CS 484 Introduction to Machine Learning

Fall Semester 2022 Midterm Test Answer Key

Question 1 (5 points)

Which of the following statement(s) best describes Machine Learning?

Multiple Choice:

(A) Machine learning is an automated process that uses algorithms to identify patterns within data, and those patterns are then used to create a data model that can make predictions.

(B) Machine learning is an idea that systems can learn from data, identify patterns, and make decisions with minimal human intervention.

(C) A computer program is said to learn from experience E for some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E. Machine learning refers to the field of study concerned with these programs or systems.

(D) All of the Above

(E) None of the Above

The correct answer is D.

Question 2 (5 points)

What does the term Label mean in the context of Machine Learning?

Multiple Choice:

(A) A cute sticker that is affixed to the hardware on where the machine learning process is running.

(B) A user-defined name that is attached to a version of your machine learning computer codes.

(C) A system-level command that creates, changes, or deletes a logical label on your dataset.

(D) A label is what a machine learning algorithm will predict or forecast. Put it another way, it is the target or response field.

(E) There is no such term in the context of Machine Learning.

The correct answer is D.

Question 3 (5 points)

Suppose the itemset {A, B, C, D, E} has a Support value of 1, then what is the Lift value of this rule {B, D} ➔ {A, C, E}?

Multiple Choice:

(A) 0

(B) 0.5

(C) 1

(D) 2

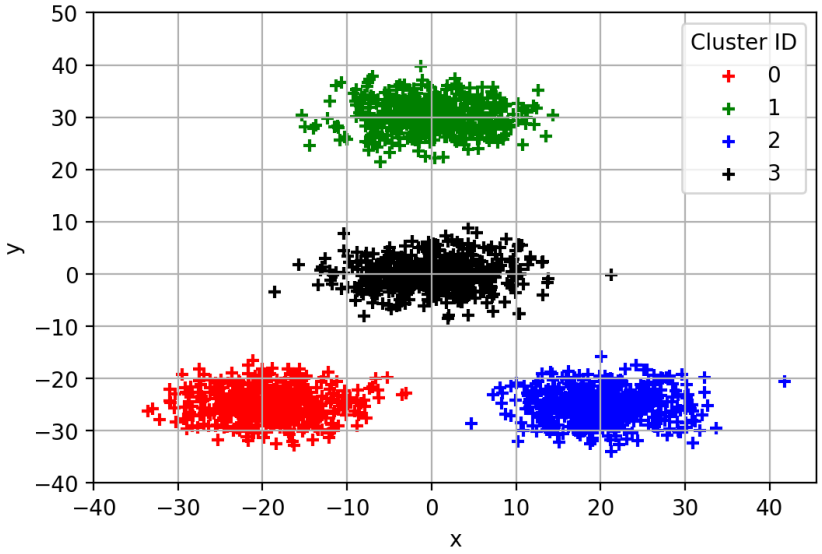
(E) Cannot be Determined

The correct answer is C. It is known that adding more items to an itemset may lower the Support, therefore, Support of {A, C, E} $\geq$ Support of {A, B, C, D, E} and Support of {B, D} $\geq$ Support of {A, B, C, D, E}. Since the question says that Support of {A, B, C, D, E} = 1, we have Support of {A, C, E} $\geq$ 1 and Support of {B, D} $\geq$ 1. As Support cannot exceed 1, therefore, Support of {A, C, E} = 1 and Support of {B, D} = 1.

The Confidence value of the rule {B, D} ➔ {A, C, E} is Support of {A, B, C, D, E} / Support {B, D} = 1. The Expected Confidence value of the rule {B, D} ➔ {A, C, E} is Support of {A, C, E} = 1. Finally, the Lift is Confidence / Expected Confidence = 1. Exact answer required.

Question 4 (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 4-cluster solution?



(A) Close to the negative one

(B) About zero

(C) Close to one

(D) Close to four

(E) Cannot be determined

The correct answer is C. The Silhouette value is bounded between -1 and 1, thus the answer D is not correct. Since the scatterplot clearly shows four non-overlapping clusters of observations, the Silhouette value for this 4-cluster solution should be close to the maximum Silhouette value which is one. Therefore, the correct answer is C. Indeed, the Silhouette value is 0.7527.

Question 5 (5 points)

Suppose there are 100 unique items in the universal set, how many 7-itemset can we possibly generate?

(A) 100

(B) 75,287,520

(C) 9,034,502,400

(D) 16,007,560,800

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

The correct answer is D. The number of 7-itemset is the number of ways of choosing 7 items out of a pool of 100 items. The answer is $C(100, 7) = \frac{100!}{(100-7)!\,7!} = \frac{100!}{93!\,7!} = \frac{100\times99\times98\times97\times96\times95\times94}{7\times6\times5\times4\times3\times2\times1} = 16{,}007{,}560{,}800$. Exact answer required.

Question 6 (5 points)

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

(A) 0

(B) 0.5

(C) 0.75

(D) 1.0

(E) 2.0

The correct answer is C. The highest possible Entropy value is $1 - \frac{1}{K}$ where K is the number of target categories. In this case, $K = 4$, so the highest Entropy value is $1 - \frac{1}{4} = \frac{3}{4} = 0.75$. Exact answer required.

Question 7 (5 points)

The file 1000Values.csv contains 1000 numeric values. Use the Shimazaki and Shinomoto (2007) method and try d = 0.1, 0.2, 0.25, 0.5, 1, 2, 2.5, 5, 10, 20, 25, and 50. What will you recommend for the histogram bin width?

The following table contains the d and the $C(d)$ values based on the Shimazaki and Shinomoto method.

Bin Width	Number of Bins	Criterion
0.1	653	222.7497
0.2	327	-20.2803
0.25	262	-53.1493
0.5	131	-137.3865
1	66	-170.6571
2	33	-189.4974
2.5	27	-192.7634
5	14	-200.6121
10	7	-191.7327
20	4	-135.6787
25	3	-176.1721
50	2	-62.9616

We should recommend the bin width that corresponds to the lowest criterion. The answer is d = 5.

Exact answer required.

Question 8 (5 points)

I invited ten 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, Wings
Carl	Cheese, Ice Cream, Soda, Wings
Danny	Cheese, Ice Cream, Salsa, Tortilla, Wings
Emily	Pizza, Salsa, Soda, Tortilla, Wings
Frank	Cheese, Cracker, Ice Cream, Soda, Wings
Gary	Cracker, Tortilla
Henry	Ice Cream, Pizza, Tortilla
Irene	Cheese, Cracker, Soda
Jack	Cheese, Cracker, Pizza, Salsa, 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}.

The Conference of this rule is Prob(Cheese, Soda, Wings) / Prob(Cheese, Soda) = (4/10) / (5/10) = 4/5. The Expected Conference of this rule is Prob(Wings) = 7/10. Therefore, the Lift = Conference / (Expected Conference) = (4/5) / (7/10) = 8/7 = 1.1429. Accepted interval is (1.1428, 1.1430).

Question 9 (5 points)

Suppose we trained a classification tree using 27,513 observations. The target field has five categories whose frequencies are listed below. What is the Gini Index value of the root node?

Target Category	I	II	III	IV	V
Frequency	6,606	11,324	7,280	2,080	223

The following table shows the steps in getting the answer.

Target Category	I	II	III	IV	V	Total
Frequency	6,606	11,324	7,280	2,080	223	27,513
Proportion (p)	0.2401	0.4116	0.2646	0.0756	0.0081	1.0
p^2	0.05765026	0.16940406	0.07001432	0.00571545	0.00006570	0.30284979

The Gini Index value of the root node is $1 - \sum p^2 = 0.69715021$. The answer is 0.6972. Accepted interval is (0.6971, 0.6973).

Question 10 (5 points)

We observed 27,513 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 II. What is the estimated Intercept of Category V?

Target Category	I	II	III	IV	V
Frequency	6,606	11,324	7,280	2,080	223

According to the model formulation, the Intercept of Category V is the natural logarithm of the ratio of the probability of Category V to the probability of Category II. An intuitive estimate of this ratio of probabilities is the ratio of the respective counts. Therefore, the estimated Intercept of Category V is $\log_e(223/11324) = -3.92750787$. The correct answer is -3.9275. Accepted interval is (-3.9276, -3.9274).

Questions 11 and 12

We performed a cluster analysis with the Chebyshev distance on a data that has five interval variables.

We found two clusters and the following table shows the cluster centroids.

Cluster	X1	X2	X3	X4	X5
0	6.34	6.82	7.21	7.18	7.47
1	8.04	8.56	9.42	8.08	7.70

We now have a new observation: X1 = 9.7, X2 = 10.7, X3 = 11.4, X4 = 7.8, and X5 = 6.5.

Question 11 (5 points)

Which cluster should we assign this new observation to?

The first step is to subtract the centroids from the new observation.

Observation – Centroid	X1	X2	X3	X4	X5
0	3.36	3.88	4.19	0.62	-0.97
1	1.66	2.14	1.98	-0.28	-1.20

The Chebyshev distance is the maximum of the five absolute differences of each cluster. Here are the new observation's Chebyshev distances from the cluster centroids.

Cluster	Chebyshev distances
0	4.19
1	2.14

Since Cluster 1's Chebyshev distance is smaller, we should assign the new observation to Cluster 1.

Question 12 (5 points)

Also, what is the Chebyshev distance from the new observation to the assigned Cluster?

The Chebyshev distance from the new observation to Cluster 1 is 2.14. Exact answer required.

Questions 13, 14, and 15

We are going to train a classification tree on 5,000 observations. We will use the Entropy criterion for growing the tree. The target field has five categories, namely, A, B, C, D, and E. The ordinal feature has four categories where I < II < III < IV. Instead of a casewise dataset, the data have been aggregated and shown in the following table.

Feature	Target Field				
	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

Question 13 (5 points)

Which is the optimal split in the first layer of the classification tree?

Multiple Choice:

(A) {I} + {II, III, IV}

(B) {I, II} + {III, IV}

(C) {I, II, III} + {IV}

(D) None of the Above

The correct answer is C. The root node entropy value is 2.091789726281832. Since Feature is an ordinal variable with four levels, there are only three splits. Here are the possible splits and their entropy values.

Split	Entropy Value
{I} + {II, III, IV}	2.089478999488883
{I, II} + {III, IV}	2.0779208573939933
{I, II, III} + {IV}	2.0654819381740097

The split {I, II, III} + {IV} has the lowest Entropy value and thus it is the optimal split in the first layer.

Question 14 (5 points)

Suppose we continue to split the first layer and create the second layer. What is the optimal split in the second layer?

Multiple Choice:

(A) {I} + {II, III}

(B) {I, II} + {III}

(C) {II, III} + {IV}

(D) {II} + {III, IV}

(E) None of the Above

The correct answer is B. For the node {IV} in the first layer, no further split can be done. For the node {I, II, III}, there are the two possible splits in the second layer and their Entropy values.

Split	Entropy Value
{I} + {II, III}	2.0668992211483874
{I, II} + {III}	2.0213409565840124

The split {I, II} + {III} has the lowest Entropy value and thus it is the optimal split in the second layer.

Question 15 (5 points)

What is the Misclassification Rate of this two layers classification tree?

Here is the summary of the classification tree.

	Target						
Feature	A	B	C	D	E	Predicted Class	Number of Correctly Classified Observations
I, II	139	489	690	542	156	C	690
III	33	228	623	755	363	D	755
IV	90	290	349	213	40	C	349

The Accuracy is (690 + 755 + 349) / 5000 = 1794 / 5000 = 0.3588. The Misclassification Rate is 1 – 0.3588 = 0.6412. Accepted interval is (0.6411, 0.6413).

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?

The iteration converged in 7 iterations. The model summary is below.

```
                          MNLogit Regression Results
==============================================================================
Dep. Variable:                      y   No. Observations:                   20
Model:                        MNLogit   Df Residuals:                       18
Method:                           MLE   Df Model:                            1
Date:                Tue, 13 Oct 2020   Pseudo R-squ.:                 0.3771
Time:                        12:24:58   Log-Likelihood:                -7.6099
converged:                      True   LL-Null:                       -12.217
Covariance Type:            nonrobust   LLR p-value:                 0.002401
==============================================================================
       y=1       coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -4.5047      2.000     -2.253      0.024      -8.424      -0.586
x1             1.4675      0.668      2.197      0.028       0.158       2.777
==============================================================================
```

The predicted event probabilities of the test data are:

x	0	1	2	3	4
y	1	0	1	0	1
Pr(y = 1)	0.0109	0.0458	0.1722	0.4744	0.7966
Predicted Category	0	0	0	1	1

For the test data, the misclassification rate is 3/5 = 0.6. Exact answer required.

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

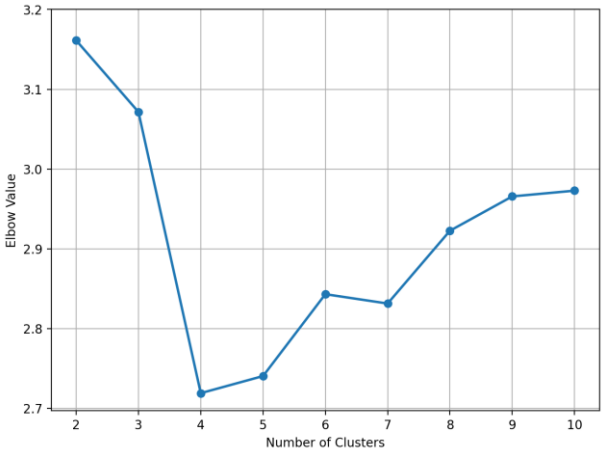
You will use the K-Means Clustering algorithm to identify clusters in the entire data with the following specifications.

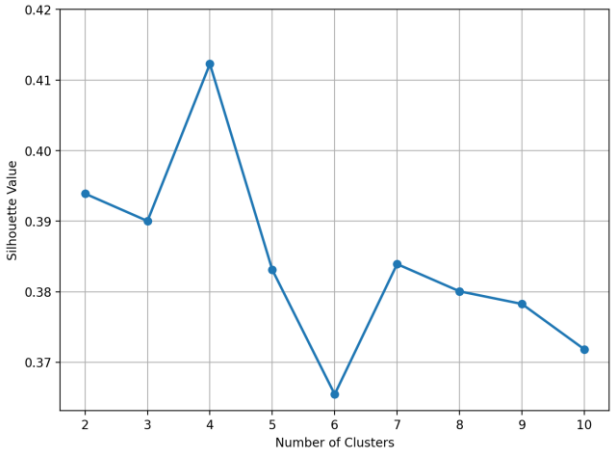
1. Use $\log_e$(N_POTHOLES_FILLED_ON_BLOCK), $\log_e$(1 + N_DAYS_FOR_COMPLETION), LATITUDE, and LONGITUDE (i.e., you need to perform the transformations before clustering)
2. The maximum number of clusters is 10 and the minimum number of clusters is 2
3. The random seed is 2022484
4. Use both the Elbow 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.

The Elbow and the Silhouette charts are shown below.





When the number of clusters is 4, an elbow appeared in the Elbow chart and a local maximum appeared in the Silhouette chart. Therefore, based on these two charts, we determined that the optimal number of clusters is 4.

Question 18 (5 points)

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

When the number of clusters is 4, the Calinski-Harabasz score is 17357.3587. Accepted interval is (17357.3586, 17357.3588).

Questions 19 and 20

In the automobile industry, a common question is how likely a policyholder will file a claim during the coverage period. You will analyze the **policy_2001.csv** that contains data on 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)

You will train a multinomial logistic model according to the following specifications.

- The optimization algorithm is the Newton-Raphson method
- The maximum number of iterations is 200
- The relative error in parameter estimates acceptable for convergence is 1E-8
- The Intercept term must be included in the model
- Use the All Possible Subset method to search for the optimal model.

Question 19 (5 points)

Based on the Akaike Information Criterion, which predictors are selected into the final logistic model?

The logistic model that yields the lowest Akaike Information Criterion value is Intercept + BLUEBOOK_1000 + MVR_PTS + TRAVTIME.

Question 20 (5 points)

What is the Bayesian Information Criterion value of the final logistic model?

The Bayesian Information Criterion of the final logistic model is 732.5683. Accepted interval is (732.5682, 732.5684).