.992	-0.775	0.439
age λ_3	2.846	1.611	1.767	0.077
I(age^2) λ_4	-0.024	0.017	-1.470	0.142
marital_statusmarried:age λ_5	2.024	1.716	1.179	0.238
marital_statusnever_married:age λ_6	2.230	1.820	1.225	0.221
marital_statusmarried:I(age^2) λ_7	-0.025	0.018	-1.412	0.158
marital_statusnever_married:I(age^2) λ_8	-0.032	0.020	-1.607	0.108

   a. Write down the fitted equation for $\hat{y}$. Define any indicator variables as needed. (4 pts)

$$\hat{y} = 25.293 + (-19.780)\lambda_1 + (-31.760)\lambda_2 + 2.846\lambda_3 + (-0.024)\lambda_4 + 2.024\lambda_5 + 2.230\lambda_6 + (-0.025)\lambda_7 + (-0.032)\lambda_8.$$

   b. One column of the output represents the t-value or t-statistic. What is the usefulness of this t-value? (2 pts)

t-statistic helps us in formulating confidence intervals for our estimates based on probabilities that gives us an idea of the range of possible values of our estimates based on a certain confidence level. −1

−1

c. Write down the form of a predicted value for somone that is `married` and has an age of 30. No need to simplify. (2 pts)

$$\hat{y} = 25.293 + (-19.780)(1) + (-31.760)(0) + 2.846(30) + (-0.024)(30^2) + 2.024(30)(1) + 2.230(30)(0) + (-0.025)(30^2)(1) + (-0.032)(30^2)(0).$$

d. Write down the form of a predicted value for somone that is `divorced` and has an age of 30. No need to simplify. (2 pts)

$$\hat{y} = 25.293 + (-19.780)(0) + (-31.760)(0) + 2.846(30) + (-0.024)(30^2) + 2.024(30)(0) + 2.230(30)(0) + (-0.025)(30^2)(0) + (-0.032)(30^2)(0)$$

f. Conceptually, what does including an interaction between `marital_status` and age and an interaction between `marital_status` and age squared do to our model as compared to a model without those interactions (that still includes a main effect for `marital_status` and a linear and quadratic term for age)? (3 pts)

Including an interaction between marital-status & age & between marital status & age^2 says that the effect of age & age^2 on the response differs depending whether the person is married, not married or divorce. A model without these interactions doesn't assume any effect of these predictors together on the response. Each predictor independently has an effect on the response variable.

g. The F-statistic for the global model test is 46.26 on 8 numerator and 2991 denominator degrees of freedom. The p-value for the test is very close to zero.

i. Write down the null and alternative hypotheses for this global test. (3 pts)

$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$

H_1: At least one of $\beta_j \neq 0$ for $j = 1, \ldots, p$.

ii. We see a significant global test but none of the coefficient tests are significant. What do you think could be causing this issue? (3 pts)

This means that the model in general is significant for predicting the response but the variables selected are not significant. The existence of interaction terms and higher degree terms in the predictors might be a reason for seeing this in the tests. because.. — /

h. What type of plot might we look at to investigate the homogenous error variance (i.e. the assumption of equal error variance)? (3 pts)

Residual plots → plotting residuals vs the fitted values. To see normality assumption of our error we can also plot a Q-Q plot. — /