**Department of Statistics**

**Quaid-i-Azam University, Islamabad**

1. Which of the following are supervised learning problems? More than one box can be checked.

1. Predict whether a website user will click on an ad
2. Find clusters of genes that interact with each other (FALSE)
3. Classify a handwritten digit as 0-9 from labeled examples
4. Find stocks that are likely to rise

2. True or False: The only goal of any supervised learning study is to be able to predict the response very accurately.

1. TRUE b) FALSE

3. True or False: A fitted model with more predictors will necessarily have a lower Training Set Error than a model with fewer predictors.

a) TRUE b) FALSE

4. While doing a homework assignment, you fit a Linear Model to your data set. You are thinking about changing the Linear Model to a Quadratic one. Which of the following is most likely true:

a) Using the Quadratic Model will decrease your Irreducible Error.

b) Using the Quadratic Model will decrease the Bias of your model.

c) Using the Quadratic Model will decrease the Variance of your model

d) Using the Quadratic Model will decrease your Reducible Error

5. You are doing an analysis in R and need to use the 'summary()' function, but you are not exactly sure how it works. Which of the following commands should you run? (There is more than one correct answer, so any one these will earn the point).

1. help(summary) b) ?summary c)main(summary) d) ?summary()

6. The sample size n is extremely large, and the number of predictors p is small:

a) Flexible is better b) Flexible is worse

7. The number of predictors p is extremely large, and the sample size n is small:

a) Flexible is better b) Flexible is worse

8. The variance of the error terms, i.e. σ2=Var(ϵ), is extremely high:

a) Flexible is better b) Flexible is worse

9. The relationship between the predictors and response is highly non-linear:

a) Flexible is worse b) Flexible is better

10. Why is linear regression important to understand? Select all that apply:

1. The linear model is often correct
2. Linear regression is very extensible and can be used to capture nonlinear effects
3. Simple methods can outperform more complex ones if the data are noisy
4. Understanding simpler methods sheds light on more complex ones

11. Which of the following are true statements? Select all that apply:

1. A 95% confidence interval is a random interval that contains the true parameter 95% of the time
2. The true parameter is a random value that has 95% chance of falling in the 95% confidence interval I perform a linear regression and get a 95% confidence interval from 0.4 to 0.5.
3. There is a 95% probability that the true parameter is between 0.4 and 0.5.
4. The true parameter (unknown to me) is 0.5. If I sample data and construct a 95% confidence interval, the interval will contain 0.5 95% of the time.

12. We run a linear regression and the slope estimate is 0.5 with estimated standard error of 0.2. What is the largest value of b for which we would NOT reject the null hypothesis that β1=b? (assume normal approximation to t distribution, and that we are using the 5% significance level for a two-sided test; need two significant digits of accuracy)

Answer :…………

13. Which of the following indicates a fairly strong relationship between X and Y?

1. R2=0.9 b) The p-value for the null hypothesis β1=0 is 0.0001 c)The t-statistic for the null hypothesis β1=0 is 30

14. Suppose we are interested in learning about a relationship between X1 and Y, which we would ideally like to interpret as causal.

True or False? The estimate β^1 in a linear regression that controls for many variables (that is, a regression with many predictors in addition to X1) is usually a more reliable measure of a causal relationship than β^1 from a univariate regression on X1.

1. True b) False

15. Judging from the plots on page 2 of the notes, which should be the better predictor of Default: Income or Balance?

1. Income b) Balance c) Both are equally good d)Not enough information is given to decide.

16. Suppose we collect data for a group of students in a statistics class with variables X1= hours studied, X2=undergrad GPA, and Y= receive an A. We fit a logistic regression and produce estimated coefficients β^o=−6,β^1=0.05,β^2=1.

Estimate the probability that a student who studies for 40h and has an undergrad GPA of 3.5 gets an A in the class (within 0.01 accuracy): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

17. How many hours would that student need to study to have a 50% chance of getting an A in the class?

18. Suppose that in Ad Clicks (a problem where you try to model if a user will click on a particular ad) it is well known that the majority of the time an ad is shown it will not be clicked. What is another way of saying that?

1. Ad Clicks have a low Prior Probability b) Ad Clicks have a high Prior Probability
2. c)Ad Clicks have a low Density d)Ad Clicks have a high Density.

19. Which of the following statements best explains the relationship between Quadratic Discriminant Analysis and naive Bayes with Gaussian distributions in each class?

1. Quadratic Discriminant Analysis is a more flexible class of models than naive Bayes
2. Quadratic Discriminant Analysis is a less flexible class of models than naive Bayes
3. Quadratic Discriminant Analysis is an equivalently flexible class of models to naive Bayes
4. For some problems Quadratic Discriminant Analysis is more flexible than naive Bayes, for others the opposite is true.

20. In ch4.R, line 13 is "attach(Smarket)." If that line was omitted from the script, which of the following lines would cause an error?:

1. line 15: mean(glm.pred==Direction)
2. line 18: glm.fit=glm(Direction~Lag1+Lag2+Lag3+Lag4+Lag5+Volume, data=Smarket, family=binomial, subset=train)
3. line 22: Direction.2005=Smarket$Direction[!train]
4. line 30: table(glm.pred,Direction.2005)