71

A.12

ST 563 601 – SPRING 2025 – POST
Exam #1

Student's Name: Jaswinder Kaur

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, _____ 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

Jaswinder Kaur

Student Signature *Date*

Exam must be turned in by: _____

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)

Statistical inference can be used to determine the relationships ~~among~~ different predictors. Determining the predictors that are majorly effecting the response variable.
One such example :- Using different traits like seed number, spike length to determine their effect on yield. Here yield is response variable & all different traits are predictor variable we can use Regression models
eg:- simple linear regression / multiple linear regression

- Predictive Modeling (4 pts)

Here in such cases, we are not much concerned about determining the underlying relationship between different variable instead we want to use all the predictor variable to predict the accurate response variable on an unknown test dataset
eg:- Predicting total yield in wheat cultivar using no. of grains per plot, spike length etc.
for these we can use machine learning models such as? -2

- Pattern Finding (4 pts)

-1 eg:- These are basically classification problem. which involve clustering different items based on the common pattern they share
eg:- Clustering the customers based on their age to determine the demographic that response better to deal.
we can use PCA for these & can use clustering models.
response variable -1

-4

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)

Flexible models tends to work best for dataset with non-linear relationship & bigger dataset, as flexibility decreases the squared bias tends to increase because as flexibility decreases models tend to perform & give more correlated prediction leading to increased bias

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

With increase in flexibility variance increases, because model tend to learn the structure of data in better way, leading to overfitting & even picking knitty gritty details & Hence variance increases

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

With increase in flexibility of model training error decreases as model learns about the data patterns & since over time it start overfitting & hence further decreasing the training error

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

In flexible models test error will increase because, the model overfits in training set so peiforms worst on test set. So as flexibility increases, Test error increases in shape -2

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

Tuning parameter or hyperparameter are those parameters which are used to train the training subset of data. When fitting models the data is split into training & testing subset. the training subset is further splitted into validation & internal training subset. Hyperparameters are used to find best model for this internal training subset. Once we got hyperparameter eg $k=30$, which gives minimum test MSE. The's hyperparameter is then used in same model to test on testing subset & further entire dataset for prediction OK

-2

whereas regular parameter are either slope or inter cept showing effect of one variable on another. This hyperparameters are only for tuning the traning & test dataset

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)

→ Non - flexible model → LASSO

→ Ridge regression

→ forward selection

→ Reverse selection

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.

   False

   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

-4

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 Splitting entire dataset into testing & Training dataset (usually 80/20, 70/30 or 60/40)

② a) ~~Training data set is further split into training & validation set.~~

b) Now this training set is further split in folds for cross validation. 5 folds where in each model 4 folds used for training one left for testing. rotate fold left out

c) In these appropriate value of k will be chosen. we create a grid containing values of k from 1 to 100. each model will be run for entire grid & the output from each model we get test MSE for all models MSE will be average & best value of k that gives best test MSE will be chosen.

d. This same value of k will be tested on validation Test if needed the steps from b to d can be repeated but don't touch original testing dataset.

3) Once we got accurate value of k, it will be tested on tested dataset & Based on test MSE → we can either be happy & do it over entire dataset
↳ or repeat

how is this

④ same will be repeated with LASSO, for multiple values of λ on the internal traing dataset.
At end best value of λ will be used to predict on entire dataset.

-2

-1

-2 compare, best model

-5

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)

In ordinary least square, the sum of squares are reduced to get better MSE. Here shrinkage factor is applied over entire set of parameters hence better modelling approach. (Maybe) The variance can be reduced & better models are obtained

-1

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 longer value of λ → There will be more constrained & more parameter will be ~~reaching~~ closer to zero as compared to when we have less λ or near zero λ.
The increase in λ will make model less flexible
higher value of λ will reduce the MSE making better model But extremly high value of λ will lead to non-flexible model with more bias.

-1

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)	25.293	38.116	0.664	0.507
marital_statusmarried x_1	-19.780	40.405	-0.490	0.624
marital_statusnever_married x_1	-31.760	40.992	-0.775	0.439
age x_2	2.846	1.611	1.767	0.077
I(age^2) x_2^2	-0.024	0.017	-1.470	0.142
marital_statusmarried:age $x_1 \times x_2$	2.024	1.716	1.179	0.238
marital_statusnever_married:age	2.230	1.820	1.225	0.221
marital_statusmarried:I(age^2) $x_1 \times x_2^2$	-0.025	0.018	-1.412	0.158
marital_statusnever_married:I(age^2)	-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)

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_2^2 + \hat{\beta}_4 x_1 * x_2 + \hat{\beta}_5 x_1 * x_2^2 + \dots$

x_1 = marital status x_2 = age

requires 2 dummy variables for marital status

now, x_1 is an categorical so we have dummy for that

married: yes 0, NO 1

never married: 0, 1

& divorce is reference here

-3

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

t statistics → $\frac{\hat{\beta}_0 - 0}{SE}$, Is useful in deterning the intervals & distribution which can be used further for confidence interval.

-1

-4

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)

$\hat{Y} = 25 + (-19.78) \times 1 + 0 + 2.48 \times (30) + (-0.02) \times (30^2 + (2.02) \times 30 \times 1 + (-0.025) \times 1 \times 30^2$

-2.84?

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)

$\hat{Y} = 25 + (-0.02) \times (30^2 + 2.48 \times (30)$

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)

Not much, because the interaction b/w marital status & age are non significant. -3

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)

$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 \cdots = \beta_8 = 0$

$: \beta_1 \neq \beta_2 \neq \cdots \neq \beta_8 \neq 0$

at least 1 is not 0 -1

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)

global is showing that there is some interaction among all the parameters but individual terms like interaction b/w age & marital status or quadratic eqn don't have interaction

model is over parameterized

why? -3

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)

The variance plot we get from plot(model), after plotting the model you will get the plot -2

-9