ECE8813: Statistical Natural Language Processing, HW2 Problems

1. Given *n* independently and identically distributed (iid) samples, (x_i, y_i) from a bivariate Gaussian source: (1) derived the likelihood function of the observations; (2) draw a scatter plot of the samples on a 2-D *x*-*y* plane; (3) can you plot a 3-D histogram of the data to get a sense about the value of the unknown parameters? (4) derive the maximum likelihood (ML) estimates for the five parameters in the mean vector and the covariance matrix; (5) given the data set in hw2-binormal.txt which lists 5000 pairs of data, estimate the above set of five parameters using ML as above with 500 pairs of samples, then the next 500, until all 5000 pairs are exhausted, compare the ten sets of estimated parameters. Can you guess what were the parameters used to generate the data? What can you say to summarize your observation?

2. A discrete random vector $R=(r_1, r_2, \ldots, r_m)$ has a multinomial distribution: (1) what is the expected mean vector, E[*R*]? (***Note***: we assume the sum of all *m* variables is *n)*; (2) derive the expected covariance, $E\{(R\text{-}E[R])^t(R\text{-}E[R])\}$? (3) given a sequence of *n* observations from the above multinomial distribution the total number of samples that belongs Category *i* is Count(r_i)=q_i, what is the maximum likelihood for r_i? (***Note***: sum of p_i is equal to 1, you need to use this constraint and Lagrange theory to derive the formula for the ML estimate); (4) can you consider the 27 symbols in Lab1 to follow a multinomial distribution, why? (5) when estimating p_i=*P*(observing i^{th} symbol) in Lab1 are you using the same ML formula obtained in Item (3),? If not, do you see any similarity? (5) for the case of observing letter pairs in Lab1, can we model this event as a multinomial distribution? If so, how many distinct events do we need to consider, i.e. what is the value of *m* in the case of estimating letter bigrams?

3. Assume that sentence beginning and ending marks, Sen_B and Sen_E, are two distinct words. Using the first 2500 WSJ sentences, do the following: (1) Find out how many distinct words and word pairs? (2) Find out, based on the frequency of occurrence, the top 5 and bottom 5 words and word pairs; (3) Estimate the word unigrams and bigrams for the 20 items in Item (2); (4) Repeat the above Items (1)-(3) with the same 10K WSJ sentences; (5) What can you say about the differences in the two sets? Does increasing the dataset size gives different effects from those in Lab1?