Test 1 Solution

DSO 530: Applied Modern Statistical Learning Methods

March, 4th, 2020

You have 90 minutes to do the problems. For multiple choice questions (1-16), make sure to read the questions very carefully and **circle** the **best** answer. If you circle 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, but no electronic equipment other than a calculator is allowed. Questions 1, 17, 18, 19, 20, 21 and 22 are worth 2 points each, and the rest are worth 1 point each. The total points are 29.

Name (Last, First):

USC Student ID number:

Sign on the line below to pledge that you did not give (receive) assistance on the questions for this exam to (from) anyone else.

---

part a) multiple choices

1. Let A and B be two events. Given $P(A \text{ and } B) = 0.6$ and $P(B^c|A) = 0.7$, what is $P(B)$? (Hint: B^c denotes the complement of B and you might need $P(A \text{ and } B)/P(A) = P(B|A)$)

- A) 0.7
- B) 0.6
- C) 6/7
- D) $P(B)$ can be calculated based on the current information and it is between 0 and 0.5
- E) $P(B)$ cannot be calculated even if we have additional information, because numbers in the problem are wrong

Answer: E). $P(B|A) = 1 - P(B^c|A) = 0.3$. Then $P(A) = P(A \text{ and } B)/P(B|A) = 0.6/0.3 = 2$. No probability should be bigger than 1.

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

i) It measures linear relationship between two categorical variables.

ii) It takes values between -1 and 1.

iii) If $r = 1$, the strength of the linear relationship is stronger than that when $r = -1$.

iv) r is not a robust measure.

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

- E) Four

Answer: C). The statements ii) and iv) are correct

Questions $3-9$ are based the `Boston` 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
from sklearn.datasets import load_boston
boston_dataset = load_boston()
boston = pd.DataFrame(boston_dataset.data, columns=boston_dataset.feature_names)
boston['MEDV'] = boston_dataset.target
boston.info()
```

```
## <class 'pandas.core.frame.DataFrame'>
## RangeIndex: 506 entries, 0 to 505
## Data columns (total 14 columns):
## CRIM       506 non-null float64
## ZN         506 non-null float64
## INDUS      506 non-null float64
## CHAS       506 non-null float64
## NOX        506 non-null float64
## RM         506 non-null float64
## AGE        506 non-null float64
## DIS        506 non-null float64
## RAD        506 non-null float64
## TAX        506 non-null float64
## PTRATIO    506 non-null float64
## B          506 non-null float64
## LSTAT      506 non-null float64
## MEDV       506 non-null float64
## dtypes: float64(14)
## memory usage: 55.5 KB
```

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

- A) Yes
- B) No

Answer: B)

4. If we were to predict the `MEDV` using two other variables in the dataset, how many linear regression models can we potentially consider?

- A) 506
- B) 13
- C) $(13 \times 12)/2$
- D) $(14 \times 13)/2$

Answer: C)

5.

```
import statsmodels.formula.api as smf
result1 = smf.ols('MEDV ~ LSTAT + RM + PTRATIO', data=boston).fit()
result1.summary()
```

```
## <class 'statsmodels.iolib.summary.Summary'>
## """
##                             OLS Regression Results
## ==============================================================================
## Dep. Variable:                   MEDV   R-squared:                       0.679
## Model:                            OLS   Adj. R-squared:                  0.677
## Method:                 Least Squares   F-statistic:                     353.3
## Date:                Mon, 02 Mar 2020   Prob (F-statistic):          2.69e-123
## Time:                        22:26:01   Log-Likelihood:                -1553.0
## No. Observations:                 506   AIC:                             3114.
## Df Residuals:                     502   BIC:                             3131.
## Df Model:                           3
## Covariance Type:            nonrobust
## ==============================================================================
##                  coef    std err          t      P>|t|      [0.025      0.975]
## ------------------------------------------------------------------------------
## Intercept     18.5671      3.913      4.745      0.000      10.879      26.255
## LSTAT         -0.5718      0.042    -13.540      0.000      -0.655      -0.489
## RM             4.5154      0.426     10.603      0.000       3.679       5.352
## PTRATIO       -0.9307      0.118     -7.911      0.000      -1.162      -0.700
## ==============================================================================
## Omnibus:                      202.072   Durbin-Watson:                   0.901
## Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1022.153
## Skew:                           1.700   Prob(JB):                    1.10e-222
## Kurtosis:                       9.076   Cond. No.                         402.
## ==============================================================================
##
## Warnings:
## [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
## """
```

What percentage of the variation of the response does the regression explain?

- A) 67.9%
- B) 67.7%
- C) 353.3
- D) We are not given proper output to decide.

Answer: A)

6.

```
x_new = {'LSTAT': [5, 6], 'RM': [3,5], 'PTRATIO': [15, 20]}
x_new_df = pd.DataFrame(x_new)
prediction1 = result1.get_prediction(x_new_df)
prediction1.summary_frame()
```

```
##         mean   mean_se  ...  obs_ci_lower  obs_ci_upper
## 0  15.293508  1.709437  ...      4.484303     26.102712
## 1  19.098931  0.789436  ...      8.708321     29.489541
##
## [2 rows x 6 columns]
```

In the above implementation, we predicted the `MEDV` for two suburbs. The first suburb has `LSTAT` $= 5$, `RM` $=$ 3, and `PTRATIO` $= 15$, and the second one has `LSTAT` $= 6$, `RM` $= 5$ and `PTRAIO` $= 20$. Which suburb has a higher predicted `MEDV`?

- A) First
- B) Second

Answer: B)

7. In the dataset `Boston`, what is the absolute value of the correlation between the variables `MEDV` and `RM`? (Hint: check out the summary in problem 5)

- A) > 0.83
- B) ≤ 0.83

Answer: B) We get the $R^2 = 0.679$ in the summary. This should be bigger than the squared correlation that we are interested. $\sqrt{0.679} = 0.824$.

8. If we regress `MEDV` on `LSTAT`, `RM` and `PTRATIO` plus another variable on the whole `Boston` dataset, and get $R^2 = 0.72$. Compare to the R^2 we had previously (with `LSTAT`, `RM` and `PTRATIO` only), this R^2 is higher. Does this suggest that the new model is better?

- A) Yes, because a higher R^2 indicates a better model
- B) No, because these two R^2's are not directly comparable

Answer: B)

9. Suppose we want to randomly split `Boston` 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

Answer: D)

10. For most classification problems, as the training sample size increases to infinity, we expect that the test error will decrease to 0.

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

Answer: B)

11. Which of the following statement(s) about the Linear discriminant analysis (`LDA`) and logistic regression (`LR`) is/are correct?

- i) Both LDA and LR have linear decision boundaries.
- ii) LDA model assumes that $X|(Y=0)$ and $X|(Y=1)$ follow two Normal distributions with different means and different covariances.

Choose one of the follwoing:

- A) i)
- B) ii)
- C) i) and ii)

Answer: A)

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 20% on the training data and 38% 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.

Answer: B)

Questions $13-15$ are based the `Smarket` data that you have seen in a tutorial. This data set consists of percentage returns for the S&P 500 stock index over 1, 250 days, from the beginning of 2001 until the end of 2005. For each date, we have recorded the percentage returns for each of the five previous trading days, `Lag1` through `Lag5`. We have also recorded `Volume` (the number of shares traded on the previous day, in billions), `Today` (the percentage return on the date in question) and `Direction` (whether the market was Up or Down on this date).

```
smarket = pd.read_csv('smarket.csv')
smarket['Up'] = np.where(smarket['Direction'] == 'Up', 1, 0)
smarket.head()
```

```
##    Year   Lag1   Lag2   Lag3   Lag4   Lag5  Volume  Today Direction  Up
## 0  2001  0.381 -0.192 -2.624 -1.055  5.010  1.1913  0.959        Up   1
## 1  2001  0.959  0.381 -0.192 -2.624 -1.055  1.2965  1.032        Up   1
## 2  2001  1.032  0.959  0.381 -0.192 -2.624  1.4112 -0.623      Down   0
## 3  2001 -0.623  1.032  0.959  0.381 -0.192  1.2760  0.614        Up   1
## 4  2001  0.614 -0.623  1.032  0.959  0.381  1.2057  0.213        Up   1
```

13.

```
result = smf.logit('Up ~ Lag1 + Lag2 + Lag3', data=smarket).fit()
```

```
## Optimization terminated successfully.
##          Current function value: 0.691349
##          Iterations 4
```

```
result.summary()
```

```
## <class 'statsmodels.iolib.summary.Summary'>
## """
##                            Logit Regression Results
## ==============================================================================
## Dep. Variable:                     Up   No. Observations:                 1250
## Model:                          Logit   Df Residuals:                     1246
## Method:                           MLE   Df Model:                            3
## Date:                Mon, 02 Mar 2020   Pseudo R-squ.:                0.001619
## Time:                        22:26:01   Log-Likelihood:                -864.19
## converged:                       True   LL-Null:                       -865.59
## Covariance Type:            nonrobust   LLR p-value:                    0.4230
## ==============================================================================
##                  coef    std err          z      P>|z|      [0.025      0.975]
## ------------------------------------------------------------------------------
## Intercept      0.0742      0.057      1.310      0.190      -0.037       0.185
## Lag1          -0.0714      0.050     -1.425      0.154      -0.170       0.027
## Lag2          -0.0443      0.050     -0.885      0.376      -0.142       0.054
## Lag3           0.0089      0.050      0.178      0.859      -0.089       0.107
## ==============================================================================
## """
```

How many predictors in the above logistic regression are significant at 5% level? (Hint: look at the fourth column in the above table)

- A) 0
- B) 1
- C) 2
- D) 3
- E) 4

Answer: A)

14. In logistic regression, the default threshold of the fitted signoid function is 0.5. If we decrease the threshold from 0.5 to 0.48, how will type II error change? (Hint: type II error is defined to be the probablity of classifying class 1 instances as class 0.)

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

Answer: B)

15. We want to split the `Smarket` data into the training and test sets. Which of the following splits is the most reasonable among the three choices.

- A) Use instances in the years 2001-2004 as training as those in year 2005 as test
- B) Randomly select 80% as training as the remaining 20% as test

- C) Us instances in the years 2002-2005 as training as those in the year 2001 as test

Answer: A)

16. Which of the following is true about supervised learning?

- A) The overall classification error is the only thing people care in evaluating classifiers' performance.
- B) A classifier that is in the lower right half of the ROC space is inferior to some random guess.
- C) In linear regression, R^2 on test data should always be smaller than R^2 on training data.

Answer: B)

part b) short answer questions

17. Explain how `7-nearest-neighbors` works on a three-class classification problem where the classes are coded by $\{1, 2, 3\}$.

Answer: To classify an observation using `7-nearest-neighbors`, we look for 7 closest observations in the training set, and assign the class label according to the plurality of the labels among these 7 observations.

18. Your dataset has two predictors and the response is coded by 0 and 1. This dataset contains 4 observations: two class 1 observations (3, 8) and (2, 4) and two class 0 observations (9, 8) and (10, 3). You want to run logistic regression. Please write down the likelihood function.

Answers: The likelihood function is

$$\frac{\exp(\beta_0 + 3\beta_1 + 8\beta_2)}{1 + \exp(\beta_0 + 3\beta_1 + 8\beta_2)} \cdot \frac{\exp(\beta_0 + 2\beta_1 + 4\beta_2)}{1 + \exp(\beta_0 + 2\beta_1 + 4\beta_2)} \cdot \frac{1}{1 + \exp(\beta_0 + 9\beta_1 + 8\beta_2)} \cdot \frac{1}{1 + \exp(\beta_0 + 10\beta_1 + 3\beta_2)}$$

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

Answer: The simple linear regression model is $y = \beta_0 + \beta_1 x + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$. There are three parameters in this model.

20. Explain the difference between regression and classification. For Medicare-Medicaid fraud detection problem, should you formulate it as regression, classification or neither, and why?

Answer: In regression, the response variable is numerical and in classification, the response variable is categorical. For Medicare-Medicaid fraud detection problem, it seems that we might formulate as a classification problem, where the response variable takes the values `fraud` and `non-fraud`. However, people have a common belief that in this problem only a small fraction of the fraud were actually found; in other words, the labels in theory do exists but we do not have an accurate record in our available dataset. Hence, we cannot formulate this problem as either regression or classification.

21.

```
arr1 = np.arange(8).reshape(2,2,2)
arr2 = arr1[1].copy()
arr2[1] = 50
arr1; arr2
```

```
## array([[[0, 1],
##         [2, 3]],
##
##        [[4, 5],
##         [6, 7]]])
## array([[ 4,  5],
##        [50, 50]])
```

After executing the above codes, what are `arr1` and `arr2`? Please write down these NumPy arrays.

22.

```
X = [[1, 2, np.nan], [3, 4, 3], [8, np.nan, 5], [8, 9, 7]]
df = pd.DataFrame(X, columns=['A', 'B', 'C'])
df
```

```
##    A    B    C
## 0  1  2.0  NaN
## 1  3  4.0  3.0
## 2  8  NaN  5.0
## 3  8  9.0  7.0
```

There is one missing value in the `B` column and one missing value in the `C` column. Please use mean imputation to fill them up. Write down not only the numbers but also the mathematical derivation to get those numbers. (Don't worry about the training / test splits in this problem. Just impute `df`.)

Answer: To fill up the missing value in column `B`: $(2+4+9)/3 = 5$, and to fill up the missing value in column `C`: $(3+5+7)/3 = 5$.