CS 4641 - Homework 1

Karl Gemayel

- Submit your answers as an electronic copy on Canvas.
- No unapproved extension of deadline is allowed. Late submission will lead to 0 credit. After submitting, make sure you **quadruple** check your submission on Canvas to make sure you submitted the right file.
- Typing with Latex is **highly recommended**. Typing with MS Word is also okay (**Submitted file should be in PDF format. Please do not submit .docx - it will NOT be graded**). Also, no handwritten submissions will be accepted.
- Do not submit a zip file. Your final submission should consist of three separate files: HTML or PDF version of document answering questions 1-3, completed Jupyter notebook, HTML/PDF version of completed (and executed Jupyter notebook).
- Explicitly mention your collaborators if any. Remember what I said about collaboration though!

1 Probability

Matt, Marc, Lucas, and Jonah have just been assigned a case. They're off to examine where Daniel Geoffrey Sustar (aka D.G.Sustar) stores all the wine he makes. They've narrowed it down to four locations in the UK. Each of them will visit one of these places and survey the locals on whether or not they've seen D.G.Sustar in the area in the past year (with bottles of wine).

Matt will take the London area, Marc will investigate Ireland, Jonah will take on Wales (funny, no?), and Lucas will go forth into the Highlands (Scotland!).

Five weeks later, they have each surveyed 100 people from their respective lands. They asked the locals the following question: Have you seen D.G.Sustar in the area over the past year. The locals were allowed to answer Yes or No. The results of the survey are shown in the table below:

Surveyor	Yes	No
Matt	58	42
Marc	56	44
Lucas	58	42
Jonah	57	43

Since the survey results are very close, the quartet find 6 more people to survey. The first three have lived in both London and Ireland over the past year. The remaining three have lived in both Wales and Scotland. The question that each will be asked is: In which of the two places you have lived in do you think D.G.Sustar left his wine? The answer can be either location, or both (but cannot be neither).

Suppose that the answers of all remaining participants are independent and that each participant is equally likely to choose as their answer: location A, location B, or both, where A and B are the two locations they visited.

Note: If a person answers location A, then the Yes count for that corresponding surveyor will increase by one, and the No count for the location B surveyor increases by one. Similarly for B. If the person answers "both A and B", then the Yes count for both surveyors increases by one.

(a) What is the probability that D.G.Sustar kept his wine in London?

(b) What is the probability that D.G.Sustar kept his wine in Ireland?

(c) What is the probability that D.G.Sustar kept his wine in Scotland?

(d) What is the probability that two locations are equally likely?

2 Information Theory

Suppose the joint probability distribution of two binary random variables X and Y are given as follows.

$X \backslash Y$	0	1
0	$\frac{1}{4}$	$\frac{1}{2}$
1	0	$\frac{1}{4}$

(a) Find entropy $H(X)$ and $H(Y)$.

(b) Find conditional entropy $H(X|Y)$ and $H(Y|X)$.

(c) Find mutual information $I(X;Y)$.

(d) Find joint entropy $H(X,Y)$.

(e) Suppose X and Y are independent. Show that $H(Y|X) = H(Y)$

(f) Suppose X and Y are independent. Show that $H(X,Y) = H(X) + H(Y)$

(g) Show that $I(X;X) = H(X)$

(h) Show that $H(X,Y) - H(X) = H(Y|X)$

For all the above, show your work, step by step. No points will be given otherwise. Also, for parts e through h, you HAVE to start from the basic definitions of entropy, joint entropy, and conditional entropy. You cannot use already-proven or "intuitive" theorems about these concepts. In short - do some math.

3 Classification Methods

(a) In no more than 5 sentences, describe the main difference between bagging and boosting.

(b) ID3-trained decision trees are more resistant to outliers than soft-margin SVMs. Do you agree? Why or why not?

(c) Mark T if the statement is true, and F otherwise. Explain why in 1-2 sentences. No points if explanation is incorrect.

- Binary decision trees can model any continuous function.
- A neural network with 4 hidden layers and 200 neurons per layer can model any continuous function.

4 Programming: Neural Networks

Here, you will implement a 2-hidden layer Neural Network. The `Assignment 1 - Template.ipynb` file contains all the information you need. You will have to download the Anaconda package manager (`anaconda.com`) and use the Jupyter Notebook application in order to open it. You should use the python 2.7 version of Anaconda. Your submission should be the completed notebook **and a PDF or HTML version of it showing the results** (i.e. after you run it, just save it as a PDF/HTML). **Your submission will not be graded if the PDF/HTML file is not submitted.**