CS390DM0 Spring 2017 Midterm 1 *solutions*, 21 February, 2017
Prof. Chris Clifton

I would expect an A student to score at least 36, a B student at least 30, and a C student at least 19.

Turn Off Your Cell Phone. Use of any electronic device during the test is prohibited. As previously noted, you are allowed notes: Up to two sheets of 8.5x11 or A4 paper, single-sided (or one sheet double-sided).

Time will be tight. If you spend more than the recommended time on any question, **go on to the next one**. If you can't answer it in the recommended time, you are either giving too much detail or the question is material you don't know well. You can skip one or two parts and still demonstrate what I believe to be an A-level understanding of the material.

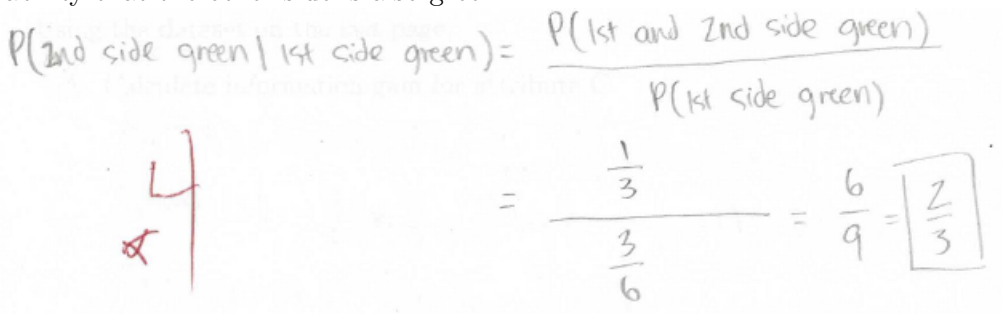
Note: It is okay to abbreviate in your answers, as long as the abbreviations are unambiguous and reasonably obvious.

In all cases, it is important that you give some idea of how you derived the answer, not simply give an answer. Setting up the derivation correctly, even if you don't carry out the calculations to get the final answer, is good for nearly full credit.

The last page of the exam contains an example database that you will use for several of the questions. You may tear it off for easy reference.

1 Probability (5 minutes, 4 points)

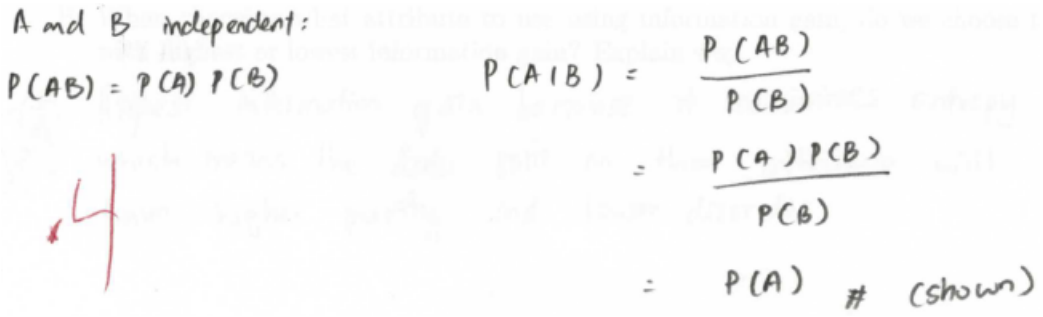
There are three cards. The first is green on both sides, the second is red on both sides, and the third is green on one side and red on the other. Consider the scenario where a card is chosen at random and one side is shown (also chosen at random). If the side shown is green, what is the probability that the other side is also green?



Courtesy of Brian Lee. *1 for setting it up as probability, 1 for noting conditional on first draw, 1 for noting that which card is the first draw is conditional on seeing green, 1 for correct*

2 Conditional Probability (5 minutes, 4 points)

Prove that if A and B are independent events, then $P(A|B) = P(A)$.

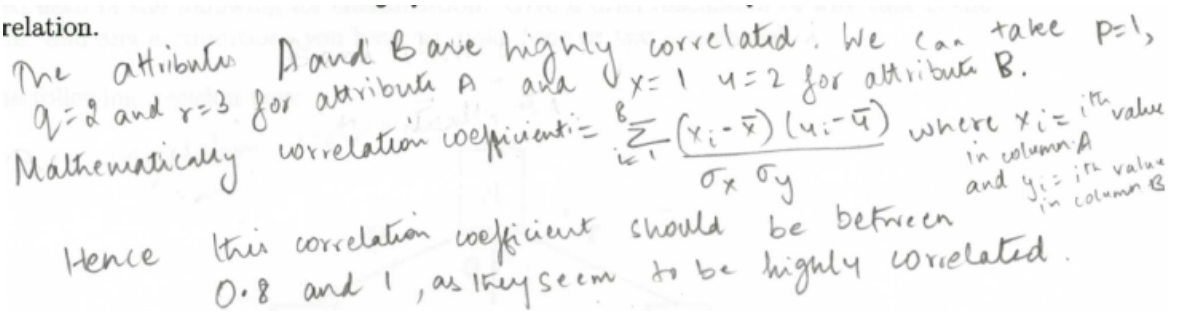


Courtesy of a student who prefers to be anonymous. *1 for setting up as proof, 1 for definition of independent, 1 for some working it out, 1 for complete*

3 Exploratory Data Analysis (10 minutes, 6 points)

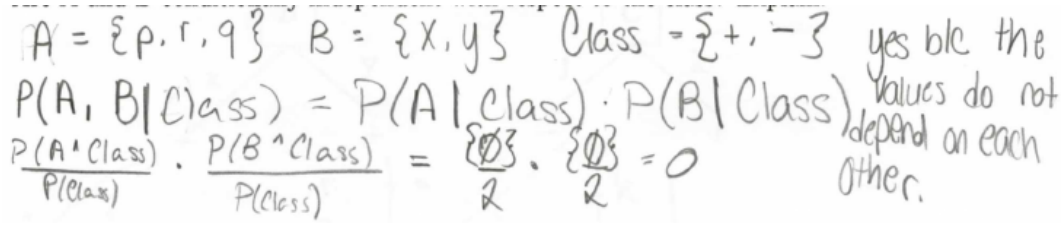
These questions all refer to the dataset on the last page.

A. Find two attributes that are highly correlated. Mathematically characterize this correlation.



Courtesy of Abhishek Shrinivasan. *1 for picking two that are correlated, 1 for a definition of correlation, 1 for working it out*

B. Are A and B conditionally independent with respect to the class? Explain.

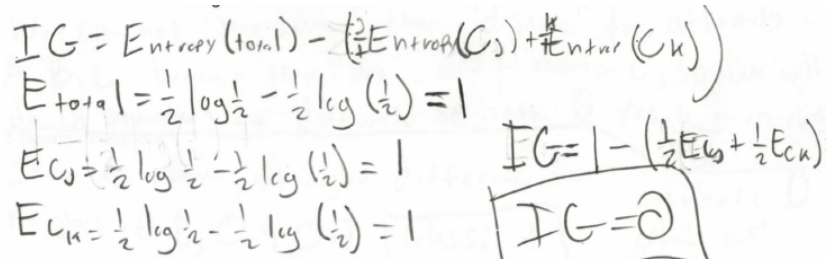


Courtesy of Simone May. *1 for showing knowledge of conditional independence, 1 for correct answer, 1 for good explanation*

4 Information Gain (8 minutes, 5 points)

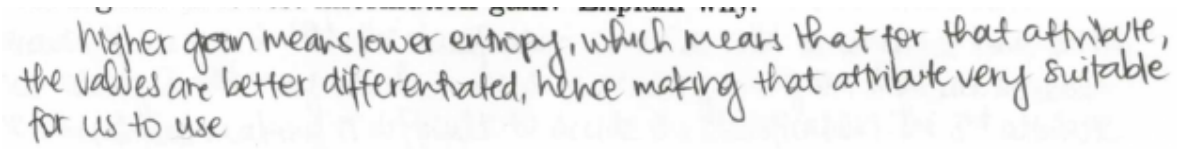
Using the dataset on the last page,

A. Calculate information gain for attribute C



Courtesy of a student who prefers to be anonymous. *1 for setting it up properly, 1 for working it out*

B. When choosing what attribute to use using information gain, do we choose the one with highest or lowest information gain? Explain why.



Courtesy of Marshia Seto. *1 for "high", 1 for good explanation*

1 for showing some understanding of IG, can come from either part.

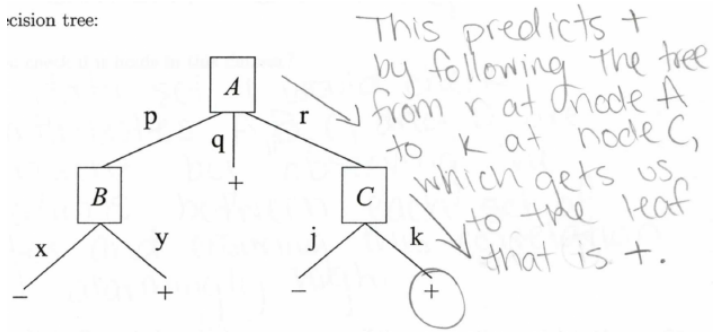
5 Classification (12 minutes, 8 points)

Given the dataset on the last page, what would we predict as the class of the instance:

r x k 0.8

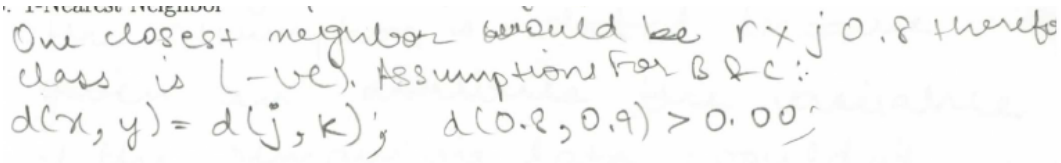
if we used each of the following for classification. Give a brief discussion of why that is the prediction, and any assumptions you have to make (one or two sentences).

A. The following decision tree:



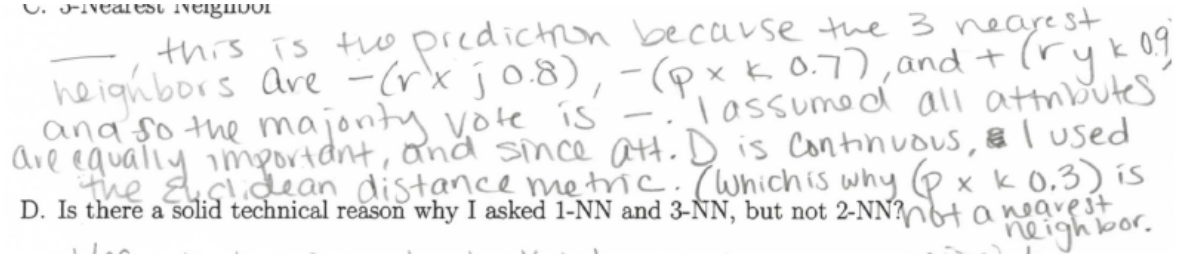
Courtesy of a student who prefers to be anonymous. *1 for correct answer, 1 for path*

B. 1-Nearest Neighbor



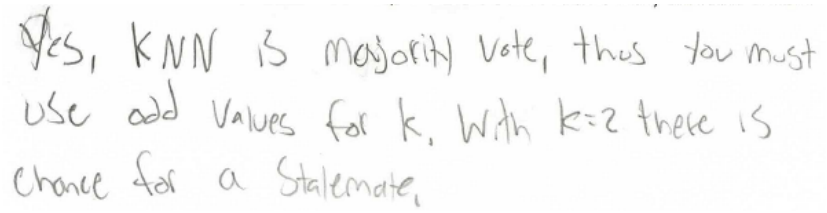
Courtesy of a student who prefers to be anonymous. *1 for correct answer, 1 for noting neighbor*

C. 3-Nearest Neighbor



Courtesy of Robyn Berkel. *1 for correct answer, 1 for noting neighbor 1 point for discussion of distance function in either part*

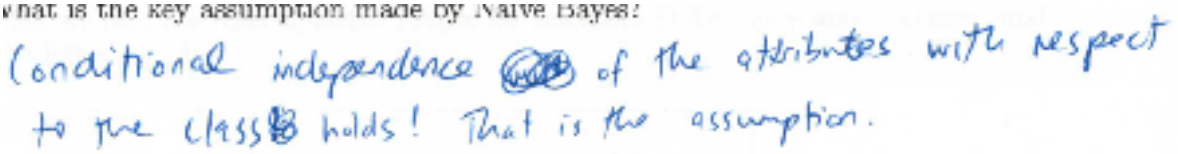
D. Is there a solid technical reason why I asked 1-NN and 3-NN, but not 2-NN?



Courtesy of a student who prefers to be anonymous. *1 for noting voting, need to break ties*

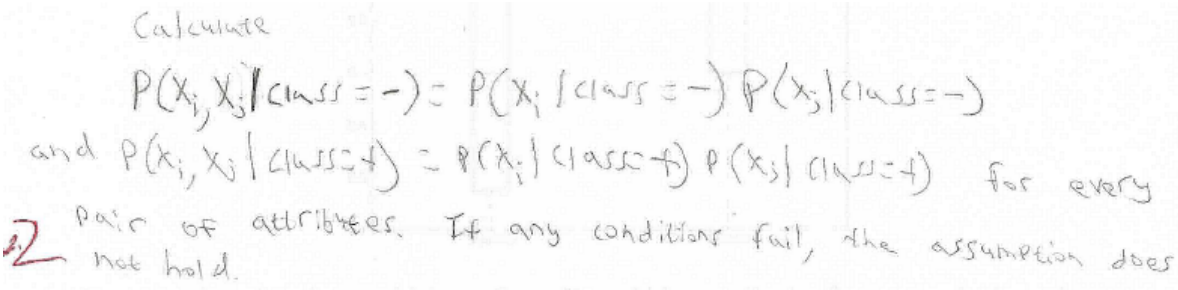
6 Naïve Bayes (8 minutes, 5 points)

A. What is the key assumption made by Naïve Bayes?



Courtesy of Avik Mikhija. *1 for conditionally independent*

B. How would you check if it holds in this dataset?



Courtesy of Caleb Beth. *1 for description. 1 for showing understanding of what CI means for NB*

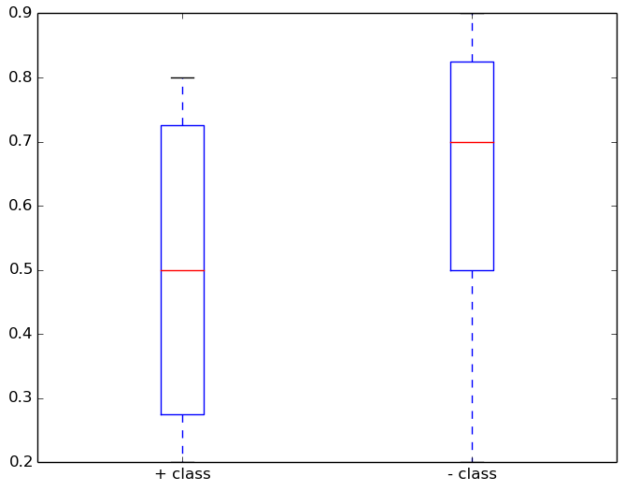
C. Is it possible for Naïve Bayes to have high accuracy even if this assumption is violated? Explain why or why not.

Yes. If the attributes are not mutually independent but are positively correlated, Naive Bayes still has robust performance most of the time.

Courtesy of a student who prefers to be anonymous. *1 for yes, 1 for either example or reasonable discussion of why*

7 Overlap of IQRs (5 minutes, 3 points)

Suppose that we plot the Inter-Quartile ranges for attribute D for the + and − classes, and we have the following plot:



Based on this plot, do you think attribute D would be an effective feature for distinguishing between the + and − class? Explain why or why not.

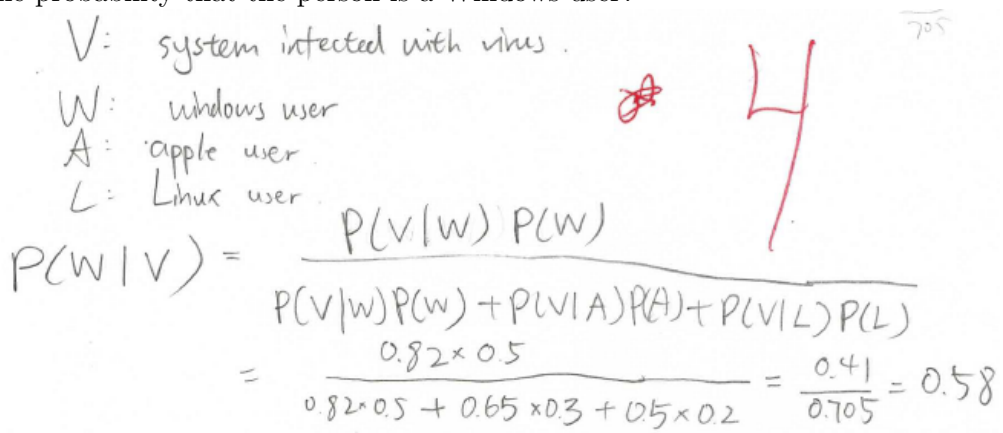
I don't believe D is a good choice since the entirety of the + class' third quartile is contained in the - class' IQR and the second quartile of the - class is contained in the + class' IQR. They overlap far to much to be effective.

Courtesy of Ian Spooner. *1 for "no", 1 for some reasonable discussion, 1 for noting overlap of distributions not working well*

8 Probability (10 minutes, 4 points)

Suppose that 30 percent of computer owners use an Apple machine, 50 percent use a Windows machine, and 20 percent use Linux. Suppose that 65 percent of Apple users have succumbed to a computer virus, 82 percent of Windows users get the virus, and 50 percent of Linux users get the

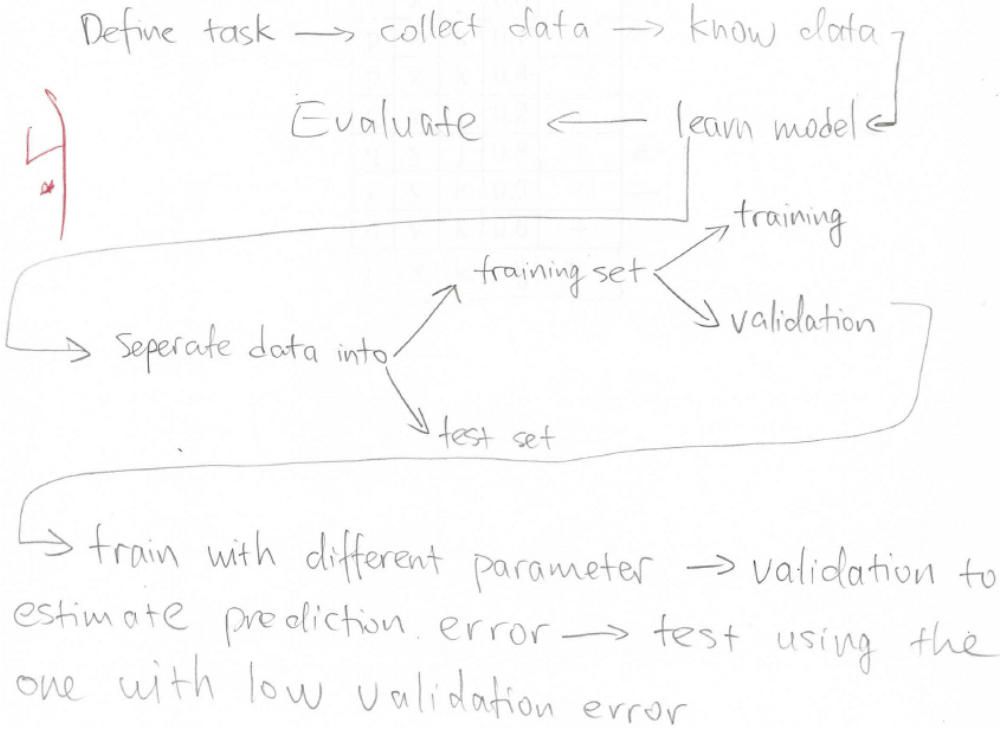
virus. We select a person at random and learn that their system was infected with the virus. What is the probability that the person is a Windows user?



Courtesy of Yixuan Ding. *1 for setting up problem, 1 for noting that having virus changes things, 1 for correct setup, 1 for answer*

9 Data Mining process (10 minutes, 4 points)

Given the dataset on the last page, I would like you to build a classifier. Discuss the process you would follow. This should be independent of the type of classifier. You want to build as accurate a classifier as possible. Also explain how you would evaluate the accuracy of the resulting classifier.



Courtesy of a student who prefers to be anonymous.

There were many points that could have been covered in answering this question, and no one answer covered everything that could have been included. However, the answer above covers several of the most important points in a particularly clear and succinct manner.

1 for division of existing data into test and training (not enough to say "I'll get more test data"

- how do you know you can do so?), 1 for discussion of computing accuracy. In addition, a point was given for each of several other relevant things, including discussion of steps, cross-validation (which counts as discussing splitting into training and test data), data exploration, model selection (discussing how or why you pick a particular model, not just picking one), data cleaning/handling missing values, or giving a formula for any of this. To get full credit, you had to mention both test/training division and discuss computing accuracy, and you couldn't state anything that was actually incorrect. E.g., a common error was saying that you would use the test set for both model selection and reporting accuracy; these would need to be two different sets (validation and test).

Dataset to be used for questions 3, 4, 5, 6, and 9.

A	*B*	*C*	*D*	*class*
p	x	j	0.2	−
r	x	j	0.8	−
p	x	k	0.7	−
p	x	k	0.3	−
q	y	j	0.2	+
q	y	j	0.8	+
r	y	k	0.9	+
q	y	k	0.6	+