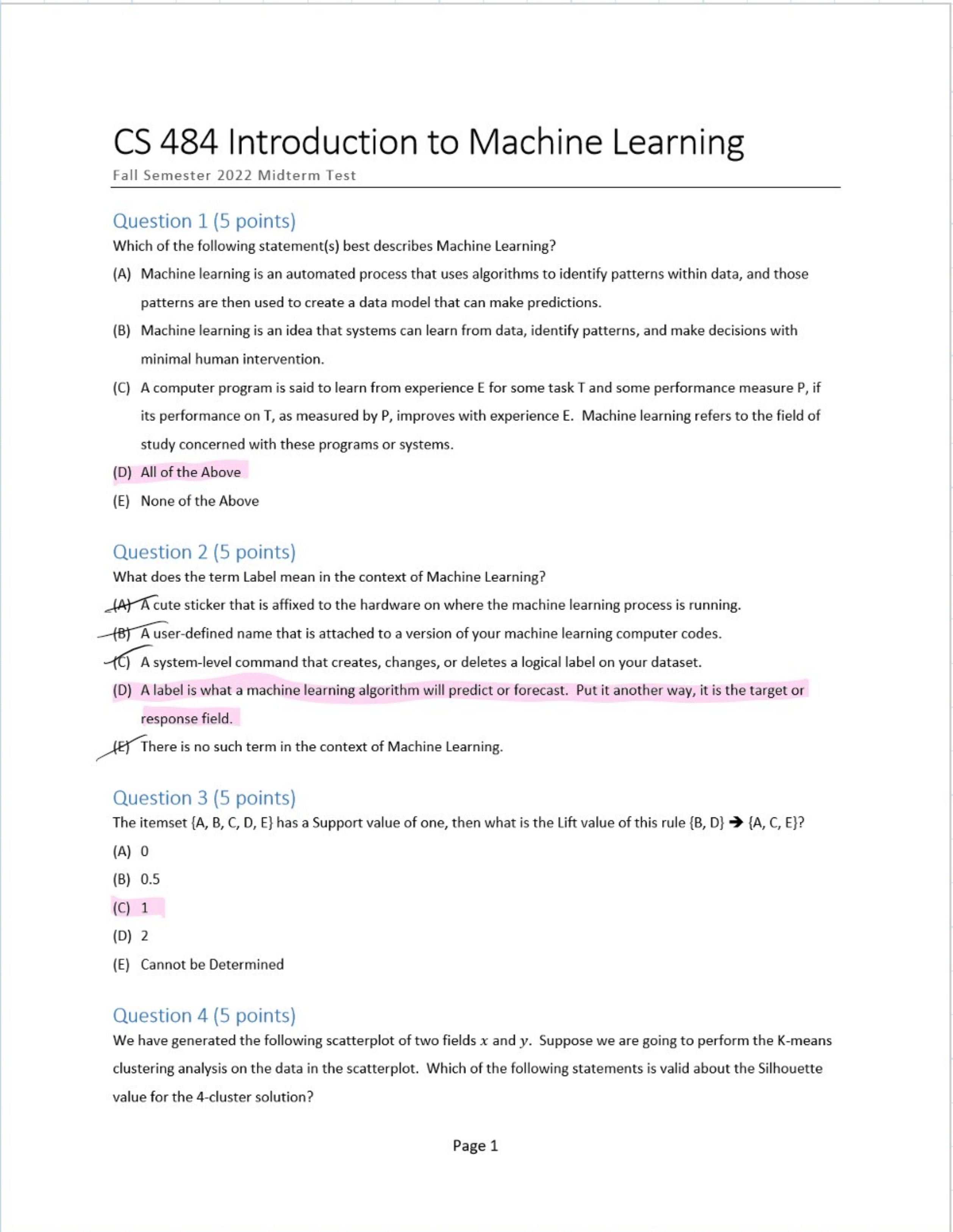
midterm

Wednesday, October 19, 2022 9:26 PM



CS 484 Introduction to Machine Learning

Fall Semester 2022 Midterm Test

Question 1 (5 points)

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

(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

Question 2 (5 points)

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

~~(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.

Question 3 (5 points)

The itemset {A, B, C, D, E} has a Support value of one, then what is the Lift value of this rule {B, D} → {A, C, E}?

(A) 0

(B) 0.5

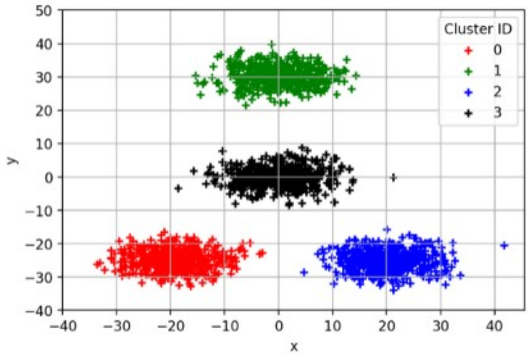
(C) 1

(D) 2

(E) Cannot be Determined

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 ⇐ all values close to center of cluster

~~(D) Close to four~~

~~(E) Cannot be determined~~

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 $= {}_{100}C_7$

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

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

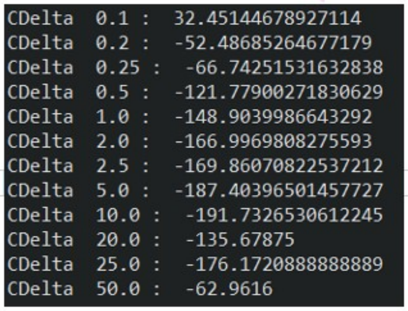
$$1-\left(\frac{1}{4}^2+\frac{1}{4}^2+\frac{1}{4}^2+\frac{1}{4}\right)=.75$$

$$1-\left(\frac{1}{4}^2+\frac{3}{4}^2\right)=.375$$

$$1-\left(\frac{2}{4}^2+\frac{2}{4}^2\right)=.5$$

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?



```
CDelta  0.1 :  32.45144678927114
CDelta  0.2 :  -52.48685264677179
CDelta  0.25 :  -66.74251531632838
CDelta  0.5 :  -121.77900271830629
CDelta  1.0 :  -148.9039986643292
CDelta  2.0 :  -166.9969808275593
CDelta  2.5 :  -169.86070822537212
CDelta  5.0 :  -187.40396501457727
CDelta  10.0 :  -191.7326530612245
CDelta  20.0 :  -135.67875
CDelta  25.0 :  -176.1720888888889
CDelta  50.0 :  -62.9616
```

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

ℓ (chz, soda) → (wine) = C((chz, soda) ∪ (wine)

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

$$L(chz, soda) \to (wings) = \frac{C(chz, soda) \to (wings)}{S(wings)}$$

$$C(chz, soda) \to (wings) = \frac{S(chz, wings, soda)}{S(chz, soda)}$$

$$S(chz, wings, soda) = \frac{4}{10} = \frac{2}{5} \quad S(chz, soda) = \frac{5}{10} = \frac{1}{2}$$

$$C(chz, soda) \to (wings) = \frac{2/5}{1/2} = \frac{2}{5} \times \frac{2}{1} = \frac{4}{5}$$

$$S(wings) = \frac{7}{10}$$

$$L(chz, soda) \to (wings) = \frac{4/5}{7/10} = \frac{4}{\overset{1}{\cancel{5}}} \times \frac{\overset{2}{\cancel{10}}}{7} = \frac{8}{7} \approx 1.14286$$

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}. $1.14286 \approx 1.14$

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

$$.24010^2 + .41159^2 + .26460^2 + .07560^2 + .00811$$
$$= .69715 \approx .70$$

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

$223/11324 = 0.19693 \quad \ln(0.19693) = -3.92749 \approx -3.93$

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.

Chebyshev = $\max(|x_1 - x_2|)$

Cluster	X1	X2	X3	X4	X5	
0	6.34 3.36	6.82 4.18	7.21 4.19	7.18 .62	7.47 .97	max = 4.19
1	8.04 1.66	8.56 2.14	9.42 1.98	8.08 .28	7.70 1.2	max = 2.14

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

9.7 10.7 11.4 7.8 6.5

Question 11 (5 points)

Which cluster should we assign this new observation to?

distance of observation to C_1 is smaller than distance to C_0

Question 12 (5 points)

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

2.14

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	.3052
II	74	185	160	55	16	.098
III	33	228	623	755	363	.4004
IV	90	290	349	213	40	.1964
	.0524	.2014	.3324	.302	.1118	

Entropy = $-\sum(p \log_2 p)$

root node entropy = 2.09179

Question 13 (5 points)

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

(A) {I} + {II, III, IV} root − .88755 = 1.20424

(B) {I, II} + {III, IV} root − .97279 = 1.11900

(C) {I, II, III} + {IV} root − .71467 = 1.37712

(D) None of the Above

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?

(A) {I} + {II, III} root − 1.02326 = 1.06853

(B) {I, II} + {III} root − 1.05709 = 1.0347

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

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

(E) None of the Above

Question 15 (5 points)

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

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?

0.63

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.

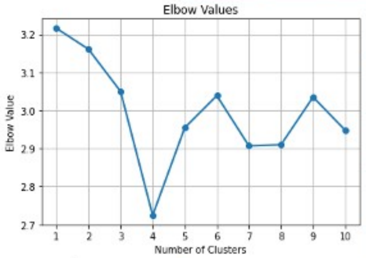
You are provided with this CSV file **ChicagoCompletedPotHole.csv** for analyzing the city's efforts to fill up street potholes. The data contains 17,912 observations. Each observation represents a completed request which was created between December 1, 2017 and March 31, 2018 and was completed between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

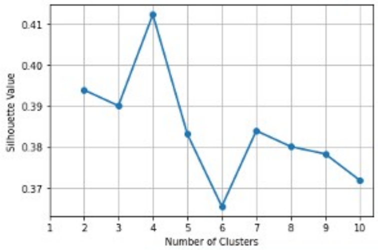
Name	Level	Description
1) CASE_SEQUENCE	Nominal	A unique index for identifying an observation
2) WARD	Nominal	Chicago's ward number from 1 to 50
3) CREATION_MONTH	Nominal	Calendar month when the request was created
4) N_POTHOLES_FILLED_ON_BLOCK	Interval	Number of potholes filled on the city block
5)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

Name	Level	Description
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.

4

Question 18 (5 points)

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

16458

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

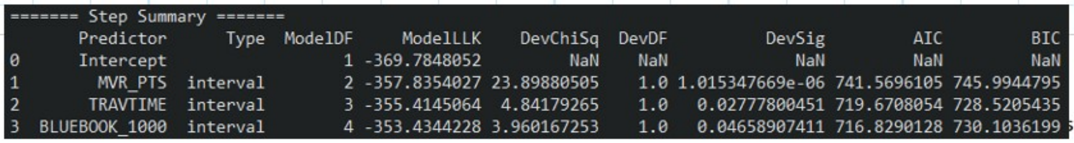
- CREDIT_SCORE_BAND: Credit Score Tier ('450 – 619', '620 – 659', '660 – 749', and '750 +')

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



```
======= Step Summary =======
      Predictor      Type  ModelDF     ModelLLK     DevChiSq  DevDF          DevSig          AIC          BIC
0     Intercept                  1 -369.7848052          NaN    NaN             NaN          NaN          NaN
1       MVR_PTS  interval        2 -357.8354027  23.89880505    1.0 1.015347669e-06 741.5696105 745.9944795
2      TRAVTIME  interval        3 -355.4145064   4.84179265    1.0   0.02777800451 719.6708054 728.5205435
3  BLUEBOOK_1000  interval       4 -353.4344228  3.960167253    1.0   0.04658907411 716.8290128 730.1036199
```

- Use the All Possible Subset method to search for the optimal model.

Question 19 (5 points)

- 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?

MVR_PTS, TRAVTIME, BLUEBOOK_1000

Question 20 (5 points)

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

730.10362