CS2810 Mathematics of Data Models	Gold
Spring 2019	April 4, 2019

Homework 09

Due: April 10, 9PM
Point total: 60

Instructions:

- Submit your PDF and/or .py file to Blackboard by the due date and time. Please do not zip your files together, as this interferes with Blackboard's preview functionality. Always show all your work, and for full credit, you must use the method that the problem instructs you to use (unless none is mentioned). Handwritten or typeset solutions are both acceptable, but unreadable submissions will be penalized. You may discuss problems with other students, but you may not write up solutions together, copy solutions from a common whiteboard, or otherwise share your written work or code. Do not use code or language that is copied from the Internet or other students; attribute the ideas *and* rephrase in your own words.

Problem 1 (10 pts [3,4,3])

i. Why doesn't $p > 0.05$ imply that the null hypothesis is true?

Solution: A high p-value could simply mean that we don't have enough samples to make a decision. For example, p will generally be high if $N = 2$. Also, notice that $p < 0.5$ still implies that the data was somewhat unlikely under the null hypothesis, just not conclusively so.

ii. If variance increases but everything else about an experiment stays the same, including the p-value threshold for significance, does the chance of Type II error increase, decrease, or stay the same? What about Type I error?

Solution: The chance of Type II error increases, since it's harder to get a low p-value with higher variance (the size of the standard error is larger, resulting in smaller t-values). The chance of Type I error remains the same, because that's determined by the threshold for significance.

iii. In tables of critical values for the t distribution, the column for the two-tailed $p < 0.10$ is the same as the column for the one-tailed $p < 0.05$. In fact, the critical values for the one-tailed distribution for $p < a$ are always the same as the critical values for the two-tailed distribution for $p < 2a$. Why?

Solution: The p values for the one-tailed distribution represent the area under the PDF curve for values that are more extreme in one direction than the critical value c, while the p values for the two-tailed distribution represent the area under the PDF curve that are more extreme than either the critical value c or its negative $-c$. Since the distribution is symmetric about the origin, the area for the two-tailed distribution that is more extreme than the critical value is twice that of the one-tailed distribution, which only uses one of the tails instead of both. Since the p-value for a critical value is this area, it follows that the p-value for the two-tailed distribution is twice that of the one-tailed distribution.

Problem 2 (15 points [4,4,3,4])

A machine learning algorithm generating recommendations for your e-commerce website has decided that people are more likely to buy *Aquaman* on Blu-Ray if they already have purchased the latest season of *Doctor Who* on Blu-Ray. Pulling all 647 customer records since both were available, you find that 121 customers have bought both, 79 have bought just Aquaman, 130 have bought Doctor Who, and 317 have bought neither.

i. Estimating probabilities from the data, what is Pr(Buy Aquaman), and what is Pr(Buy Aquaman | Buy Doctor Who)?

Solution: Pr(Buy Aquaman) is (121+79)/647 = 0.31, while Pr(Buy Aquaman | Buy Doctor Who) is 121/(121+130) = 0.48.

ii. Draw the 2x2 table of expected values, if Pr(Buy Aquaman) and Pr(Buy Doctor Who) were independent.

Solution: We calculate Pr(Buy Doctor Who) = 251/647 = 0.388, and realize that Pr(No Aquaman) = 1 - .309 and Pr(No Doctor Who) = 1 - .388. We multiply the relevant "yes" or "no" probabilities by N to get the following values:

	Yes Dr	No Dr
Yes Aquaman	77.57	122.35
No Aquaman	173.47	273.61

iii. Perform a chi-square test to determine whether these two variables are not independent.

Solution: Computing $(O-E)^2/E$ for each square of the table, we get 24.3+15.36+10.89+6.88, which is easily enough to be significant with only one degree of freedom (the threshold is 3.84).

iv. The e-commerce AI has also decided that people who like the band Honest Bob and the Factory-to-Dealer-Incentives are more likely to like the band The Slip. Your data for this is rather more sparse, with 2 buying both, 4 buying just Honest Bob, 3 buying just The Slip, and 638 buying neither. Determine whether the association between the bands is significant.

Solution: Pr(Honest) = 6/647 = 0.0093, Pr(Slip) = 5/47 = 0.0077. This makes the expected number of people with both 0.0093*0.0077*647 = 0.05. This makes the chi-square value for that square alone $(2 - 0.05)^2/0.05 = 76.05$, which is enough to be significant right there (chi-square values can't be negative). These bands are rare enough that to see even two co-purchases is significant. (The other expected values are 4.94, 5.97, and 636, making the other chi-square values 1.76, 0.65, and 0.006, respectively.)

Problem 3 (10 pts [2,4,4])

Players of an online game complain that the random number generator must be broken, since it seems to generate extreme values too often. You call the integer-generating function 100 times for numbers between 0 and 4 inclusive and get 23 0's, 19 1's, 22 2's, 18 3's, and 18 4's.

i. How many degrees of freedom should you use in your test?

Solution: Four, since there are five numbers but one is determined by all the others (and N).

ii. Can you reject the null hypothesis that the random number generator is fair?

Solution: Doing a chi-square test, we find the chi-square values are 9/20 + 1/20 + 4/20 + 4/20 + 4/20 = 22/20, which is not enough to exceed the critical value of 9.49. There's no reason to think the random number generator isn't fair.

iii. You then try calling the code that should generate normally distributed numbers and get 70 values within 1 standard deviation (in either direction), 15 values that are between 1 and 2 standard deviations away, and 15 values that are 2 or more standard deviations away. Can we reject the null hypothesis that these values are normally distributed? (You can assume 68 percent of the probability is within 1 standard deviation for a normal distribution. Also, we know the mean and variance without calculating them from values here, so we don't lose any degrees of freedom besides the one for the last column.)

Solution: The expected values here are 68, 27, and 5, giving chi-square values of $2^2/68 + 12^2/27 + 10^2/5 = 23.4$. The threshold for 2 degrees of freedom is 5.99. We can reject the hypothesis that this code generates a truly normal distribution.

Problem 4 (10 pts [4,4,2])

A decision tree algorithm that is supposed to decide whether a loan will be paid back is trying to decide which feature to use to build its next decision branch. (The YES and NO classifications below refer to this outcome of whether a loan was paid back.) The possible features are as follows:

- Good borrower credit history: 21 YES examples, 5 NO examples when credit was good. When credit was bad, 5 YES examples, 9 NO examples.
- Borrower is employed: When borrower was employed, 30 YES, 2 NO. When borrower was not employed, 4 YES, 4 NO.

Calculate the expected entropy after splitting on each feature, and identify which choice is the better option for building a tree. Then use a chi-square test to determine whether the decision should actually be pruned instead of being used at all. (You may use the Python function `scipy.stats.chi2_contingency()` if you wish.)

Solution: Credit history feature: Pr(good credit) = 26/40. Entropy of good credit is $-21/26 \log_2 21/26 - 5/26 \log_2 5/26 = 0.71$. Entropy of bad credit is $-5/14 \log_2 5/14 - 9/14 \log_2 9/14 = 0.94$. Expected entropy is therefore 26/40 * 0.71 + 14/40 * 0.94 = 0.79.

Employed feature: Pr(employed) = 32/40. Entropy of employment is $-30/32 \log_2 30/32 - 2/32 \log_2 2/32 = 0.34$. Entropy of unemployment is $-4/8 \log_2 4/8 - 4/8 \log_2 4/8 = 1$. Expected entropy is then 32/40*0.34 + 8/40*1 = 0.47.

So the employment feature is better. The result of a chi-square test is a chi value of about 6.5, significant with $p < 0.05$, so the feature should not be pruned.

Problem 5 (15 points [7,5,3])

In this problem, you'll have a chance to fill in the key mathematical parts of a decision tree classifier, and also see how pruning can prevent overfitting. You'll need the `adult.data.csv` datafile you used for the t-test homework, as well as the decision tree starter code I've provided in `decision_tree.py`. (For simplicity, all numerical features here are equality checks, just like the strings.)

i. Fill in the code for `calcExpectedEntropy()`, which is the function that handles the core choice of splitting decision. You may want to create a `calcEntropy()` helper that just finds the entropy of a list of labels of 1 or 0; your `calcExpectedEntropy()` function will then call `split()` and find the entropies of the two target lists that result, returning their probability-weighted sum. (Note: Check for empty lists, which always have entropy 0 and probability 0.)

ii. Fill in the code for `chiSignificant()`, which determines whether the association between the target label and the feature value really is statistically significant. (You'll want to call `scipy.stats.chi2_contingency()` here; you don't need any arguments besides the 2x2 table, and you only need the p-value.)

iii. Train two versions of the tree, pruned and unpruned. For each tree, train on the first 1000 training examples of the adult dataset (`adult_features[:1000], adult_targets[:1000]`) and test on the next 1000 (`adult_features[1000:2000], adult_targets[1000:2000]`) using `measureAccuracy()`. Report both accuracy figures in your PDF, labeling them as "unpruned" and "pruned."

Solution: The unpruned accuracy is 0.738 and the pruned accuracy is 0.795.

Turn in your python code along with your PDF submission.