CS57300 Spring 2022 Midterm *solutions*, February 24, 2022
Prof. Chris Clifton

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).

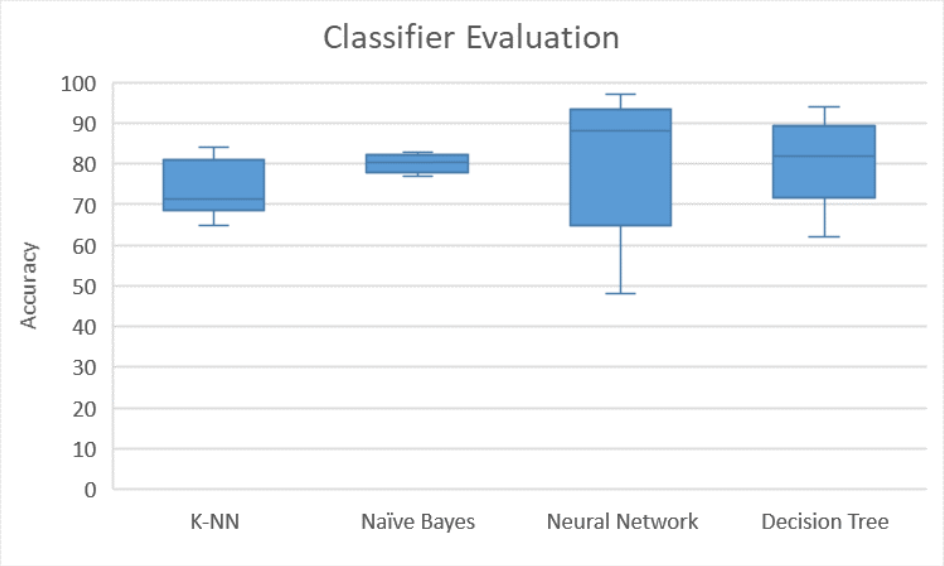
Time will be tight. If you spend more than the recommended time on any question, **go on to the next one**. If you can't answer it in the recommended time, you are either giving too much detail or the question is material you don't know well. You can skip one or two parts and still demonstrate what I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

1 Interpreting visualization (9 minutes, 7 points)

A. Given the following chart, if you were to build a classifier using one of the four techniques, which would you expect to give the highest accuracy? Explain.



Expected accuracy would be the mean, which isn't quite shown for this graph. The high median suggests that Neural Networks would have the highest mean, but it could possibly be Naive Bayes.

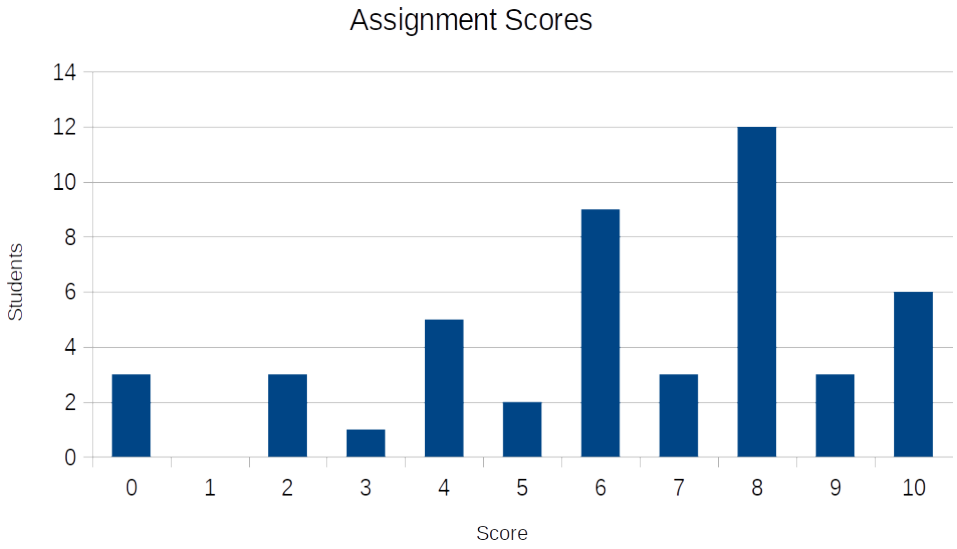
Scoring: 1 NN, 1 correct explanation

B. From the above chart, if you were asked to build a classifier that was at least 75% accurate, which technique would you choose? Explain.

Naive Bayes is the only one that shows consistently high accuracy, the others have higher variance and often (>25% of the time) have lower accuracy.

Scoring: 1 NB, 1 correct explanation

C. The following chart shows a significant variation between odd and even values. Do you think this likely reflects an interesting/true real-world situation that tells us something about how students are doing in the course? Explain.



While it might say something about how scoring is done (e.g., partial credit that gives odd scores is rare), the odd/even disparity isn't helpful in determining how students are doing.

Scoring: 1 for no with reasonable explanation.

D. Referring to the above histogram, can you suggest changes to the histogram (while still keeping it as histogram) that would make it better for giving a general idea of how students are doing in the course?

Simply widening the bins (e.g., bin size of two or four) would give a better view of how students are doing.

Scoring: 2 for grouping ranges, 1-2 for other good idea

2 Choosing a visualization (7 minutes, 6 points)

A. You are given data consisting of student test scores and undergraduate GPAs, and if they complete a Computer Science Ph.D. in six years or not. How would you visualize this data to determine if there are interesting relationships between these factors and timely completion of the Ph.D.? You can either explain (e.g., name a type of visualization), or draw a sample.

A scatterplot is appropriate for showing two continuous dimensions, and we can use color or shape of the points (e.g., + and -) to represent timely completion of the Ph.D.

Scoring: 2 scatterplot, 1 for capturing all 3 dimensions

B. Would you choose a different visualization if instead of test scores and GPA, you had undergraduate major and GPA?

Since there isn't an order to major, a scatterplot is no longer appropriate. A bar chart showing average GPA, or a box/whisker plot with the GPA range for each major, would be better - and placing the "complete in six" and "not complete in six" side-by-side in each UG major column would allow seeing how completion compared with all data.

Scoring: 2 for capturing category vs. continuous, 1 for good choice.

3 Hypothesis testing (4 minutes, 3 points)

You are given a Naive Bayes classifier that is highly accurate in predicting CS57300 class performance based on undergraduate GPA. From this, would the relationship between GPA and class performance be a descriptive, non-directional relational, directional relational, or directional causal hypothesis? Explain.

This suggest that the probability of good class performance can be estimated based on GPA, implying a correlation. But it doesn't suggest that it is causal. Simply that we know this from Naive Bayes doesn't tell us directional. Naive Bayes operates on categorical data, it could be that people with undergrad B's do well in CS57300, but with A's and C's as undergrads don't.

Scoring: 1 for showing some understanding of relationships, 1-2 for understanding what Naive Bayes being accurate means.

4 Classification: general, metrics (10 minutes, 5 points)

Consider police officers detecting drunk drivers. Assume their breathalyzer reports false drunkenness in 5% of the cases (i.e., displays the driver is drunk when the driver is sober), but never fail to detect a truly drunk person. Suppose one in a thousand drivers is driving drunk.

A. Assume the police officers stop a driver at random and the breathalyzers indicates that the driver is drunk, how high is the probability the driver really is drunk? Explain how you arrived at your answer.

$P(Drunk|Positive) = \frac{P(Positive|Drunk)*P(Drunk)}{P(Positive)}$ **.** $P(Positive)$ **is the sum of the probability of testing a Drunk person (1/1000) times the probability the test is positive (1), plus the probability of testing a sober person (1-1/1000) times the probability that reading is positive (.05), so** $\frac{1*.001}{\frac{1}{1000}*1+(1-\frac{1}{1000})*.05} = \frac{.001}{.001+.999*.05} = \frac{1}{1+999*.05} \approx \frac{1}{51}$

Scoring: 1 for probabilities, 1 for correctly understanding base rate / the difficulty of rare events, 1 for correctly setting up calculation.

B. You can think of the breathalyzer as a classifier, that classifies someone as drunk or sober. What do you think would be a good metric to determine which is the best breathalyzer: accuracy, F1 score, or something else? Explain.

The challenge is the varied base rate - accuracy doesn't work well with skewed base rates. F1 score is better. You can also create a customized score based on the relative cost of Type I and Type II errors.

Scoring: 1 for F1 scoring or something else that accommodates for base rate, 1 for showing understanding of base rate.

5 Linear Regression (9 minutes, 7 points)

Blood alcohol level is zero when you have no drinks, and it increases as you consume more alcohol. We have measured five individual's alcohol consumption (number of drinks) and blood alcohol level as below.

Number of Drinks	Blood Alcohol Level
2	3
2	4
3	8
4	12
5	17

Consider you want to model the relationship using linear regression and you came up with two representations, $y = 4x - 4$ and $y = 3x$.

A. Compute the mean squared error of the two representations.

$y = 4x - 4$

$= \frac{((4*2-4)-3)^2+0+0+0+(16-17)^2}{5}$ **= 2/5**

$y = 3x$

$= \frac{((3*2)-3)^2+2^2+1^2+0+2^2}{5}$ **= 18/5**

Scoring: 1 for understanding error, 1 for correctly calculating error in some for, 1 each for correct MSE

B. Which is better? Do you think it is optimal? Explain.

$y = 4x-4$ **has lower mean-squared error. As a linear model, it may be optimal for the data given, although a slightly increased slope could be better - we can use a linear regression to obtain an optimal linear model. That said, it is clearly unusable for smaller values, as it gives a negative blood alcohol level with 0 drinks -** $y = 3x$ **is probably better for less than two drinks, but not as good with two or more. Note that there may be better higher-order (non-linear) models. However, *no* model will have 0 error, as there are two different outcomes for 2.**

Scoring: 1 for y=4x-4, 1 for showing some idea of optimal, 1 for idea of how to prove optimal or improve.

6 Decision Trees (6 minutes, 4 points)

You developed a decision tree to predict data. The depth of the tree is more than 20, and in the training data, there are very few instances of the data at each leaf (1-3).

A. What is a likely problem with using this tree as a classifier? Explain.

Since there are few instances at a leaf, a bad sample could result in the wrong decision at that leaf - the tree is likely to overfit the data.

Scoring: 1 for overfit, 1 for why; 1-2 for other non-performance issue with good explanation.

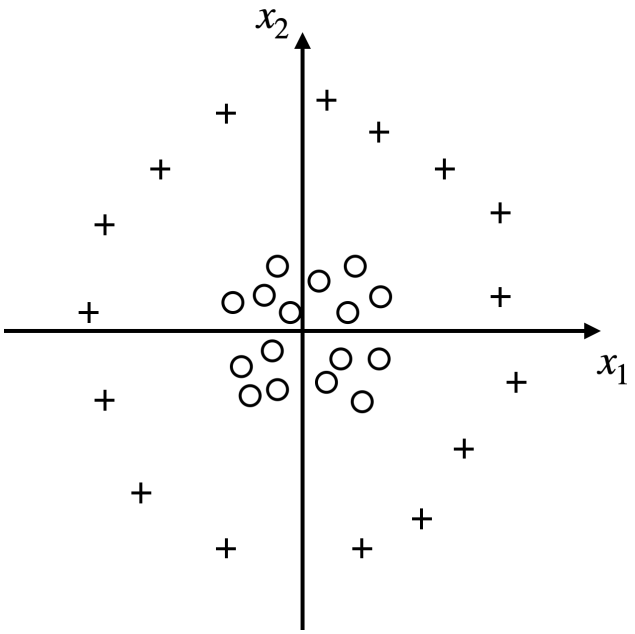
B. Suggest a method to tackle this problem and briefly describe how it works.

Pruning. This can be done in several ways; postpruning can be done by comparing accuracy at a higher node with the average accuracy at lower nodes, and where the lower nodes do not significantly improve accuracy, removing them.

Scoring: 1 for "pruning", 1 for some idea of how to do it. 1-2 for other reasonable idea that fits with decision tree learning.

7 Support Vector Machine (SVM) (8 minutes, 5 points)

Given the following data, with two dimensions as independent variables and + or **o** as the class:



A. What will happen if you use a linear SVM to separate + classes from **o** classes?

Since this data is not even remotely linearly separable, the resulting SVM classifier will give very poor results.

Scoring: 1 for not linearly separable, 1 for understanding this won't work with SVM

B. Describe how you might use a SVM as a classifier to predict if an item is + or **o**.

We can "invent" a higher dimension where the data is linearly separable. For example, $x_1 + x_2$ as a third dimension would allow us to separate with a plane. All we need is a function that gives us the distance between points in that higher dimension (the kernel function).

Scoring: 1 for kernel function, 1 for showing understanding of how kernel trick works, 1 for choosing reasonable function.

8 Naive Bayes (8 minutes, 4 points)

A Naive Bayes classifier assumes attributes are learned independently, a full Bayes classifier computes probability similarly, but accounts for all possible interactions (combinations of attributes), but otherwise estimates the probability similarly.

A. Assuming you have two binary attributes and a trinary (say, red/yellow/green) attribute, and a binary class label. How many parameters would you need to estimate?

One for each attribute value (2+2+3) for the positive class, plus the same for the negative class (although 7 or 14 were acceptable, since the probabilities are inverse and thus knowing one set gives you the other.)

Scoring: 1 for showing idea of independent, 1 for correct.

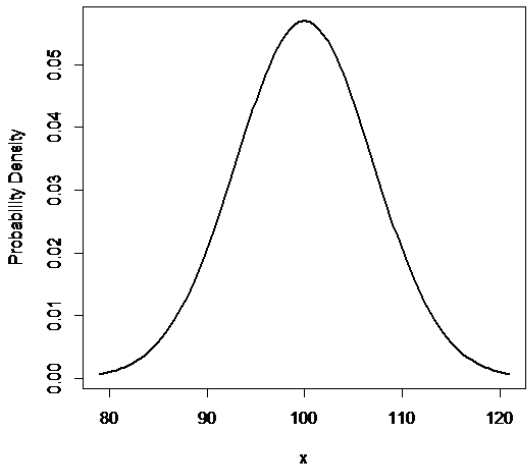
B. Would your answer be higher or lower for a full Bayes classifier? Explain. For full credit, give the number of parameters you would need to estimate.

Higher - we need to estimate the probability for each combination of attributes, not for each independently. This gives 2*2*3 for each class.

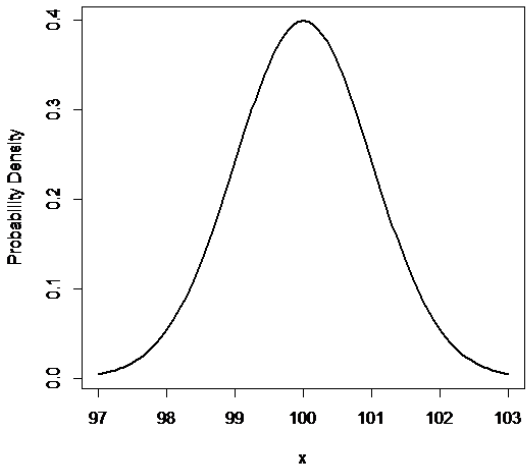
Scoring: 1 for higher with explanation, 1 for reasonably close.

9 Populations/Sampling (6 minutes, 4 points)

You are learning a classifier to address two different problems, and are given a sample of each data to learn the classifier. You are also provided with the following sampling distributions:



A



B

A. Do you think a classifier learned from sample A or sample B is more likely to overfit? Explain.

People interpreted this graph in two different ways. The intention was that this be a distribution of the mean (or some other statistic) on a sample, across many samples - which would show that sample A is less likely to match the general population and overfit to the specific population. However, some interpreted it as a PDF of a particular sample, in which case sample B, being highly concentrated around a particular value, may well overfit.

Scoring: 2 for understanding A less likely to give sample that matches general population (or B having a very tight distribution, ifyou interpreted this in the second way), 1 for how this leads to overfitting.

B. Supposed you are told the two samples come from the same underlying population. Which sample do you think is larger? Explain.

Again, if this is a distribution across samples, then B is likely to be larger, since the samples are more similar. But if viewed as a distribution of the particular sample, then A might be larger, to be able to get some outlying values.

Scoring: 1 for understanding relationship between sample size and variance, 1 for relationship between variance and distribution.

10 Probabilistic Classifiers (8 minutes, 3 points)

Assume you are given a K-Nearest Neighbor classifier, that has a fairly large value of K (there is a lot of data, so the large value of K works well.) Describe how you could use this to give a probabilistic classifier.

With large K, we will presumably have a distribution of points at different classes among the K nearest neighbors. We can use that number (divided by K) as an estimate of the probability that a test point belongs to each class.

Scoring: 1 for showing understanding K-NN, 1 for understanding probabilistic classifier, 1 for idea of using varying classes in k nearest neighbors.