

MSCI 623 Test #2

Available: 9am EDT, Tuesday, June 15, 2021

Due: 9am EDT, Wednesday, June 16, 2021

Instructions: There are 9 questions for a total of 37 points. This test must be done individually, with no collaboration allowed. Everyone taking this test is expected to act with the highest integrity and not cheat on this test in any way, shape, or form. Suspected violations will be reported to the Associate Dean.

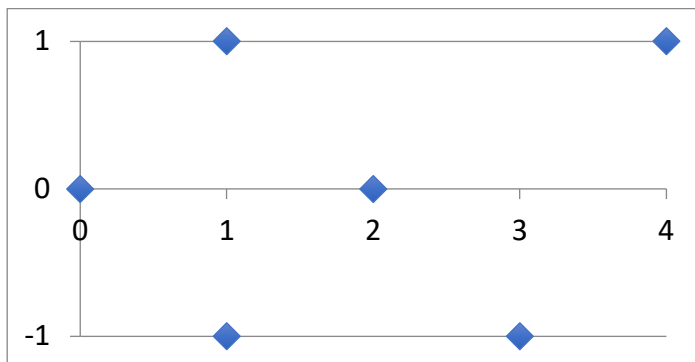
Please show all your work. Correct answers with no work shown will receive a grade of zero.

Allowed aids: course textbook; course notes and other course material posted on Learn; external material linked to in the lecture slides.

Please submit your solutions, in pdf format, to Learn, under Submit -> Dropbox -> Test 2.

Question 1: Linear models [4 points]

Suppose we have the following (x,y) datapoints: (0,-1), (1,0), (1,2), (2,3), (3,4), (4,8). Suppose we are given a linear model of the form $y=mx+b$ that uses x to predict y . Suppose that running the model on the above data points gives the residual plot shown below (with x on the horizontal axis and residuals on the vertical axis). Without making any other assumptions, what is the linear model?



Question 2: Logistic regression [3 points]

Suppose we build a logistic regression model to classify patients into HEALTHY or NOT-HEALTHY based on their age in years above sixty (call it p ; for example, for a 62-year-old person, years above sixty = 2), fitness (call it q ; zero if the patient exercises regularly, one otherwise), and cholesterol level (call it r). Suppose the model is

$$z = \frac{1}{1 + e^{1 - 2p + q - 0.5r}}$$

Will this model make the same or different prediction for the following patients?

Patient 1: 60-year-old person who exercises regularly, with cholesterol level 3.0

Patient 2: 60-year-old person who does not exercise regularly, with cholesterol level 3.5

Question 3: Naïve Bayes Classifier [4 points]

Suppose we have the following data set in which A and B are features and C is the class label.

A	B	C
No	30	No
No	35	No
No	60	Yes
Yes	52	Yes
No	50	Yes
Yes	40	Yes
Yes	48	Yes
No	45	No
Yes	32	No

For a new datapoint with A = Yes and B = 45, would Naïve Bayes predict C = Yes or C = No?

Question 4: Decision Trees [4 points]

Consider the following dataset in which F, G and H are features and C is the class label. Explain which root node would be chosen by the entropy-based decision tree classifier described in the lecture slides.

F	G	H	C
Maybe	Yes	Yes	Yes
Yes	Yes	No	No
No	Yes	No	No
Yes	Yes	Yes	Yes
Maybe	Yes	No	No
Yes	Yes	Yes	Yes
Maybe	Yes	Yes	No
No	Yes	No	No
Maybe	Yes	Yes	Yes

Question 5: Cross-validation [4 points]

Suppose you are the manager of a data analytics company. One of your junior analysts has applied a decision tree classification algorithm to a labeled data set D and calculated two things:

- 1) the accuracy of the decision tree on D using 10-fold cross validation, and
- 2) the accuracy of the decision tree on D using an 80%-train / 20%-test split of the data

She says to you: “With 10-fold cross validation, we get a 12% improvement in accuracy compared to the 80/20 split. I think that for classifying new data, we should use 10-fold cross validation”. You realize that the analyst found something interesting, but she is also lacking a fundamental understanding of data mining. What is incorrect in the analyst’s statement that “I think that for classifying new data, we should use 10-fold cross validation”?

Question 6: Leave-one-out cross-validation [4 points]

Consider a Boolean classification problem with a yes/no class label (and some features). Consider a classifier called VERY_SIMPLE that works as follows. Let C be the class label that occurs more frequently in the training dataset (if both yes and no are equally frequent, let C = "yes" to break the tie). When given a new datapoint to classify, VERY_SIMPLE will always output C, regardless of the values of the features of the new datapoint.

Suppose we have a dataset with exactly one half of the records with class "yes" and exactly one half with class "no". Compute the accuracy of VERY_SIMPLE on this dataset using leave-one-out cross validation.

Question 7: The PRISM Rule-Based Classifier [6 points]

Consider the following dataset in which F, G and H are features and C is the class label. Give the output of the PRISM algorithm on this data set.

F	G	H	C
No	Yes	Yes	Yes
No	No	Yes	No
Maybe	Yes	Yes	Yes
No	Yes	Yes	Yes
Yes	Yes	No	No
Maybe	Yes	No	No
Maybe	Yes	Yes	Yes
Yes	Yes	No	No
No	Yes	No	No

Question 8: Confusion Matrix Analysis [4 points]

Suppose you are the manager of the loan department of a bank. Your data analyst examines previous loan repayments, and, based on some training dataset, suggests two classification algorithms to determine whether a new customer should get a loan. Suppose the confusion matrices for these two algorithms are as follows:

ALGORITHM 1	Predicted paid_loan = yes	Predicted paid_loan = no
Actual paid_loan = yes	295	5
Actual paid_loan = no	95	605

ALGORITHM 2	Predicted paid_loan = yes	Predicted paid_loan = no
Actual paid_loan = yes	205	95
Actual paid_loan = no	5	695

Under what business circumstances would it make sense to use Algorithm 2 instead of Algorithm 1?

Question 9: Overfitting [4 points]

Suppose we have a table T with a primary key k and a Boolean column c. Suppose we learn a model from T that uses k to predict c. Explain why such a model is likely to overfit the data.