CMSC 858L — Spring 2009 — Homework #1

You may work together on the homeworks, but you must write up your own solutions. Neatness counts — homeworks that are unreadable (by me) will receive a score of 0. No late homeworks will be accepted. If you use outside sources (papers, books, classmates, etc.) you must cite them. Use can use any of the formulas on the Information Theory Hint sheet without proof (and you probably will have to).

**Due:** at the beginning of class on March 12, 2009.

1. Prove $H(X) \geq H(X \mid Y)$.

2. **True or False.** $H(X \mid Y = y)$ can be larger than $H(X)$. Explain your answer.

3. Show that if $H(X \mid Y) = 0$ then for all $y$, there exists only one $x$ such that $p(x, y) > 0$.

4. A **metric** is a function $\mu(x, y)$ such that: (i) $\mu(x, y) \geq 0$; (ii) $\mu(x, y) = \mu(y, x)$; (iii) $\mu(x, y) = 0$ if and only if $x = y$; and (iv) $\mu(x, y) + \mu(y, z) \geq \mu(x, y)$.

   (a) Show that mutual information $I(X; Y)$ is not a metric. (b) A **pseudometric** is a function $\mu$ that obeys all the properties of a metric except that $\mu(X, Y) = 0$ does not necessarily imply that $X = Y$. Is $I(X; Y)$ a pseudometric?

5. When $X, Y$ are random variables, we say $X = Y$ if their probability distributions are the same up to relabeling. Prove that $VI(X, Y)$ satisfies the properties (i)–(iii) of metrics above. Extra credit: prove $VI$ satisfies the triangle inequality (iv).

6. Describe an algorithm to compute the number of shortest paths from a node $u$ to every other node in a graph $G$ in time $O(n + m)$, where $n$ is the number of nodes and $e$ is the number of edges.

7. Suppose you have a hierarchical decomposition $T$ of a set of proteins where $T$ was derived without using any network information. Let $G$ be an interaction network where the nodes of $G$ are the proteins at the leaves of $T$. Explain how to find a clustering induced by a node-cut of $T$ with maximum modularity in $G$.

8. Give a small example where the greedy graph summarization algorithm does not find the optimal summary.