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)

- Predictive Modeling (4 pts)

- Pattern Finding (4 pts)

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)

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

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

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

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

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)

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

    a. Ordinary least squares performs variable selection.

    b. Ordinary least squares performs shrinkage of coefficient estimates.

    c. Best subset selection performs variable selection.

    d. Best subset selection performs shrinkage of coefficient estimates.

    e. Ridge Regression performs variable selection.

    f. Ridge Regression performs shrinkage of coefficient estimates.

    g. LASSO performs variable selection.

    h. LASSO performs shrinkage of coefficient estimates.

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)

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)

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)

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	-19.780	40.405	-0.490	0.624
marital_statusnever_married	-31.760	40.992	-0.775	0.439
age	2.846	1.611	1.767	0.077
I(age^2)	-0.024	0.017	-1.470	0.142
marital_statusmarried:age	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)	-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)

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

   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)

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)

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)

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)

    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)

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)