# Statistical Natural Language Processing : Assignment 1

## 1 Probability Basics

1. Assume two random variables $X$ and $Y$ are independent and continuous with probability density functions $f_X$ and $f_Y$ respectively.

   (a) Show that probability denisty function of $X + Y$ at $z$ is $f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(z - y) f_Y(y) dy$

   (b) If $f_X = \mathcal{N}(X; \mu_1, \sigma_1^2)$ and $f_Y = \mathcal{N}(Y; \mu_2, \sigma_2^2)$, Show that $f_{X+Y} = \mathcal{N}(X + Y; \mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$

   (c) Let $X_1, \ldots, X_n$ be idependent and identically distributed continuous random variables with c.d.f $F_X$ and p.d.f $f_X$. Obtain the distribution for the $i^{th}$ smallest random variable among $X_1, \ldots, X_n$.

2. The moment generating function (MGF) $\phi(t)$ of a continuous random variable $X$ with p.d.f $f_X$ is defined as $\phi(t) = \mathbb{E}(\exp(tX)) = \int_{\infty}^{\infty} \exp(tX) f_X(x) dx$.

   (a) Show that $\phi'(0) = \mathbb{E}(X)$ and $\phi''(0) = \mathbb{E}(X^2)$ where $\phi'(t)$ and $\phi''(t)$ represents the first and second derivatives of the moment generating function w.r.t $t$. This shows that various moments of $X$ can be obtained by taking higher order derivatives of the moment generating function. Note that moment generating function uniquely determines the distribution.

   (b) Obtain moment generating function for a Normal random variables and use it to show $1(b)$.

   (c) Show that the sum of squares of $n$ independent standard Normal random variables is chi-squared distributed with $n$ degrees of freedom. Let $X_1, \ldots, X_n$ be independent and Normally distributed random variables with mean $\mu$ and variance $\sigma^2$. Let $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ be sample mean and $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$ be sample variance. Prove that $\frac{(n-1)S^2}{\sigma^2}$ is Chi-square distributed with $n - 1$ degrees of freedom.

## 2 Linguistic Essentials

Read Chapter 3 in the Foundations of Statistical Natural Language Processing book.

1. What are the parts of speech tags (based on Brown Tagset) of the words in the following paragraph?

   A gnostic was seated before a grammarian. The grammarian said, ‘A word must be one of three things: either it is a noun, a verb, or a particle. The gnostic tore his robe and cried, “Alas! Twenty years of my life and striving and seeking have gone to the winds, for I laboured greatly in the hope that there was another word outside of this. Now you have destroyed my hope.’

2. Provide all possible parse trees (along with their meanings) and attachment ambiguities in the following sentences. Provide an example of attachment ambiguity in your native language.

   This is the dog that worried the cat that killed the rat that ate the malt that lay in the house that Jack built.

# 3 Text Processing

Provided is a dataset containing tweets from Twitter and a categorization of tweets into two classes. The file has the following format (Note that tweet text can span multiple lines )
{tweet text}, label

Write a python program which process the data in the provided file (Input) and extract the following information (submit separate codes for each task).

1. Extract 20 most frequent hashtags, usermentions and URLs mentioned in the data set.
2. Preprocess data to remove stopwords and punctuations and represent text in lower case. Extract 20 most collocating word bigrams and trigrams based on (a)Frequency (b)t-score (c)chi-square measure and (d)pointwise mutual information. (without using the collocation package in NLTK ! ).
3. Train a Naive Bayes classifier using the provided data which can classify text into either of two classes. The input to the classifier will be a file containing tweets in the format '{tweet text}' and output should be a file containing labels assigned to the text in each line. You could try different representations of the text ( unigrams, bigrams or extracted features) and submit the program which you expect to give best performance on the unknown test data.