Test 1

DSO 530: Applied Modern Statistical Learning Methods

2022

You have 90 minutes to do the problems. For multiple choice questions (1-16), 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 (17-22), **write concisely, and clearly**. This test is open notes. You can read the class slides, notes, and python tutorials as printouts or on your computer, but you should **not** search on-line, open jupiter notebooks or use Python, or watch class recordings. Questions 1, 17-22 are worth 2 points each, and the rest are worth 1 point each. The total points are 29. Do not write your answers on this test paper. Instead, write your answers on some **blank papers**. **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_test1.pdf. Then, **upload** this pdf file to Blackboard (like you did for HW1 and Quiz1). Finally, **send a public message on Zoom to sign off**. For example: Alex James signs off at 8:15 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. Let Y be the child's height, and X_1 and X_2 be the heights of both parents. Based on a training set, we used some fitting method to get the linear regression equation $Y = 0.5X_1 + 0.5X_2$. Now we have a test dataset of five people. The heights of these five people, along with the heights of their respective parents, are given in the table below. What is the out-of-sample R^2?

Table 1: Table for five people's heights(Y) and their parents heights (X_1 and X_2). All units are centimeters

Y	X_1	X_2
168	160	174
181	180	178
172	172	170
170	168	180
174	175	170

- A) 100%
- B) 90.3%

- C) 80.6%
- D) 75.8%
- E) 55.9%

2. How many of the following statements is/are correct?

i) Let A and B be two disjoint events. $P(A) + P(B)$ could be more than 1.

ii) The correlation r takes values between -1 and 1.

iii) In statistical hypothesis test, we reject the null hypothesis if p-value is larger than α.

iv) For simple linear regression, the relation $R^2 = r^2$ holds for out-of-sample R^2 (Here r is the correlation between the response and the input variable).

Choose one of the following:

- A) None
- B) One
- C) Two
- D) Three
- E) Four

Questions $3 - 8$ are based the `housing` dataset. As we have seen this dataset multiple times in lectures and tutorials, we will skip the description.

```
import numpy as np
import pandas as pd
housing = pd.read_csv("Housing.csv")
housing.info()
```

```
## <class 'pandas.core.frame.DataFrame'>
## RangeIndex: 506 entries, 0 to 505
## Data columns (total 6 columns):
##  #   Column   Non-Null Count  Dtype
## ---  ------   --------------  -----
##  0   crim     506 non-null    float64
##  1   zn       506 non-null    float64
##  2   river    506 non-null    int64
##  3   rm       506 non-null    float64
##  4   ptratio  506 non-null    float64
##  5   medv     506 non-null    float64
## dtypes: float64(5), int64(1)
## memory usage: 23.8 KB
```

3. Based on the above information, do we have the missing data issue in the `housing` dataset?

- A) Yes
- B) No

4. If we were to predict the `medv` using *four* features in the dataset, how many linear regression models can we potentially consider?

- A) 506
- B) 20
- C) 10
- D) 5

5.

```
import statsmodels.formula.api as smf
result1 = smf.ols('medv ~ river + ptratio', data=housing).fit()
result1.summary()
```

```
## <class 'statsmodels.iolib.summary.Summary'>
## """
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                   medv   R-squared:                       0.271
## Model:                            OLS   Adj. R-squared:                  0.268
## Method:                 Least Squares   F-statistic:                     93.46
## Date:                Fri, 04 Mar 2022   Prob (F-statistic):           3.06e-35
## Time:                        17:54:35   Log-Likelihood:                -1760.3
## No. Observations:                 506   AIC:                             3527.
## Df Residuals:                     503   BIC:                             3539.
## Df Model:                           2
## Covariance Type:            nonrobust
## ==============================================================================
##                  coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------
## Intercept     60.9579      3.041     20.048      0.000      54.984      66.932
## river          4.1735      1.389      3.005      0.003       1.445       6.902
## ptratio       -2.0977      0.163    -12.874      0.000      -2.418      -1.778
## ==============================================================================
## Omnibus:                       77.406   Durbin-Watson:                   0.759
## Prob(Omnibus):                  0.000   Jarque-Bera (JB):              145.280
## Skew:                           0.883   Prob(JB):                     2.84e-32
## Kurtosis:                       4.942   Cond. No.                         162.
## ==============================================================================
##
## Notes:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
## """
```

What are the independent variables (i.e., features) in this regression?

- A) medv, river, ptratio
- B) river, ptratio
- C) river only
- D) ptratio only

6. Now you plan to regress `medv` on `rm`, `crim` and `zn`. Before we implement this regression, John says he expects a higher (in-sample) R^2 than what you got for regressing `medv` on `river` and `ptratio` only, because the new model has 3 predictors while the existing one just has 2. Do you agree with John's reasoning?

- A) Yes, because the more predictors in the regression model, the larger the in-sample R^2
- B) No, because the new model does not include the 2 predictors in the existing model.

7. Based on the regression output from Problem 5, what is the absolute value of the correlation between the variables `medv` and `ptratio`?

- A) ≥ 0.53
- B) < 0.53

8. Suppose we want to randomly split `housing` into training and test parts with 30% as the test data. You split the data twice, both using the `train_test_split` function from `sklearn.model_selection`. For the first split, you set random state equals 0; and for the second split, you set random state equals 1. Do you expect that the training data sets from these two splits are the same?

- A) Yes, because each split gives 30% to test data
- B) Yes, because the value of random state does not influence the splits.
- C) Yes, and the reasons offered in both A) and B) are correct.
- D) No

9. Let h be a classifier for binary classification. Recall that $P(h(X) \neq Y|Y = 0)$ is type I error and $P(h(X) \neq Y|Y = 1)$ is type II error. Suppose $P(Y = 1) = P(Y = 0) = 1/2$. Consider two scenarios: (i) $X|(Y = 0) \sim \mathcal{N}(0, 1)$ and $X|(Y = 1) \sim \mathcal{N}(2, 1)$; (ii) $X|(Y = 0) \sim$ Uniform$[0, 1]$ and $X|(Y = 1) \sim$ Uniform$[2, 3]$. In which of these two scenarios does there exist a classifier, such that both type I error and type II error are zero?

- A) i) only
- B) ii) only
- C) Both i) and ii)
- D) Neither i) nor ii)

10. The LDA classifier does not assume a probabilistic model.

- A) The statement is true.
- B) The statement is false.

11. How many of the following statement(s) about logistic regression is/are correct?

- i) The logistic regression model is $P(Y = 1|X = x) = e^{\beta_0+\beta_1 x}/(1 + e^{\beta_0+\beta_1 x}) + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

- ii) Logistic regression is always better than LDA.

Choose one of the following:

- A) i)
- B) ii)
- C) i) and ii)
- D) neither i) nor ii)

12. Suppose that we take a dataset, 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 25% on the training data and 40% on the test data. Next we use 5-nearest neighbors (i.e., $K = 5$) and get training error a and test error b. Although we do not know the exact numbers for a and b, we know that $(a + b)/2 = 16\%$. Based on these results, which method should we prefer to use for classification of new observations?

- A) Logistic regression
- B) 5-nearest neighbors
- C) the two methods are actually equally good.

Questions $13 - 15$ are based an `email spam` data. For simplicity, one can think of the first 57 columns as engineered features from the original emails, while the last column indicates whether an email is spam (class 1) or non-spam (class 0). There are 1813 spam emails. There is no missing data in this dataset.

```
df_spam = pd.read_csv("spambase.data", header = None)
df_spam.shape
```

```
## (4601, 58)
```

13. We want to split the `email spam` data into the training and test sets. Should we specify `stratify=y` in `train_test_split`?

- A) Yes, because we are dealing with a classification problem
- B) No, because we are dealing with a regression problem

14. Do you agree with the statement: KNN with $K = 4601$ is a good classification method for this dataset?

- A) Yes
- B) No

15. If we train logistic regression on this dataset, the default threshold of the fitted sigmoid function is 0.5. If we decrease the threshold from 0.5 to 0.4, how will type II error change?

- A) Increase
- B) Decrease
- C) Changing the threshold should not have any impact on type II error.

16. In classification problems, *interpolation* means all training data are classified correctly. Among the following classification methods, which one gives interpolation?

- A) logistic regression
- B) KNN with $K = 1$
- C) KNN with $K = 3$
- D) linear discriminant analysis

part b) short answer questions

17. Explain how `9-nearest-neighbors` classification method works in a 3-class problem where the classes are coded by $\{1, 2, 3\}$.

18. Your dataset has two features (i.e., two predictors or two independent variables) and the response is coded by 0 and 1. This dataset contains 4 observations: (3, 8, 1), (2, 4, 1), (9, 8, 0) and (10, 3, 0), where in each triple the the first two coordinates are feature values and the last one is the class label. You want to run logistic regression. Please write down the likelihood function.

19. The ROC space is a unit square (recall that the horizontal axis is for type I error, and the vertical axis is 1-type II error). Draw an ROC space and draw the line that represents all kinds of random guesses.

20. Write down the simple linear regression model. How many parameters are there in this model?

21. In a cancer diagnosis problem, we coded the cancer class as 0 and normal class as 1. You trained a logistic regression classifier based on some training data. Suppose the type I error of this classifier is 0.5 and the type II eror of this classifier is 0.2. Are you satisfied with these errors and why? If not, how will you adjust the threshold (either increase or decrease) of the fitted sigmoid function to cater your concern.

22.

```
X = [[1, 2, np.nan], [3, np.nan, 3], [np.nan, 7, 5], [8, 9, 7]]
df = pd.DataFrame(X, columns=['A', 'B', 'C'])
df
```

```
##      A    B    C
## 0  1.0  2.0  NaN
## 1  3.0  NaN  3.0
## 2  NaN  7.0  5.0
## 3  8.0  9.0  7.0
```

There is one missing value in the `C` column. To use KNNImputer with $k = 2$ to fill it up, we need to calculate the distance between the 0th row and other rows. Denote the squares of these distances by $d^2(0,1)$, $d^2(0,2)$ and $d^2(0,3)$. Please find $d^2(0,1)$, $d^2(0,2)$ and $d^2(0,3)$. Show the calculation steps as well as the final results. (Hint: Tutorial 2)