IST 5535 Exam 1 Exercise

Notes: This exercise provide some exercise questions for students to prepare for Exam 1. It does not suggest anything that will be actually covered in the real exam.

Part I: Multiple Choice

1. We collect a set of data on the top 500 firms in the US. For each firm we record revenue, cost, number of employees, and the CEO salary. We are interested in understanding which factors affect CEO salary. Which of the following statement is NOT correct?

A) This is a regression problem.
B) Sample size n = 500.
C) Number of predictors p = 4.
D) The purpose of analysis is for statistical inference.

2. Under what circumstance might a more flexible approach be preferred to a less flexible approach?

A) Sample size is small.
B) The number of predictors is small.
C) The goal is for explanation rather than prediction.
D) The Bayes decision boundry is very non-linear.

3. Which of the following statement is NOT correct?

A) Bayes classifier is gold standard for a classification task.
B) Bayes classifier has zero error rate.
C) We can train machine learning algorithms to approximate the Bayes classifier but we could not directly train a Bayes classifier.
D) Bayes classifier has the right level of model complexity.

4. Which of the following statement about simple linear regression is NOT correct?

A) The number of predictors p = 1.
B) Regression R squred equals to the sqaure of correlation coefficient.
C) The regression line always passes through the piont ($\bar{x}, \bar{y}$).
D) The simple linear regression can do statistical control of confounding factors.

5. The relationship between number of beers consumed (x) and blood alcohol content (y) was studied by using least squares regression. The following regression equation was obtained from this study:
y= -0.0127 + 0.0180x
The above equation implies that:

A) Each beer consumed increases blood alcohol by 1.27%.
B) On average it takes 1.8 beers to increase blood alcohol content by 1%.
C) Each beer consumed increases blood alcohol by an average of amount of 1.8%.
D) Each beer consumed increases blood alcohol by exactly 0.018.

6. Which of the following statements about regression analysis is NOT true?

A) Regression is about estimating relationships between dependent and independent variables.
B) Parametric regression models have no assumption regarding the functional form y=m(x) + e
C) Regression intends to summarize observed data as simply and usefully as possible
D) None of the above

7. Assume we are doing a multiple linear regression analysis:
$Y= beta_0 + beta_1X_1 + beta_2X_2 + beta_3X_3$
What can you say about their relationship if $beta_1$ is estimated as 2.0?

A) The relationship between X_1 and Y is significant.
B) The estimated value of Y increases by an average of 2 units for each increase of 1 unit of X_1, holding X_2 and X_3 constant.
C) The estimated value of Y increases by an average of 2 units for each increase of 1 unit of X_1, without regard to X_2 and X_3.
D) The estimated average value of Y is 2 when X_1 equals to zero

8. What is the purpose of using the set.seed() function in the following R section?

```
---------------------------
set.seed (200)
x <- rnorm(100)
eps <- rnorm(100, mean=0, sd=0.5)
y <- -1 + 0.5*x + eps
---------------------------
```

A) To obtain 200 random numbers generated.
B) To set the mean of random numbers as 200.
C) It has no special purpose.
D) To make the result reproducible.

9. Below is a confusion matrix of a classification algorithm:

	Yes	No
Yes	9627	228
No	40	105

Rows are prediction and columns are reference or ground truth. Positive class is Yes. What is the sensitivity of the algorithm?

A) 0.9732
B) 0.3153
C) 0.9959
D) 0.7241

10. Below is a confusion matrix of a null classifier:

	Yes	No
Yes	100	900
No	0	0

Rows are prediction and columns are reference or ground truth. Positive class is No. What is the specificity of the null classifier?

A) 1.00
B) 0.10
C) 0.90
D) 0.50

11. Which of the following method is NOT appropriate to handle imbalanced datasets?

A) Use AUC rather than accuracy to measure performance.
B) Over-sample the majority class.
C) Use different threshold for prediction.
D) Customize the cost function to assign larger penalty to misclassified minority class.

12. Which of the following is a non-parametric method?

A) Linear regression
B) Logistic regression
C) LDA
D) QDA
E) kNN

13. Assume we have y, X1, and X2 in a data frame df, where y is a binary variables containing values of 0 and 1. What is the correct R code to analyze the impact of X1 and X2 on y?

A) model <- lm(y ~ X1 + X2, data=df)
B) model <- glm(y ~ X1 + X2, data=df)
C) model <- glm(y ~ X2 + X1, family=binomial(link='logit'), data=df)
D) None of the above

Part II: Essay

1. Explain the practical definition of machine learning.

2. Explain the difference between KNN classification and KNN regression methods.

3. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

4. Explain the difference among four major data structures in R: vector, matrix, data frame, and list.

5. Explain the trade-off between variance the bias and its implication for machine learning.

6. Explain the meaning of the following R code block.

```
library(ISLR)
library(dplyr)

    Auto %>%
      select(-c(origin, name)) %>%
      slice(-c(10:20)) %>%
      sapply(sd)
```

7. Below table shows the results of regressing sales on two advertising channels including TV and radio.

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

   Explain the synergy effect in the regression result.

8. The dataset Boston has a structure like below:

```
library(MASS)

data(Boston)
str(Boston)
```

```
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ black  : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

(a) Explain the meaning of the following R code block.

```
library(dplyr)

Boston %>%
filter(rm > 7) %>%
nrow
```

(b) Explain the meaning of the following R code block.

```
Boston$class <-  ifelse(Boston$crime >= 0.26, 'Yes', 'No')
```

9. I collect a set of data (n = 100 observations) containing a single predictor and a quantitative response. I then fit a linear regression model to the data, as well as a separate cubic regression, i.e.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

(a) Suppose that the true relationship between X and Y is linear, i.e. $Y = \beta_0 + \beta_1 X + \varepsilon$. Consider the training residual sum of squares (RSS) for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(b) Answer (a) using test rather than training RSS.

(c) Suppose that the true relationship between X and Y is not linear, but we don't know how far it is from linear. Consider the training RSS for the linear regression, and also the training RSS for the cubic regression. Would we expect one to be lower than the other, would we expect them to be the same, or is there not enough information to tell? Justify your answer.

(d) Answer (c) using test rather than training RSS.

10. Below is the result of linear regression on the mtcars dataset.

Multiple Linear Regression

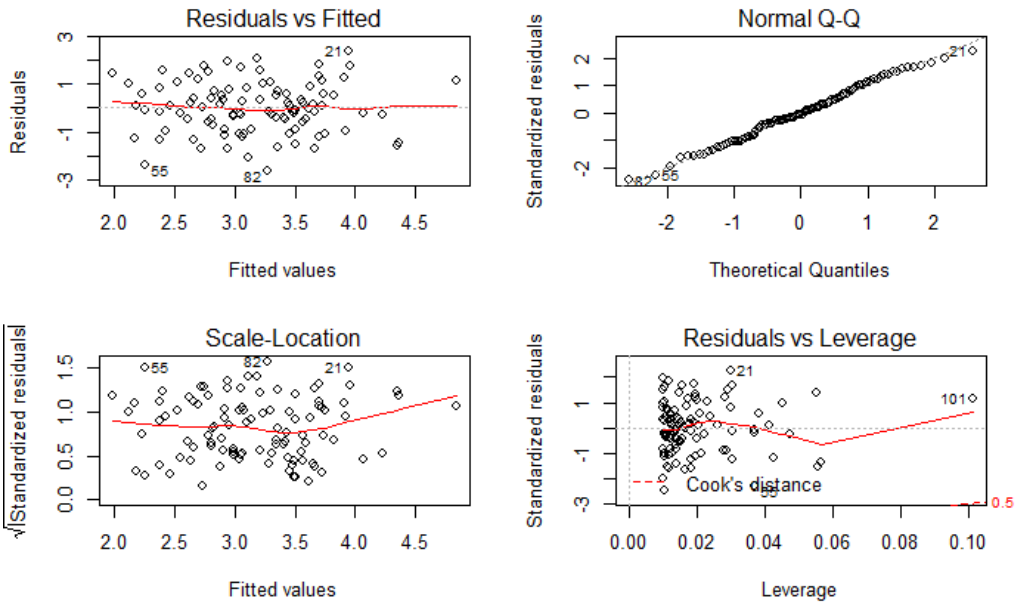
	Dependent variable:		
	mpg		
	(1)	(2)	(3)
wt	-2.8786**	-3.2381**	-3.9165***
	(0.9050)	(0.8899)	(0.7112)
factor(am)1	2.0837	2.9255*	2.9358*
	(1.3764)	(1.3971)	(1.4109)
hp	-0.0375***	-0.0176	
	(0.0096)	(0.0142)	
qsec		0.8106	1.2259***
		(0.4389)	(0.2887)
Constant	34.0029***	17.4402	9.6178
	(2.6427)	(9.3189)	(6.9596)
Observations	32	32	32
R2	0.8399	0.8579	0.8497
Adjusted R2	0.8227	0.8368	0.8336
Residual Std. Error	2.5375 (df = 28)	2.4348 (df = 27)	2.4588 (df = 28)
F Statistic	48.9600*** (df = 3; 28)	40.7354*** (df = 4; 27)	52.7496*** (df = 3; 28)
Note:			*p<0.05; **p<0.01; ***p<0.001

(a) Describe the null hypotheses to which the p-values given in the above table. Explain what conclusions you can draw based on these p-values.

(b) If the purpose is for explanation, which model is more preferred? Explain why.

(c) If the purpose is for prediction, which model is more preferred? Explain why.

11. Below shows diagnostic plots of a multiple linear regression (p = 2, n = 100).



Is there any outliers and high-leverage points? Explain reasons.

12. Explain confounding effect using a practical example. How can we deal with confounding effect in regression analysis?

13. Explain the Bayes optimal classifier.

14. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

15. Formally specify how the logistic regression is modelled. What is the model specification? What is the meaning of a parameter in logistic regression model?

16. This problem relates to the QDA model, in which the observations within each class are drawn from a normal distribution with a class-specific mean vector and a class specific covariance matrix. We consider the simple case where p = 1; i.e. there is only one feature.
    Suppose that we have K classes, and that if an observation belongs to the kth class then X comes from a one-dimensional normal distribution, $X \sim N(\mu_K, \sigma_K^2)$. Recall that the density function for the one-dimensional normal distribution is given in the following equation.

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_K} exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_K)^2\right)$$

    Prove that in this case, the Bayes' classifier is not linear. Argue that it is in fact quadratic.

17. Suppose we collect data for a group of students in a machine learning class with variables X_1 =hours studied, X_2 =undergrad GPA, and Y = receive an A. We fit a logistic regression and produce estimated coefficient, $\hat{\beta}_0 = -5$, $\hat{\beta}_1 = 0.04$, $\hat{\beta}_2 = 1$.

    (a) Estimate the probability that a student who studies for 40 hours and has an undergrad GPA of 3.5 gets an A in the class.

    (b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?

18. Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next we use 1-nearest neighbors (i.e. K = 1) and get an average error rate (averaged over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

19. On average, what fraction of people with an odds of 0.37 of defaulting on their credit card payment will in fact default? Suppose that an individual has a 16% chance of defaulting on her credit card payment, what are the odds that she will default?

20. Suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube centered around the test observation that contains, on average, 20% of the training observations. For p = 1, 2, 4, and 8, what is the length of each side of the hypercube? What conclusions you can get from the example.

    Note: A hypercube is a generalization of a cube to an arbitrary number of dimensions. When p = 1, a hypercube is simply a line segment, when p = 2 it is a square, and when p = 8 it is an 8-dimensional cube.

21. The table below provides a training data set containing six observations, three predictors, and one quantitative response variable.

Observation	X_1	X_2	X_3	Y
1	0	3	0	5
2	2	0	0	4.5
3	0	1	3	3
4	0	1	2	3.5
5	-1	0	1	4
6	1	1	1	5.5

    Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

(b) What is our prediction with K = 1? Why?

(c) What is our prediction with K = 3? Why?

(d) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small? Why?

22. The table below provides a confusion matrix. Calculate overall accuracy, sensitivity, specificity, false positive rate, and false negative rate. Assume class Yes is the positive class.

		Truth	
		No	Yes
Prediction	No	9,300	200
	Yes	150	100

23. What is an imbalanced dataset? What problems does an imbalanced dataset have for machine learning? What methods can be used to deal with an imbalanced dataset?

24. Explain the reasons why we should not include all predictors in a regression model.

25. Provide an example to explain multicollinearity in regression analysis. What is the problem for regression analysis? How the multicollinearity can be detected? How the multicollinearity can be handled in regression analysis?

26. Write down the result of the following R chunk.

```
sum <- 0
 i <- 0
 while(i < 20){
    i <- i +1
    if (sum >= 100)
       break
    sum <- sum + i
 }
 cat("The i is",i,"\n")
 cat("The sum is",sum,"\n")
```