Test 2

DSO 530: Applied Modern Statistical Learning Methods

2022

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 class slides, notes, python tutorials on your computer or in print, but you should **not** search on-line, open jupyter 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 lastname_firstname_uscIDnumber_test2.pdf. Then, **upload** this pdf file to Blackboard (like you did for test1). Finally, **send a public message on Zoom to sign off**. For example: Alex James signs off at 8:35 pm. Note that after you sign off, any submission to Blackboard will be considered as improper conduct.

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

part a) multiple choices

1. In a binary classification problem, you apply logistic regression and calculate the type I error = 0.2 and type II error = 0.5. Moreover, you know that the sample size of class 0 observations is 100 and the sample size of class 1 observations is 50. What is the precision? (Hint: you can check out Tutorial 7 for related definitions)

- A) 2/5
- B) 1/2
- C) 1/5
- D) 5/9
- E) It cannot be calculated without additional information.

2. After implementing the code

```
import numpy as np
np.random.seed(2)
x = np.random.standard_normal(100)
np.random.seed(2)
y = -3 * np.random.standard_normal(100)
```

What is the absolute value of the correlation between x and y?

- A) 1
- B) -1
- C) 3
- D) -3
- E) It cannot be determined given the above information.

3. Which of the following is/are correct?

i) The ridge regression aims to minimize $RSS + \lambda \sum_{j=0}^{p} \beta_j^2$.

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

- A) i)
- B) ii)
- C) i), ii)
- D) None

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

i) Out-of-sample R^2 and adjusted R^2 are the same concept.

ii) For classification problems, out-of-sample R^2 is a commonly-used metric to evaluate an algorithm's performance.

iii) Both in-sample R^2 and out-of-sample R^2 are smaller than or equal to 1.

- 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 n is larger than the number of features p, forward stepwise selection searches through more models compared to backward stepwise selection.

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

iii) When $p = 5$ and $n = 100$, let N_b =the number of models best subset selection searches through, and N_f =the number of models searched by forward stepwise selection. Then $N_b > 3N_f$.

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

6. Which of the following is/are true?

i) For all sample sizes, the maximum number of principal components (PCs) of a data set equals the number of features.

ii) Support vector machine can create 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)

7. Which of the following is/are true about k-means clustering?

i) k-means clustering is a supervised learning technique.

ii) To use k-mean clustering, you need to specify the number of clusters.

iii) k-means clustering is the same as the k-nearest neighbors.

- 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 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 tends to perform better than ridge when the true model is sparse.

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

9. Which of the following denotes a hyperplane in the p-dimensional space?

- A) $\beta_0 + \beta_1 X_1 \cdots + \beta_p X_p = 0$
- B) $\beta_0 + \beta_1 X_1^2 \cdots + \beta_p X_p^2 = 0$

10. Which of the following is/are true about random forests?

i) Random forests are bagged decision tree models that split on the first m feature on each split.

ii) Bagging is a special case of random forests.

iii) Random forests can be used for both regression and classification.

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

11. Which of the following is/are the source(s) of randomness in random forests?

i) randomness due to sampling.

ii) randomness due to creating bootstrap samples.

iii) randomness due to random selection of m out of p features in each split.

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

12. Which of the following is/are true about the maximal margin classifier?

i) Maximal margin classifier classifies a test observation based on which side of the maximal margin hyperplane it lies.

ii) Here, *margin* refers to the maximal distance from the training observations to the hyperplane.

iii) Maximal margin classifier has non-linear decision boundaries.

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

13. Which of the following statements are/is true?

i) Validation set approach is NOT a special case of cross validation

ii) Leave-one-out-cross-validation (LOOCV) is a special case of 10-fold CV.

iii) Both 5-fold CV and 10-fold CV are commonly used.

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

14. We generate a bootstrap sample (sample size 5) from a set of 5 observations $\{a_1, a_2, a_3, a_4, a_5\}$. What is the probability that neither a_1 nor a_2 is in this bootstrap sample?

- A) 60%
- B) 40%
- C) 21.6%
- D) 20.5%
- E) 7.8%

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

- A) True
- B) False

16. Which of the following is/are true?

i) The Bayes classifier is the best classifier no matter what evaluation metric we consider.

ii) Both boosting and random forests use multiple trees for prediction

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

Problems **17-19** 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.

```
from sklearn.model_selection import train_test_split
X = caravan.drop(['Purchase'], axis=1)
y = caravan['Purchase']
X_train, X_test , y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0, stratify = y)
```

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

- A) The test set consists of 30% of the data
- B) The test set consists of 70% of the data

18.

```
from sklearn.preprocessing import MinMaxScaler
#### i
mms=MinMaxScaler(); X_train_norm=mms.fit_transform(X_train)
X_test_norm=mms.transform(X_test)
```

```
#### ii
mms=MinMaxScaler(); X_train_norm=mms.fit_transform(X_train)
X_test_norm=mms.fit_transform(X_test)
```

If we want to normalize the features, which is recommended?

- A) i
- B) ii

19. On this dataset, suppose you want to do some feature selection for logistic regression. Among best subset selection and forward stepwise selection, which one do you recommend?

- A) best subset selection
- B) forward stepwise selection

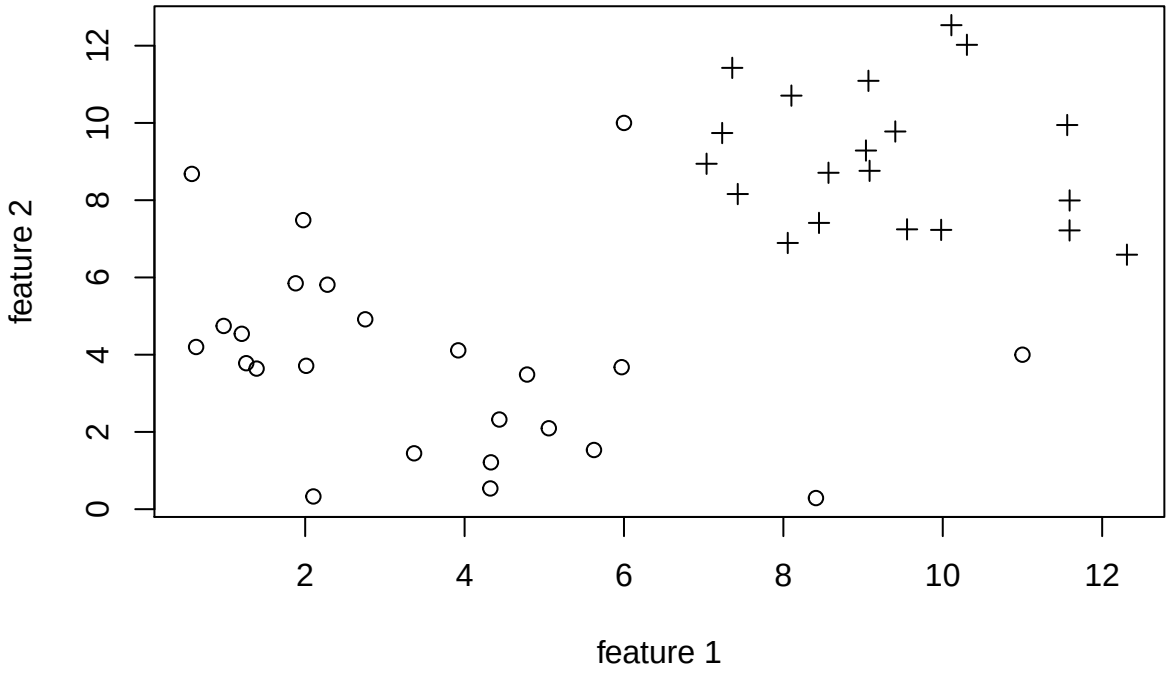
20. A recent survey conducted by the personnel manager of a major enterprise resources planning (ERP) company showed that 35% of the employees were dissatisfied with their salary, 85% were satisfied with their work assignments, 15% were dissatisfied with their work hours, 17% were dissatisfied with both 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 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.

21. "In binary classification, classification error equals the sum of type I error and type II error." The statement is

- A) True
- B) False

22. Given a dataset with two dimensional feature space and two labels (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. If we were to build a support vector classifier based on these observations, which of the following about C is correct?

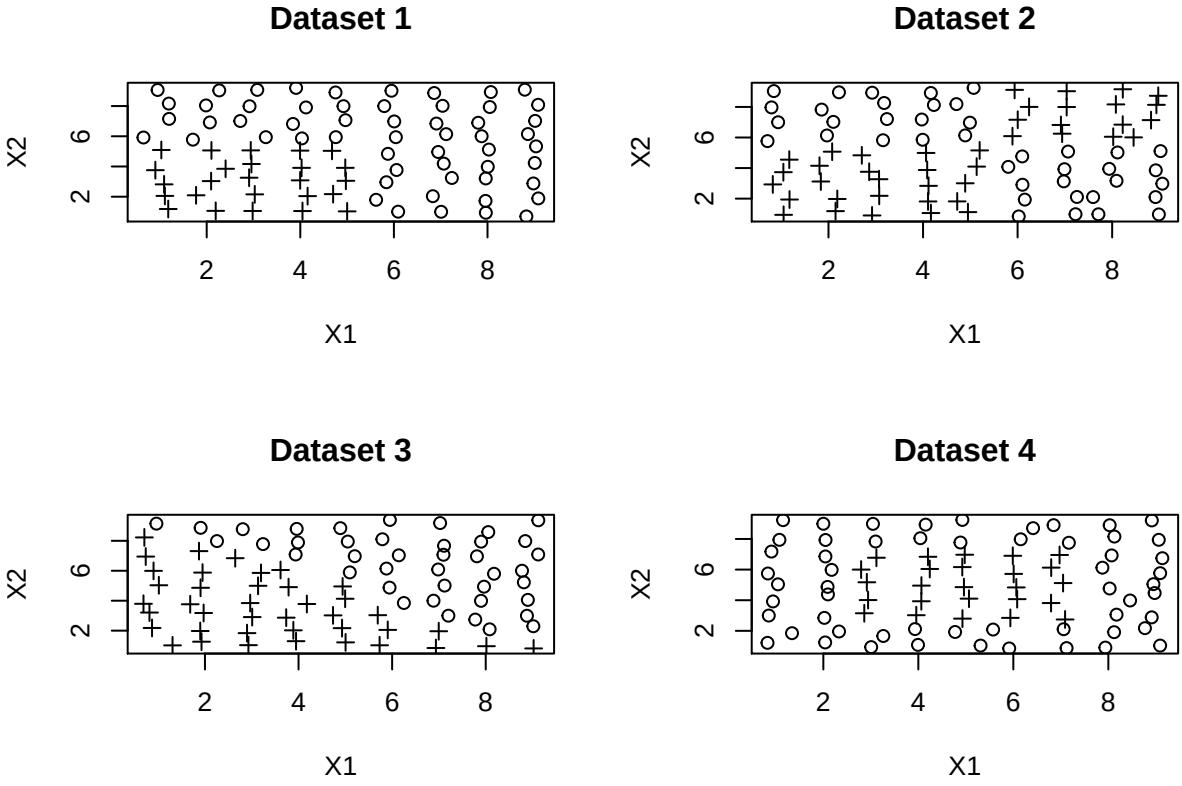


- A) We should set C to be 0
- B) We should set C to be positive
- C) We cannot make a decision based on the plot

23. You have applied the least squares approach to linear regression to a data set that has 3 predictors, each of which you know is important for prediction. The sample size of your data set is $n = 50{,}000$. Now your boss just learned a fancy technique LASSO, and would like you to redo the work with this newer technique. Will you follow his advice?

- A) Yes, because fancy techniques are always better than the simple ones.
- B) No, and I would argue that it is not necessary to use this more advanced technique in this example.

24. 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 number of terminal nodes is smaller than or equal to 4.



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

25. "We can construct a decision tree where the number of terminal nodes is smaller than the number of internal nodes." This statement is

- A) True.
- B) False.

part b) Questions **26-31** are **short answer** questions. Answer questions clearly and concisely. Writing irrelevant or convoluted responses could result in deduction in points.

26. 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, as the one variable increases, the other decreases, Spearman's rank correlation will be negative. Given this description, based on four pairs of $x - y$ observations: $(0, 1)$, $(1, 3)$, $(2, 7)$, $(3, 15)$, what will be the signs of Spearman's rank correlation and Pearson's correlation for x and y respectively? (You just need to state "positive" or "negative" without reasoning.)

27. Suppose you have got the regression equation $\log y = 3 + 90 \log x$. 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. In a regression problem, the number of total features $p = 20$, and the sample size $n = 500$. (i) If you do backward stepwise selection for this problem, how many models will be investigated? (ii) If you do best subset selection for this problem, how many models will be investigated? (You only need to write down two numbers, and do not need to write down the derivations)

29. Write down the AIC and C_p criteria for linear regression. Explain why they will choose the same model.

30. A dataset contains five points: $\{x_1, x_2, x_3, x_4, x_5\}$, where $x_1 = (0, 0)$, $x_2 = (1, 1)$, $x_3 = (100, 100)$, $x_4 = (101, 101)$, $x_5 = (123, 123)$. We wish to apply the K-means algorithm to this dataset with $K = 2$. Suppose the initial group assignment is $\{x_1, x_3\}$ in one group and $\{x_2, x_4, x_5\}$ in another. Compute the centroid in each group after the initial assignment. (You need to show the centroids as well as how you get those centroids.)

31. When we built a decision tree for 4-class classification, we found out that on the 1st terminal node (or region R_1 if you think in terms of the feature space partition), there are five class 1 instances, five class 2 instances, ten class 3 instances, and ten class 4 instances. What is the Gini index for this node? (You need to show the steps to get the answer)