CS 484 Machine Learning

Autumn 2020 Midterm Test

Instruction

1. Calculate your answers using all the available precision
2. If the final numeric answer has more than four decimal places, then round the numeric answer to the nearest fourth decimal place. Otherwise, please give the exact value.

Question 1 (5 points)

Based on the MNLogit output below, what statistical phenomenon may have happened?

```
Optimization terminated successfully.
         Current function value: 0.479393
         Iterations 24
                          MNLogit Regression Results
==============================================================================
Dep. Variable:                    BAD   No. Observations:                 4981
Model:                        MNLogit   Df Residuals:                     4973
Method:                           MLE   Df Model:                            7
Date:                Mon, 02 Mar 2020   Pseudo R-squ.:                0.06869
Time:                        11:21:49   Log-Likelihood:               -2387.9
converged:                       True   LL-Null:                      -2564.0
Covariance Type:            nonrobust   LLR p-value:                4.104e-72
==============================================================================
         BAD=1       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const             -0.8412        nan        nan        nan         nan         nan
REASON_DebtCon    -0.5134   1.26e+06  -4.06e-07      1.000  -2.48e+06   2.48e+06
REASON_HomeImp    -0.3278   1.37e+06  -2.39e-07      1.000  -2.69e+06   2.69e+06
JOB_Mgr           -0.0073   7.27e+05  -1.01e-08      1.000  -1.42e+06   1.42e+06
JOB_Office        -0.7000   7.27e+05  -9.63e-07      1.000  -1.42e+06   1.42e+06
JOB_Other         -0.0939   7.27e+05  -1.29e-07      1.000  -1.42e+06   1.42e+06
JOB_ProfExe       -0.5037   7.27e+05  -6.93e-07      1.000  -1.42e+06   1.42e+06
JOB_Sales          0.4530   7.27e+05   6.23e-07      1.000  -1.42e+06   1.42e+06
JOB_Self           0.0106   7.27e+05   1.46e-08      1.000  -1.42e+06   1.42e+06
DEROG              0.6917      0.049     14.167      0.000       0.596       0.787
==============================================================================
```

(A) The complete separation – maybe this

(B**) The quasi-complete separation**

(C) The iteration did not converge

(D) Nothing unusual has happened

Question 2 (5 points)

Suppose we train a classification tree on a nominal target field that has five categories. What is the highest Gini Index value that we can see in any node?

(A) 0

(B) 0.5

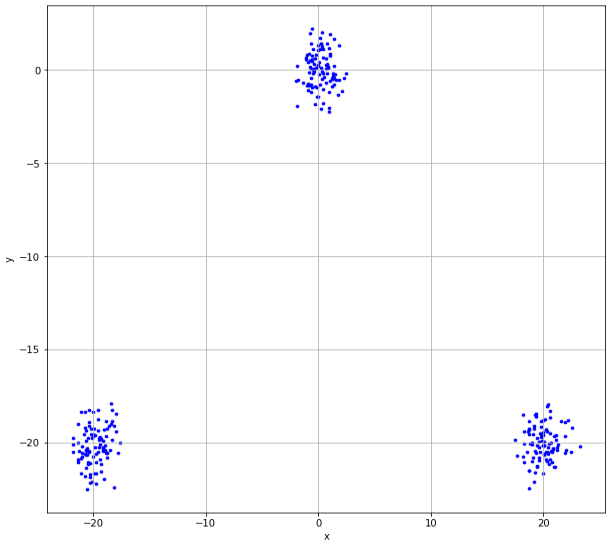
(C) 0.8

(D) 1

(E) 2.3219

Question 3 (5 points)

We have generated the following scatterplot of two fields x and y. Suppose we are going to perform the K-means clustering analysis on the data in the scatterplot. Which of the following statements is valid about the Silhouette value for the 3-cluster solution?



(A) Close to the negative one

(B) About zero

(C) Close to one

(D) Close to three

(E) Cannot be determined

Question 4 (5 points)

Suppose there are 100 unique items in the universal set, how many 5-itemset can we possibly generate?

(A) 100

(B) 120

(C) 75,287,520

(D) 9,034,502,400

(E) 1,267,650,600,228,229,401,496,703,205,375

Question 5 (5 points)

The CSV file 2020S2MT_Q5.csv contains 100 values. Suppose we constructed a histogram on these 100 values with a bin-width of 5. What is the empirical density for the bin where the mid-point is x = 97.5?

0.0303030303 WRONG -- bruh not this? I calculated it using python and similar way to HW1

But when I did it manually I got 15 values within the range … so 15/100? maybe

Question 6 (5 points)

The support of the Consequent of an association rule X → Y is numerically equal to the Expected Confidence of the association rule.

True or False?

True

Question 7 (5 points)

Suppose we trained a binary classification model and the Area Under Curve value is one. Then what is the Sensitivity value when the False Positive Rate is 37.5%?

(A) 0%

(B) 37.5%

(C) 50%

(D) 100%

(E) Cannot be determined – WRONG, NOT THIS – but searched online, AUC one doesn't mean accuracy 100% or whatever

Question 8 (5 points)

Suppose we trained a classification tree using 5,000 observations. The target field has five categories whose frequencies are listed below. What is the Entropy value of the root node?

Target Category	I	II	III	IV	V
Frequency	262	1,007	1,662	1,510	559

E = -summation(probabilityofcategory * log2(probabilityofcategory))

= -((262/5000)log2(262/5000)+(1007/5000)log2(1007/5000)+(1662/5000)log2(1662/5000)+(1510/5000)log2(1510/5000)+(559/5000)log2(559/5000))

= 0.22292 + 0.465609 + 0.528186 + 0.5216686 + 0.3534

= **2.0917897**

Question 9 (5 points)

We observed 5,000 observations for a target field that has five categories. The categories are I, II, III, IV, and V. The following table shows their frequencies. We trained a multinomial logistic model that contains only the Intercept terms.

Suppose the reference target category is Category III. What is the estimated Intercept of Category V? (Hint: Suppose the J^{th} target category is the reference category of the target field, then a model that contains only the Intercept terms theory has this formulation: $\log_e\left(\frac{\pi_{ij}}{\pi_{i3}}\right) = \boldsymbol{\beta}_j$ for $j = 1, \ \ldots, K$. Think of an intuitive way to estimate the probabilities and solve for the Intercept terms.)

Target Category	I	II	III	IV	V
Frequency	262	1,007	1,662	1,510	559

0.2526 – using python and using 3 as value in y.where – but different values result in different coefficient of category 5 when doing y.where(…)??? ALSO WRONG

Question 10 (5 points)

I invited six friends to my home to watch a basketball game. My friends brought snacks and beverages along. The following table lists the items my friends brought.

Friend	Items
Andrew	Cheese, Cracker, Salsa, Soda, Tortilla, Wings
Betty	Cheese, Soda, Tortilla
Carl	Cheese, Ice Cream, Soda, Wings
Danny	Cheese, Ice Cream, Salsa, Tortilla, Wings
Emily	Salsa, Soda, Tortilla, Wings
Frank	Cheese, Cracker, Ice Cream, Soda, Wings

I noticed that a few of my friends brought Cheese, Soda, and Wings together. Since I prefer to spend your money on other food besides Wings, I am curious to know how likely my friends will bring Wings if they have already brought Cheese and Soda. Therefore, please help me determine the Lift of this association rule {Cheese, Soda} ==> {Wings}.

X = {Cheese, Soda}, Y = {Wings}, X -> Y

Lift(X -> Y) = Support(X, Y)/(Support(X) * Support(Y))

= (3/6)/((4/6)*(5/6)) = **9/10 = 0.9**

Question 11 (5 points)

Given the following confusion matrix

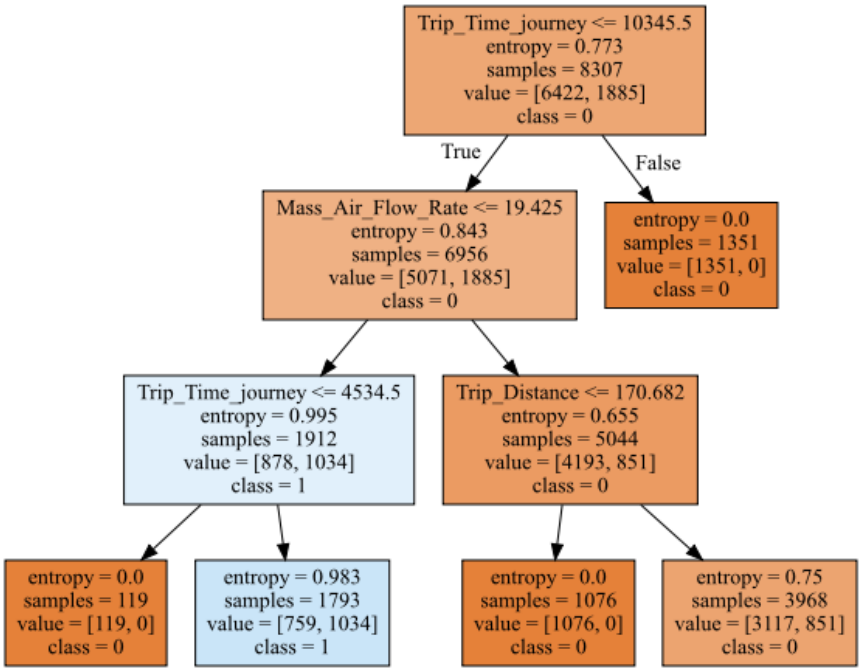
	Predicted Event	Predicted Non-Event
Observed Event	18	7
Observed Non-Event	12	22

What is the F1 Score value?

F1 = tp/(tp+0.5(fp+fn)) = 18/(18+0.5(12+7)) = **0.654545**

Question 12 (5 points)

We are given the following classification tree diagram. The target variable has two categories 0 and 1.



Suppose we are interested in predicting the class 1 of the target variable. In other words, the predicted event probability is the Prob (class = 1). What is the Root Average Squared Error?

??? get all the samples that resulted in class 0 (thus error), in leaf nodes, then divided by the total amount – then square root

$$\sqrt{\frac{\left(1-\frac{1034}{1793}\right)^2+\left(0-\frac{3117}{3968}\right)^2}{2}} = \sqrt{\frac{(1-0.57668)^2+(0-0.785534)^2}{2}} = \sqrt{\frac{(0.42332)^2+(0.785534)^2}{2}} = 0.630977$$

or 0.33555 – similar but values squared are 759/1793 and 851/3968 – think this is more right

0.33555 IS WRONG

Questions 13, 14, and 15

We are going to train a multinomial logistic regression on 5,000 observations. The target field has five categories, namely, A, B, C, D, and E. The categorical feature has four categories, namely, I, II, III, and IV. Instead of a casewise dataset, the data have been aggregated and shown in the following table.

	Target Field				
Feature	A	B	C	D	E
I	65	304	530	487	140
II	74	185	160	55	16
III	33	228	623	755	363
IV	90	290	349	213	40

We will use the Deviance Test to determine if the categorical feature is predictive for the categorical target. We will first train a model, say M_0, with only the Intercept terms. The log-likelihood value of the model M_0 is -7249.5908 and the degree of freedom is 4. Next, we train another model M_1 with the Intercept terms and the categorical feature.

Question 13 (5 points)

What is the log-likelihood value of the model M_1?

-6918.516197

(Hint: When a multinomial logistic model includes only one categorical predictor, what are the predicted probabilities without training the model?)

Question 14 (5 points)

What is the degree of freedom of the model M_1?

12 – WRONG, ????, Got loglikelihood right, and says 12 on the mnlogit output… bruh

Question 15 (5 points)

What is the significance value of the Deviance Test?

(A) 1.4536E-130

(B) 5.5356E-134

(C) 1.3696E-63

(D) 1.0

(E) None of the Above

Question 16 (5 points)

You are going to build a logistic model using the 20 observations below. The binary target field is y, and the interval predictor is x.

x	0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4
y	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	1	1	1

The specifications are:

1. The target event category is 1
2. The Intercept term is included
3. The optimization method is Newton
4. The maximum number of iterations is 100
5. The tolerance level is 1e-8.

After you have built your model, you will apply them to the following test data and then calculate the misclassification rate metric. An observation will be classified as an event if the predicted event probability is greater than or equal to 0.3.

x	0	1	2	3	4
y	1	0	1	0	1

What is the Misclassification Rate when the Logistic model is applied to the test data?

3/5, 0.6 misclassifcation rate

	0	1
0	0.989064	0.010936
1	0.954229	0.045771
2	0.827752	0.172248
3	0.525554	0.474446
4	0.203402	0.796598

Questions 17 and 18

You can use Chicago's 311 Service Request to report street potholes. After a request has been received, the Department of Transportation will first assess the severity of the pothole, and then schedule a road crew to fill up the pothole. After the pothole is filled, the service request will be closed.

You are provided with this CSV file **ChicagoCompletedPotHole.csv** for analyzing the city's efforts to fill up street potholes. The data contains 17,912 observations. Each observation represents a completed request which was created between December 1, 2017 and March 31, 2018 and was completed between December 4, 2017 and September 12, 2018. The data has the following seven variables.

Name	Level	Description
1) CASE_SEQUENCE	Nominal	A unique index for identifying an observation
2) WARD	Nominal	Chicago's ward number from 1 to 50
3) CREATION_MONTH	Nominal	Calendar month when the request was created
4) N_POTHOLES_FILLED_ON_BLOCK	Interval	Number of potholes filled on the city block
5) N_DAYS_FOR_COMPLETION	Interval	Number of days elapsed until completion
6) LATITUDE	Interval	Latitude of the city block
7) LONGITUDE	Interval	Longitude of the city block

You will use the K-Means Clustering algorithm to identify clusters in the entire data with the following specifications.

1. Use $\log_e$(N_POTHOLES_FILLED_ON_BLOCK), $\log_e$(1 + N_DAYS_FOR_COMPLETION), LATITUDE, and LONGITUDE (i.e., you need to perform the transformations before clustering)
2. The maximum number of clusters is 10 and the minimum number of clusters is 2
3. The random seed is 20201014
4. Use both the Elbow and the Silhouette methods to determine the number of clusters

Question 17 (5 points)

What is the optimal number of clusters? Please give the number of clusters as an integer.

4

Question 18 (5 points)

What is the Calinski-Harabasz Score for that optimal number of clusters?

17357.358719049626

Questions 19 and 20

In the automobile industry, a common question is how likely a policyholder will file a claim during the coverage period. Your task is to build a logistic model. To avoid discriminating policyholders, we will use predictors that can be verified and are related to the risk exposures of the policyholders. The CSV file policy_2001.csv contains data about 617 policyholders. We will use only the following variables.

Target Variable

- CLAIM_FLAG: Claim Indicator (1 = Claim Filed, 0 = Otherwise) and 1 is the event value.

Nominal Predictor

- CREDIT_SCORE_BAND: Credit Score Tier ('450 – 619', '620 – 659', '660 – 749', and '750 +')

Interval Predictors

- BLUEBOOK_1000: Blue Book Value in Thousands of Dollars (min. = 1.5, max. = 39.54)
- CUST_LOYALTY: Number of Years with Company Before Policy Date (min. = 0, max. ≈ 21)
- MVR_PTS: Motor Vehicle Record Points (min. = 0, max. = 10)
- TIF: Time-in-Force (min. = 101, max. = 107)
- TRAVTIME: Number of Miles Distance Commute to Work (min. = 5, max. ≈ 93)

Since the tools may not take the nominal predictor as is, you will first derive the dummy indicators from the nominal predictors and then use the dummy indicators in building the models. You will build the logistic model according to the following specifications.

- The optimization algorithm is the Newton-Raphson method
- The maximum number of iterations is 100
- The relative error in parameter estimates acceptable for convergence is 1E-8
- The Intercept term must be included in the model
- Use the Forward Selection method to enter predictors, the alpha level is 0.05.

You will divide the data into the Training and the Test partitions. You will build the logistic model using the Training partition. Later, you will evaluate the model based on the Test partition.

Data Partition

- The Training partition consists of 67% of the original observations, the remaining 33% goes to the Test partition.
- Use the CLAIM_FLAG as the stratum variable.
- The random seed is 20201014.

Question 19 (5 points)

Which predictors are selected into the final logistic model?

Using R but not newton or whatever and no alpha whatever – BLUEBOOK_1000, MVR_PTS, TRAVTIME

Using python same – MVR_PTS, BLUEBOOK_1000, TRAVTIME

Question 20 (5 points)

What is the Area Under Curve value of the Testing Partition?

0.538623 --- WRONG bruh