

**Course:**

Intro to Data Science – DS-GA-1001 /

Data Mining for Business Analytics - INFO-GB.3336.11

Fall 2014

**Instructor:**

Brian Dalessandro

**Homework 3 – Due 10/22/2014 at 5 pm**

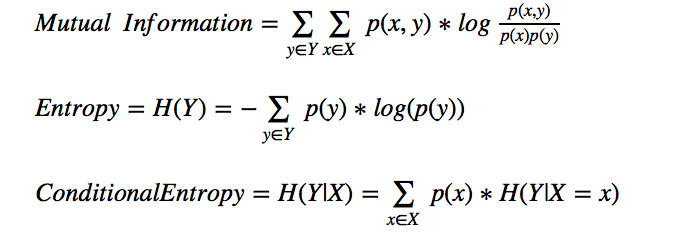
**(Submissions must be a WORD or PDF document)**

Name:\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Grade Total: \_\_\_\_\_\_\_\_\_\_ out of 10**

**Part 1 (2 Points)**

**Assume that X and Y are discrete random variables. The formulas for Mutual Information, Entropy and Conditional Entropy are given by:**

****

**Show mathematically that Mutual Information = Information Gain, where Information Gain = H(Y) – H(Y|X).**

*(Give the derivation below – I’d like for this to be done electronically, so here are some options: 1) do it by hand and scan or take a photo of it, add image to this doc, 2). Use the MS Word equation editor, 3). Use the Latex equation editor in iPython notebook and put image of it here).*

**Part 2 (4 Points)**

**This is a hands-on task where we build a predictive model using Decision Trees discussed in class.**

1. Download the Cell2Cell\_data.csv from NYU Classes > Resources > Data. Also read the Cell2Cell\_info.pdf to understand the contents of the data file.
2. Load the data and prepare it for modeling:
   1. Define a training set (80% of data)
   2. Define a test set (20% of data)
3. Now build a decision tree classifier on the training data to predict the churndep target variable. Make sure to use criterion='entropy' when instantiating an instance of the DecisionTreeClassifier. Also, use all of the default options.

**Question 2c: Using the results of above show a bar plot of features ranked by feature importance (hint: check the attributes of the DecisionTreeClassifier object).**

**Is the relationship between the top 3 most important features (as measured here) negative or positive?**

**If your marketing director asked you to explain the top 3 drivers of churn, how would you interpret the relationship between these 3 features and the churn outcome? What “real-life” connection can you draw between each variable and churn.**

1. Using the classifier built in 2c:

**Question 2d: What is the accuracy on the test data?**

**Part 3 (4 Points)**

**The default options for your decision tree may not be optimal. We need to do analysis on whether tuning the parameters can improve the accuracy of the classifier. For the following options [min\_samples\_split, min\_samples\_leaf]:**

* 1. Generate a range of 10 values of each that make sense to test
  2. For each combination of 3a (there should be 100), build a new classifier and check the classifier’s accuracy on the test data.

**Question 3b: Plot the test set accuracy against the options above. Use the values of min\_samples\_split as the x-axis and generate a new series (line) for each of min\_samples\_leaf. Copy and paste the plot below.**

**Which configuration returns the best accuracy?**

**What is this accuracy?**

**If you were working for a marketing department, how would you use your churn production model in a real business environment? Explain why churn prediction might be good for the business and how one might improve churn by using this model.**