CS 484 Introduction to Machine Learning

Autumn 2020 Final Examination

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

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 computer 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)

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) 3.1416
(E) Cannot be Determined

Question 4 (5 points)

I calculated the Elbow values and the Silhouette values for my 1-cluster to 10-cluster solutions. Based on these values, what do you suggest for the number of clusters? An integer answer is expected.

Number of Clusters	Elbow Value	Silhouette
1	579857.9543	N/A
2	532455.2722	0.5391
3	493218.0813	0.5300
4	433215.8150	0.5479
5	430290.4574	0.5411
6	412804.9312	0.5140
7	409729.7423	0.5172
8	404285.7518	0.5081
9	378087.1355	0.5056
10	369686.6227	0.4984

Question 5 (5 points)

The CSV file FinalQ5.csv contains 500 observations. It has three columns.

1. TransportMode: mode of transportation, a categorical predictor that has four levels: *Bike*, *Drive*, *Public*, and *Walk*
2. CommuteMile: number of miles for commuting, an interval predictor
3. Late4Work: late for work indicator, a binary target that has two levels: *No* and *Yes*

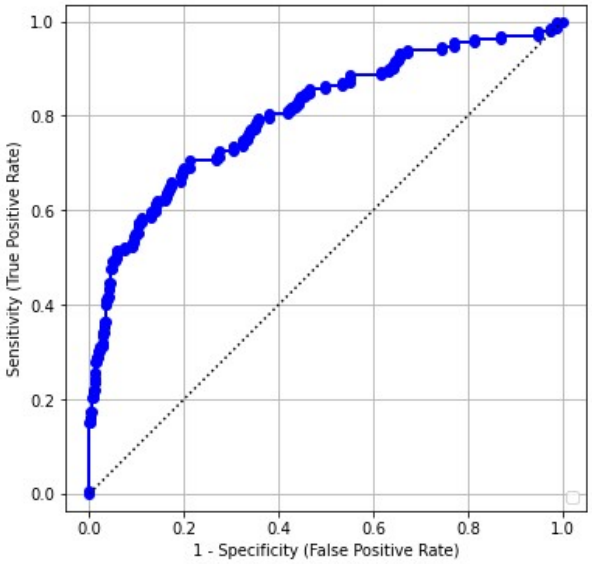
You will build a main effect logistic regression model that includes the Intercept term. Based on the model results, what phenomenon can you conclude the model is having?

Multiple Choice:

(A) No Special Phenomenon
(B) Complete Separation
(C) Quasi-Complete Separation in One Combination of TransportMode and CommuteMile
(D) Quasi-Complete Separation in Two Combinations of TransportMode and CommuteMile
(E) Quasi-Complete Separation in Three Combinations of TransportMode and CommuteMile

Question 6 (5 points)

We ran a marketing campaign to promote a product. We offered the product to 1,431 persons and 132 persons responded positively to the product. We trained a binary logistic regression model on the marketing campaign data. The following Receiver Operating Characteristics curve measures the performance of this logistic regression model. Which of the following values might be the Area Under Curve metric?



Multiple Choice:

(A) 0.3
(B) 0.5
(C) 0.6
(D) 0.8
(E) 0.95

Question 7 (5 points)

You will calculate the Cramer's V statistic to measure the association between two categorical features, namely, *Row* and *Column*. Instead of the original training data, you are given the following crosstabulation table.

Number of Observations		Column			
		1	2	3	4
Row	A	4,340	5,403	2,456	353
	B	8,095	16,156	10,798	2,371
	C	4,761	14,154	14,103	4,597
	D	813	3,636	5,307	2,657

Question 8 (5 points)

You live in the San Francisco Bay area where earthquakes are not uncommon. Your house has a security alarm system against burglary, and it can be set off occasionally by an earthquake. Historically, there is a 6% chance that your house will be burglarized and there is a 2% chance that an earthquake will occur in your area. You can assume that the occurrences of burglary and earthquake are statistically independent. Based on your experience, your alarm will sound if the following events have occurred.

Earthquake	True	True	False	False
Burglary	True	False	True	False
Probability that the Alarm will sound	0.99	0.15	0.95	0.0001

Please calculate this quantity Prob(Burglary = True and Earthquake = False | Alarm Sounded = True), i.e., the conditional probability that your house has been burglarized but no earthquake has occurred provided the alarm has been sounded.

Question 9 (5 points)

Please analyze the following eighteen observations using the Nearest Neighbors algorithm. The distance measure is the Euclidean distance and the number of neighbors is three.

ID	x1	x2	y
1	0	1	0.2
2	1	1	3.9
3	-1	1	-4.5
4	0	1	2.0
5	1	1	20.8
6	1	2	7.3
7	-1	2	-12.5
8	1	2	-5.1
9	0	1	1.1

ID	x1	x2	Y
10	0	2	13.8
11	-1	1	7.5
12	0	2	23.2
13	1	1	-7.5
14	-1	1	10.7
15	1	2	-2.7
16	-1	2	-3.2
17	0	2	1.4
18	-1	2	0.3

Based on the MEDIAN of the y field of the three nearest neighbors, what is the predicted y of this observation x1 = 1 and x2 = 2?

Question 10 (5 points)

The following table shows the observed target values and the predicted event probabilities from a model. The target is a binary variable whose values are Event and Non-Event. Please determine the threshold value that yields the highest F1 score. Please give your exact answer.

You should refrain from using the sklearn.metrics.precision_recall_curve function because it does not return all the possible thresholds.

Observed Target Value	Non-Event	Non-Event	Event	Event	Event	Non-Event	Non-Event	Event	Event	Non-Event
Predicted Event Probability	0.2	0.3	0.45	0.5	0.55	0.4	0.45	0.7	0.7	0.1

Question 11 to 13

Please use the following information for answering Questions 11, 12, and 13.

We are interested in studying the effects of the Vehicle Age on the Claim Indicator. We particularly want to optimally separate the Vehicle Age into two groups.

You are given a two-way table and are asked to build a decision tree model using the Entropy criterion to discover the groups. We will treat the Vehicle Age as an **ordinal** predictor and the Claim Indicator as a **nominal** target variable. The order of the Vehicle Age is '1 to 3' < '4 to 7' < '8 to 10' < '11 and Above'.

Vehicle Age (Number of Years)	Claim Indicator	
	No	Yes
1 to 3	1,731	846
4 to 7	1,246	490
8 to 10	1,412	543
11 and Above	2,700	690

Question 11 (5 points)

What is the Entropy value of the root node?

Question 12 (5 points)

Which is the most optimal way to form the two groups?

Multiple Choice:

(A) {1 to 3} + {4 to 7, 8 to 10, 11 and Above}
(B) {1 to 3, 4 to 7} + {8 to 10, 11 and Above}
(C) {1 to 3, 4 to 7, 8 to 10} + {11 and Above}
(D) {1 to 3, 8 to 10} + {4 to 7, 11 and Above}
(E) {4 to 7, 8 to 10} + {1 to 3, 11 and Above}

Question 13 (5 points)

What is the reduction in entropy of your most optimal split?

Question 14 to 19

Please use the following information for answering Questions 14 to 19.

The Center for Machine Learning and Intelligent Systems at the University of California, Irvine manages the Machine Learning Repository (https://archive.ics.uci.edu/ml/index.php). You will train three models using the training CSV file WineQuality_Train.csv. Then, you will select the model that yields the lowest misclassification rate on the test CSV file WineQuality_Test.csv.

Our target variable is quality_grp. It has two distinct values: 0 and 1.

Our input features are alcohol, citric_acid, free_sulfur_dioxide, residual_sugar, and sulphates. These five features are considered interval variables.

If the model algorithm can return predicted probabilities, then we will classify an observation to quality_grp = 1 if the Prob(quality_grp = 1) $\geq$ 0.1961733010776.

Question 14 (5 points)

Build a Multinomial Logistic model with the following specifications.

- Include the Intercept term
- Use the Newton-Raphson optimization method
- Set the relative error in parameters acceptable for convergence to 0.00000001

- The algorithm must converge

What is the misclassification rate on the test data?

Question 15 (5 points)

Build a Support Vector Machine classifier model with the following specifications.

- Specify the kernel to be linear
- Set the random state to 20201202
- Do not set a limit on the maximum number of iterations
- Disable probability estimates

What is the misclassification rate on the test data?

Question 16 (5 points)

Build a Multi-layer Perceptron classifier model with the following specifications.

- Try the number of hidden layers from 1 to 10 by an increment of 1
- Try the number of hidden neurons per layer from 5 to 10 by an increment of 1
- Use the Rectified Linear Unit activation function for the hidden layers
- Use the lbfgs optimizing solver
- Specify the initial learning rate to 0.1
- Set the maximum number of iterations to 5000
- Set the random state value to 20201202

You will choose the configuration (i.e., number of layers and number of neurons) that yields the smallest loss on the training data. What is the misclassification rate on the test data?

Question 17 (5 points)

Among the three models, multinomial logistic, Support Vector Machine classifier, and Multi-layer Perceptron classifier, which model gives the lowest misclassification rate on the test data?

Multiple Choice:

(A) Multinomial Logistic
(B) Support Vector Machine classifier
(C) Multi-layer Perceptron classifier
(D) Two-way Tie Multinomial Logistic and Multi-layer Perceptron classifier
(E) Three-way Tie

Question 18 (5 points)

Suppose you apply the Bagging technique on the multinomial logistic model with the specifications in Question 14. You tried the number of Bagging steps from 1 to 200 using your lucky random seed. After you have performed each number of Bagging steps, you calculate the misclassification rate on the test data. At the end of this experiment, you plot the misclassification rate on the test data against the number of Bagging steps. What will you expect from the graph?

Multiple Choice:

(A) An upward trend going to 0.5
(B) A downward trend going to 0.175
(C) A trend converging to approximately 0.245
(D) A trend converging to approximately 0.275
(E) Inconclusive

Question 19 (5 points)

You trained a classification tree on the Training data with the following specifications.

- The Splitting Criterion is the Entropy
- The maximum depth is 5
- The random state value is None

Using the threshold of 0.5 on the predicted probability Prob(quality_grp = 1), the classification tree yields an accuracy of 0.81589744 on the Test data.

Next, you are going to apply the Adaptive Boosting technique on the above classification tree model with the following specifications.

- The maximum number of Boosting iterations is 50
- Terminate the iteration if the accuracy of the classification tree on the Training data is greater than or equal to 0.9999999
- If the observed quality_grp is 1, then the absolute error is 1 – Prob(quality_grp = 1). Otherwise, the absolute error is Prob(quality_grp = 1).
- If an observation is correctly classified, then the weight is the absolute error. Otherwise, the weight is 2 plus the absolute error.

What could be the accuracy of the boosted classification tree on the test data?

Multiple Choice:

(A) 0.78
(B) 0.83
(C) 0.24
(D) 0.99
(E) 0.55

Question 20 (5 points)

If the Area Under Curve metric is one for a binary classification problem, then what can you conclude that the value of the Root Average Square Error metric?

Multiple Choice:

(A) Equal to 0
(B) Equal to 0.5
(C) Greater than 0.5

(D) Less than 0.5
(E) Inconclusive