10-704 Homework 1
Due: Thursday 2/5/2015

Instructions: Turn in your homework in class on Thursday 2/5/2015

1. **Information Theory Basics and Inequalities** C&T 2.47, 2.29

   (a) A deck of n cards in order $1, 2, \ldots, n$ is given to you. You remove one card at random and then place it again at one of the n available positions at random. What is the entropy of the resulting deck?

   (b) Let X, Y, Z be joint random variables. Prove the following inequalities and identify conditions for equality.

      i. $H(X, Y|Z) \geq H(X|Z)$
      ii. $I(X, Y; Z) \geq I(X; Z)$
      iii. $H(X, Y, Z) - H(X, Y) \leq H(X, Z) - H(X)$

   (c) Consider a random variable X supported on $\{1, \ldots, m\}$ with $\mathbb{P}(X = i) = p_i$. We will assume $p_1 \geq p_2 \geq \cdots \geq p_m$. Let $\mathbf{p} = [p_1, \ldots, p_m]$. Since $X = 1$ is the most likely assignment, the minimal probability of error predictor of X is $\widehat{X} = 1$ with probability of error $P_e = 1 - p_1$. Maximize $H(\mathbf{p})$ subject to the constraint $1 - p_1 = P_e$ to find a bound on P_e in terms of the entropy. This is Fano's inequality in the absence of conditioning.

2. **Submodular Feature Selection** Here we study the problem of trying to predict a random variable Z given a collection of random variables $X_1, \ldots, X_p$ (called features). The goal of **feature selection** is to find a small subset of the features that predict Z well.

   (a) Show that the mutual information function $f(S) = I(Z; X_s, s \in S)$ is *not* submodular. This provides evidence that greedy maximization of the mutual information functional may not be a good way to do feature selection.

   (b) Show that in the naive bayes model, greedy maximization of mutual information is a theoretically justified approach for feature selection. The naive bayes model posits that $X_i \perp X_j|Z$ for all $i \neq j$ so the distribution factors as $P(Z, X_1, \ldots, X_p) = P(Z) \prod_{i=1}^{p} P(X_i|Z)$.

3. **Unbiased Estimation of Entropy Functionals** In class we mentioned that there are no practical unbiased estimators for entropy functionals. One can however design an unbiased estimator if you are allowed to choose a set of samples of arbitrary but finite size. The problem is that there is no *a priori* bound on the sample size. In this question we will develop and analyze these estimators for the discrete setting. Let $X_1, X_2, \ldots$ denote a sequence of samples from a discrete distribution P with symbols $C_1, \ldots, C_k$ and probabilities $(p_1, \ldots, p_k)$.

(a) For $1 \le i \le k$, let N_i denote the smallest $j \ge 1$ for which $X_j = C_i$. Show that:

$$\widehat{H}_1 = \sum_{i=1}^{k} \frac{\mathbf{1}[N_i \ge 2]}{N_i - 1} \tag{1}$$

is an unbiased estimator for the entropy $H(P) = -\sum_{i=1}^{k} p_i \log p_i$. The expansion $\log(1-x) = -\sum_{j=1}^{\infty} x^j/j$ may be useful.

(b) Design an unbiased estimator for the entropy $H(P)$ based on pairing each of the first n samples with the next sample in the sequence with the same symbol. The identity $\frac{\log(1-x)}{1-x} = -\sum_{i=1}^{\infty} h_i x^i$ where $h_i = \sum_{j=1}^{i} \frac{1}{j}$ is the ith harmonic number will be useful.

4. **Estimation of KL Divergence** Describe how to estimate the KL divergence $D(p||q)$ using the first-order Von-Mises Expansion approach. Say you are given $2n$ i.i.d. samples from each distribution ($\{X_i\}_{i=1}^{2n} \sim p$ and $\{Y_i\}_{i=1}^{2n} \sim q$).