Homework 1

1 Probability and Statistics

1. In a bag there are 5 red balls and 7 green balls. Two balls are selected without replacement.
   (a) What is the probability that both are red?
   (b) What is the probability that the first is red and the second is green?

2. - 1% of women over 50 have breast cancer
   - 90% of women who have breast cancer test positive on mammograms
   - 8% of women who test positive do not have breast cancer

   (a) If you assume everyone being tested is over 50, what is the probability that a woman with a positive mammogram test result has cancer?

3. Imagine a game where you win the dollar amount of a die roll. You can choose to play this game or receive a guaranteed $3.50. Which do you choose and why?

2 Neural Nets

1. Imagine a neural network for regression with a linear layer, a ReLU activation function, and another linear layer. The first linear layer takes in 1000 inputs (features) and the second linear layer takes in 100 hidden units.

   Input
   ↓
   Linear Layer
   ↓
   ReLU Activation Layer
   ↓
   Linear Layer
   ↓
   Output

   (a) What is the purpose of the ReLU activation layer? Why is it different than just having two linear layers?
   (b) What are the input and output shapes for each layer (i.e. how many features does each layer take in and how many does it spit out)?
   (c) How many parameters are used in this model (i.e. how many weight values and how many bias values)?

(d) What are some hyperparameters we can tune in this model (give at least 3)?

2. Let $f(\vec{x}) = ReLU(\vec{x})$. What is the output of:

    (a) $f([5, -1, -3, 2, 4])$

    (b) $f([-1, -2, -5, 10, 5])$

3 NLP

1. In this part of the assignment, you will read in and do some basic manipulation of a text corpus (a fancy term for a set of documents). Included with the assignment is a file nyt.txt containing 8860 sentences taken from New York Times articles, one sentence per line. You should implement your solutions in a file called a1.py. For your final submission, please include a1.py along with a document containing your answers to the questions in this homework.

    In parts a - c, you will investigate tokenization schemes. Tokenization is the process of splitting raw text into words. In English, this involves splitting out punctuation and contractions (shouldn't becomes should 'nt) and is typically done with rules.

    (a) Use whitespace tokenization (that is, splitting the sentences up only by spaces, tabs, etc). List the top ten words that you find and their counts, as well as any patterns you see.

    (b) Now, use smarter tokenization like nltk.word_tokenize, which uses different delimiters (such as apostrophes) to break up the text. Report the top 10 words, counts, and patterns. Is this any different than whitespace tokenization?

    (c) How do you think that tokenization might affect a natural language processing system downstream (e.g. preprocessing, modelling)?

2. In parts d - e, you will investigate Zipf's Law. A word has rank n if it is the nth most common word. Zipf's Law states that the frequency of a word in a corpus is inversely proportional to its rank. Roughly speaking, this means that the fifth most common word should be five times less frequent than the most common word, and the tenth most common word should occur half as much as the fifth most common word.

    (d) Make a plot of inverse rank vs. word frequency for the tokenized data. Include your plots in your submission. Matplotlib is a good tool to use, but Excel/Matlab/Gnuplot/others are okay too.

    (e) Where does Zipf's law appear to hold? Are there any outliers? If so, list one or two and why you think they might be outliers?