Test 2

DSO 530: Applied Modern Statistical Learning Methods

April, 29th, 2020

You have 120 minutes to do the problems. For multiple choice questions (1-25), make sure to read the questions very carefully and write down the best answer. If you write down multiple answers for a question, you will receive zero for that question. For short answer questions (26-31), **write concisely, and clearly**. This test is open notes. You can read the textbook, class slides, notes, python tutorials on your computer, but you should **not** search on-line, open jupiter notebooks or use Python, or watch class recordings. Questions 1-2, 25-31 are worth 2 points each, and the rest are worth 1 point each. The total points are 40. **For both multiple choice and short answer questions, you should clearly indicate the correspondence between the question number and your answer**. All answers have to be **hand-written**. Do **not** upload your scratch paper.

If you have multiple pages in your answers, **number** these pages. Write down your **name (Last, First)** and **USC Student ID number** on top of every page of your answers.

submission instructions: **Scan** your answers into **a single pdf** document. **Name** this document by firstname_lastname_uscIDnumber_test2.pdf. **Email** this pdf file to your own usc.edu email and to dso530.2020spring@gmail.com. Then, **upload** this pdf file to https://www.dropbox.com/request/JEMsyBGrdmOiZYGgIHie. Finally, **send a public message on Zoom to sign off**. For example: Alex James signs off at 9:10pm.

additional instructions: Do not redistribute this test. **Delete** the test after you submit your answers. Also, **do not discuss or share** your answers after the test.

part a) multiple choices

1.

```
import numpy as np
np.random.seed(2);x = np.random.standard_normal(50)
## simulate 50 numbers from standard normal distribution
np.random.seed(2); y  = -0.5 * np.random.standard_normal(50)
```

Which is true about $x + 2 * y$?

- A) Each entry of $x + 2 * y$ is 0.
- B) The entries of $x + 2 * y$ cannot be determined given the above information. Likely, they follow some normal distribution with standard deviation 1.

2. Which of the following is/are true about PCA?

i) We should always perform PCA first when we do data analysis

ii) The number of PCs we can compute cannot be bigger than the number of features

iii) The first PC is the direction along which data vary the least.

- A) i), iii)
- B) ii), iii)

- C) i) ii)
- D) ii)
- E) iii)

3. Which of the following is/are correct?

i) When we use the least squares approach to fit a linear regression, we do not have to standardize the predictors.

ii) When we use the LASSO approach to fit a linear regression, we should standardize the predictors.

iii) LASSO and ridge regression can be applied to situations where the number of features is bigger than the sample size.

- A) i), ii), iii)
- B) i) ii)
- C) ii)

4. Which of the following is/are TRUE about R^2?

i) R^2 on training data can be used to compare models of different sizes (i.e., different number of predictors).

ii) For classification problems, R^2 on test data is usually used as a metric to evaluate an algorithm's performance.

iii) R^2 is a robust measure.

- A) i)
- B) ii)
- C) iii)
- D) i), ii)
- E) ii), iii)
- F) i), ii), iii)
- G) None

5. Which of the following are/is true?

i) When the sample size is larger than the number of features, forward stepwise selection searches through more models compared to backward stepwise selection.

ii) Backward stepwise regression cannot be used when the number of features is larger than the sample size.

- A) i)
- B) ii)
- C) i), ii)

- D) None

6. A and B are two events. Given that $P(A) = 0.8$ and $P(B) = 0.7$, what is $P(A \text{ and } B)$?

- A) 0.8
- B) 0.7
- C) 0.56
- D) It cannot be determined without further information

7. Which of the following is/are true?

i) In k-means algorithm, we do not need to specify the number of clusters.

ii) Support vector machine is capable of creating non-linear decision boundaries

iii) Decision trees are usually not as good as random forest for prediction

- A) i)
- B) ii)
- C) iii)
- D) i), ii)
- E) i), iii)
- F) ii), iii)
- G) i), ii), iii)

8. Which of the following are/is NOT correct about LASSO for linear regression.

i) When the tuning parameter $\lambda = 0$, LASSO reduces to the usual least squares approach

ii) As λ increases, the absolute value of each coefficient estimate monotone decreases. ("monotone" means in a single direction)

iii) LASSO encourages sparse models for proper choices of λ.

- A) i), ii), iii)
- B) i), iii)
- C) i), ii)
- D) i)
- E) ii)
- F) iii)

9. Which of the following is/are correct about AIC, BIC, C_p, and adjusted R^2 for model selection in linear regression? (Suppose each criterion chooses one model)

- A) These four criteria will give us 4 different models.
- B) These four criteria will give us at most 3 different models.

10. How many of the following classification methods are based on explicit probabilistic model assumptions?

i) logistic regression

ii) LDA

iiii) classification trees

iv) random forest

v) support vector machines

- A) Less than 2.
- B) 2
- C) 3
- D) 4
- E) 5

11. A recent survey conducted by the personnel manager of a major enterprise resources planning (ERP) company showed that 35% of the employees were dissatisfied with their salary, 85% were satisfied with their work assignments, 15% were dissatisfied with their work hours, 17% were dissatisfied with both their salary and work assignments, and 8% were dissatisfied with both their work assignments and work hours. What is the percentage of employees who are satisfied with both their salary and work assignments?

- A) 0.38
- B) 0.02
- C) 0.62
- D) The correct percentage is between 0.2 and 0.9, but it does not equal to A), B), or C)
- E) The numbers in the survey do not add up, and there must be something wrong with the summary statistics.

12. The usual correlation we learned in class is called Pearson's correlation, and it measures the linear dependence between two numerical variables. Rank correlation coefficients, such as Spearman's rank correlation, measure the extent to which, as one variable increases, the other variable tends to increase, without requiring that increase to be represented by a linear relationship. For example, if as one variable increases, the other decreases, then their Spearman's rank correlation will be negative. Given this description, based on four pairs of $x - y$ observations: $(0, 1)$, $(1, 3)$, $(2, 7)$, $(3, 100)$, what is the sign of Spearman's rank correlation for x and y?

- A) positive
- B) negative

13. Which of the following is/are true about regression trees?

i) It is a supervised learning technique

ii) It is an unsupervised learning technique

iii) We usually use Gini index or cross-entropy to grow these trees

iv) For all purposes, regression trees are inferior to random forest

- A) i)
- B) ii)
- C) i), iii)
- D) ii), iii)
- E) i), iii), iv)
- F) ii), iii), iv)

14. We generate a bootstrap sample (sample size 4) from a set of 4 observations $\{a_1, a_2, a_3, a_4\}$. What is the probability that the first observation a_1 is NOT in the bootstrap sample?

- A) 20%
- B) 75%
- C) 25%
- D) 51.2%
- E) 31.6%

15. The statement "The predictors in the k-variable model identified by best subset selection are a subset of the predictors in the (k+1)-variable model identified by best subset selection" is

- A) True
- B) False, because the predictors in k-variable model are never a subset of the predictors in the (k+1)-variable model.
- C) False, because, in some situations, the predictors in the k-variable model are not a subset of the predictors in the (k+1)-variable model.

16. Let X be a feature and $Y \in \{0, 1\}$ denote class label. Suppose the distribution of X given $Y = 0$ is Normal with mean 0 and standard deviation 1 (i.e., $N(0, 1)$), and the distribution of X given $Y = 1$ is Normal with mean 2 and standard deviation 1 (i.e., $N(2, 1)$). What is the Bayes classifier?

- A) The Bayes classifier is to assign X to class 1 if $X > 1$, and to class 0 otherwise.
- B). We need more information to tell the Bayes classifier.

Problems **17-20** are based on the dataset *caravan*. This data set includes 85 predictors that measure demographic characteristics for 5,822 individuals. The response variable is Purchase, which indicates whether or not a given individual purchases a caravan insurance policy.

```
import pandas as pd
import numpy as np
caravan = pd.read_csv('caravan.csv')
caravan.head(); caravan.shape; caravan.isnull().head()
```

```
##    MOSTYPE  MAANTHUI  MGEMOMV  MGEMLEEF  ...  AFIETS  AINBOED  ABYSTAND  Purchase
## 0       33         1        3         2  ...       0        0         0        No
## 1       37         1        2         2  ...       0        0         0        No
## 2       37         1        2         2  ...       0        0         0        No
## 3        9         1        3         3  ...       0        0         0        No
## 4       40         1        4         2  ...       0        0         0        No
##
## [5 rows x 86 columns]
## (5822, 86)
##    MOSTYPE  MAANTHUI  MGEMOMV  MGEMLEEF  ...  AFIETS  AINBOED  ABYSTAND  Purchase
## 0    False     False    False     False  ...   False    False     False     False
## 1    False     False    False     False  ...   False    False     False     False
## 2    False     False    False     False  ...   False    False     False     False
## 3    False     False    False     False  ...   False    False     False     False
## 4    False     False    False     False  ...   False    False     False     False
##
## [5 rows x 86 columns]
```

```
caravan.isnull().sum(); caravan.isnull().sum().sum()
```

```
## MOSTYPE     0
## MAANTHUI    0
## MGEMOMV     0
## MGEMLEEF    0
## MOSHOOFD    0
##             ..
## APLEZIER    0
## AFIETS      0
## AINBOED     0
## ABYSTAND    0
## Purchase    0
## Length: 86, dtype: int64
## 0
```

17. How many variables in the dataset have missing entries?

- A) 0
- B) 1
- C) ≥ 2

18.

```
from sklearn.model_selection import train_test_split
X = caravan.drop(['Purchase'], axis=1).astype('float64')
```

```
y = caravan['Purchase']
X_train, X_test , y_train, y_test = train_test_split(X, y, test_size=0.5, random_state=1)
```

The above code splits the *caravan* data into training and test sets. Which of the following is correct?

- A) The *caravan* data were split randomly into two halves as training and test sets.
- B) The first half number of rows of *caravan* is the training set and the latter half is the test set.

19.

```
from sklearn.linear_model import LogisticRegression
fit1 = LogisticRegression(random_state=1, penalty='none', max_iter=4000).fit(X_train, y_train)
fit1.score(X_test, y_test)
```

```
## 0.9350738577808313
```

Based on the output, what is the test error of the trained logistic regression classifier?

- A) 93.5%
- B) 6.5%
- C) 0
- D) 1

20. Now fit an $l1$-penalized version of the logistic regression, where the penalty level is chosen by cross-validation.

```
from sklearn.linear_model import LogisticRegressionCV
fit2 = LogisticRegressionCV(Cs=20,random_state=1, penalty='l1',solver='liblinear',cv=5)
fit2.fit(X_train, y_train)
```

```
## LogisticRegressionCV(Cs=20, class_weight=None, cv=5, dual=False,
##                      fit_intercept=True, intercept_scaling=1.0, l1_ratios=None,
##                      max_iter=100, multi_class='auto', n_jobs=None,
##                      penalty='l1', random_state=1, refit=True, scoring=None,
##                      solver='liblinear', tol=0.0001, verbose=0)
```

```
fit2.score(X_test, y_test)
```

```
## 0.9395396770869117
```

Does the $l1$-penalized logistic regression achieve an accuracy on the test data higher than that achieved by the regular (i.e., no penalty) logistic regression?

- A) Yes
- B) No

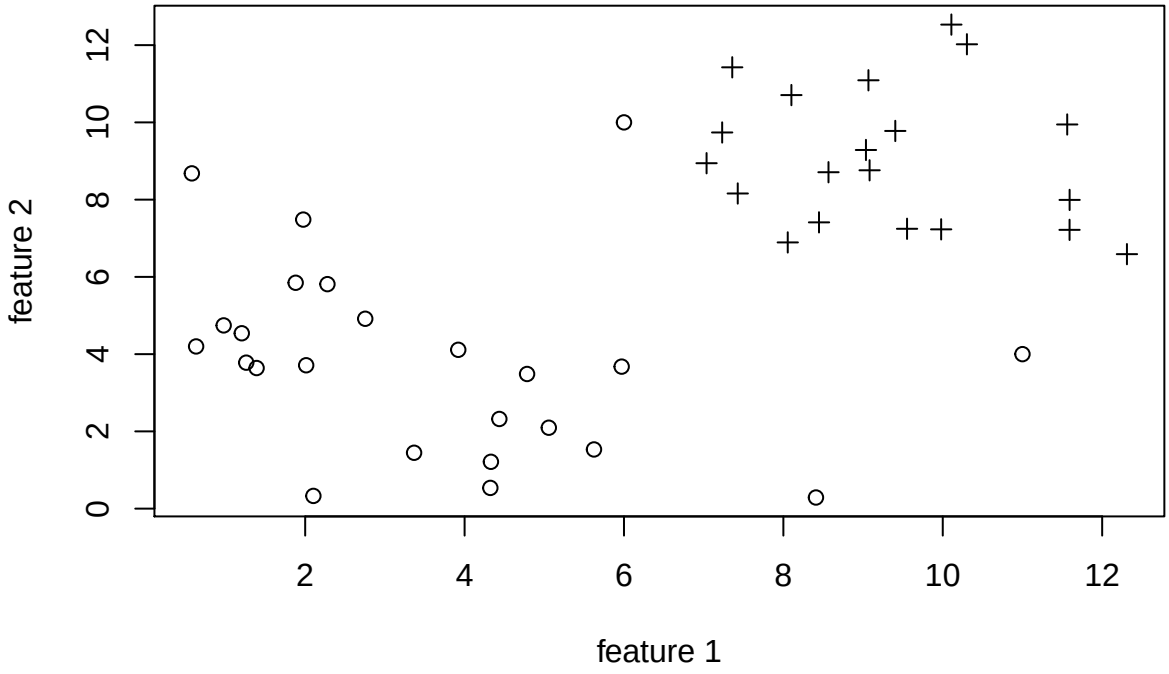
21. Which of the following statements is correct?

- A) We should always prefer more complex models.
- B) In email spam detection, we care prediction more than inference.
- C) In all real applications, the number of available features is smaller than the sample size.

22. Both random forest and boosting are ensemble methods that build a lot of trees. Which of the following is correct.

- A) Random forest builds trees sequentially.
- B) Boosting builds trees sequentially.

23. Given a dataset with a two-dimensional feature space and binary response (labeled as cross and circle), as shown in the plot below. Recall the support vector classifiers that have a tuning parameter C which bounds the sum of slack variables. Based on the plot, if we were to train a support vector classifier based on these observations, which of the following about C is correct?

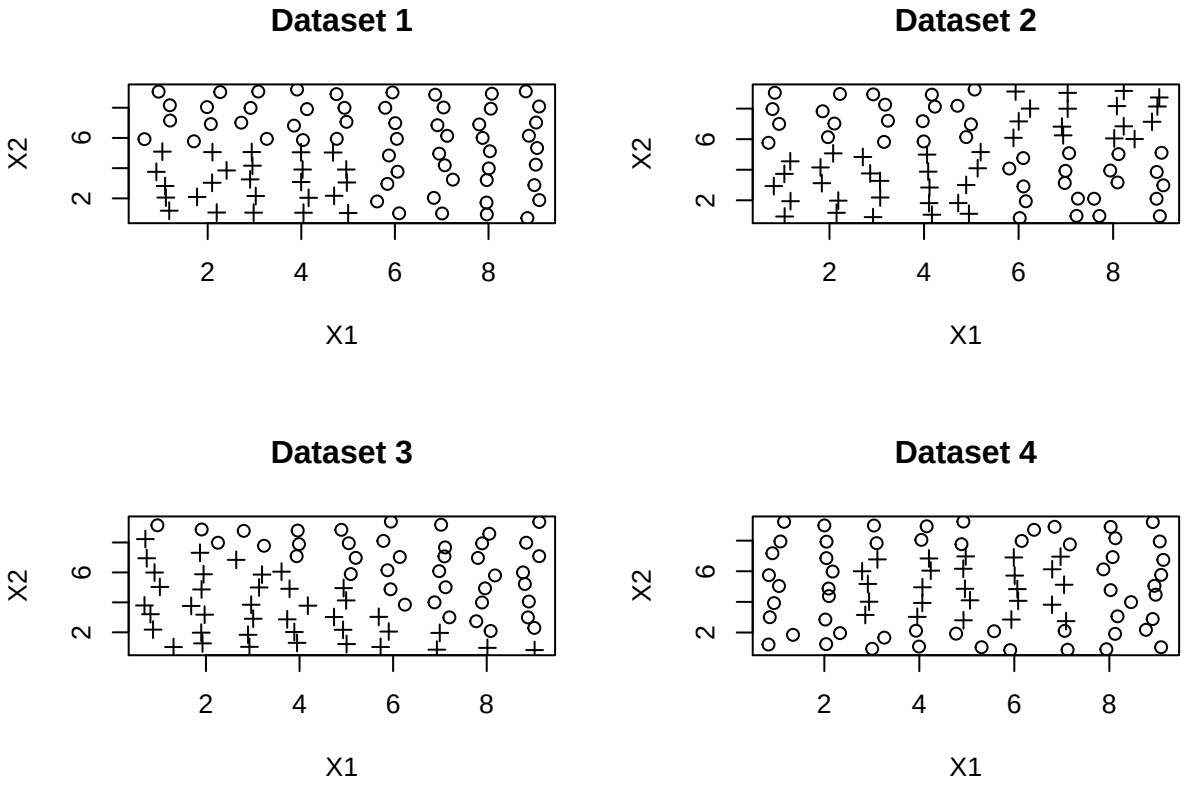


- A) We should set C to be 0
- B) We should set C to be positive

24. If a decision tree (as defined as in our class) partitions the feature space into regions $R_1 \cdots, R_J$, can any of these regions overlap?

- A) Yes.
- B) No.

25. Suppose we have continuous features X_1 and X_2 and collected labels (cross and circle) for four datasets. We want to use decision trees for classification on these datasets. At each split, the criteria for splitting is of the form $X_j > c$, where X_j is a feature to be picked up and c is a threshold to be trained. At each split, the decision should be made based on only one feature, but each feature could be used multiple times at different splits in a tree. If we only grow one tree on every dataset, which of the following dataset(s) can be perfectly classified with a tree whose depth is less than 3. (Tree depth means the maximum number of splits one instance possibly needs to go through to be classified)



- A) Datasets 1, 2, 4
- B) Datasets 1, 2, 3
- C) Datasets 2, 3, 4
- D) Datasets 1, 2
- E) Datasets 2, 4

part b) Questions **26-31** are **short answer** questions.

26. In a dataset with sample size $n = 3$ and feature dimensionality $p = 2$, the three instances are $x_1 = (0, 0)$, $x_2 = (1, 1.1)$ and $x_3 = (2, 2)$. How many PCs can we get from this dataset? Without computing the PCs, do you expect that the PVE of the first PC is larger than 50% and why?

27. Suppose you have got the regression equation $\log y = 1 + 3 \log x$. Based on this equation, how would changes in y associate with changes in x? (You need to show the mathematical derivation to get your answer.)

28. Your team calculated R^2 (computed on the training set, using the least squares approach of linear regression) to be 0.1 for a certain financial data set whose training sample size is $n = 201$. Suppose your model has 20 predictors, what is the adjusted R^2? (You should show the steps to get your answer)

29. What is the difference between bagging and random forest algorithms?

30. In addition to the randomness due to sampling, we have two other sources of randomness in the random forest algorithm. What are these sources?

31. A data set contains four observations: $\{x_1, x_2, x_3, x_4\}$, where $x_1 = (0, 0)$, $x_2 = (1, 1)$, $x_3 = (100, 100)$, and $x_4 = (101, 101)$. We apply K-means algorithm to this dataset with $K = 2$. Suppose the initial group assignment is $\{x_1, x_3\}$ in one group and $\{x_2, x_4\}$ in another. Compute the centroid in each group after the initial assignment. (You need to show the step to get your answer)