

Homework Assignment 1

February 16, 2004

Due Date: March 1

Follow the Collaboration Policy closely!! If you forgot, check it out at the course webpage by following the link at www.cse.wustl.edu/~zhang/teaching.

YOU SHOULD START AT LEAST THINKING ABOUT THESE PROBLEM RIGHT AWAY!

1. Describe the main differences between traditional statistics and Bayesian statistics. Argue when you should apply traditional statistics and when you should consider Bayesian method.
2. Let X, Y, Z be Boolean random variables. Label the eight entries in the joint distribution $P(X, Y, Z)$ as a to h . Express the statement that X and Y are conditionally independent given Z as a set of equations relating a through h . How many *nonredundant* equations are there?
3. The example on discriminating two coins we discussed in class actually has an application in sequence analysis. It is known that there are certain regions, called *GC-islands*, in a genome where the GC content - the frequencies of G and C - is higher than the GC content for the regions outside of GC-islands. We can consider the nature used two coins when constructing a genome, one for the GC-islands and the other for the other sequence regions. Assume that the GC content within GC islands is 75% (so the AT content is 25%) and the GC content outside of GC islands is 40%. You are given following sequence $\langle h, h, t, h, t, t, h, t, h, h, h, t, h, h, h, h, t, h, h, t, h, h, h, h, t, h, h, t, h, h, t, t, h, t, t, h, t, h, h, t, h, t \rangle$, where h stands for observing a base G or C, and t for a base A or T. Can you predict if there is a GC-island in this sequence? How many GC-islands can you find? Predict the locations of GC-islands if you believe there is at least one. If not, state why there should not be such an island.
4. Suppose you are a witness to an attack involving a taxi in Baghdad. All taxi in Baghdad are blue or green. You swear, under oath, that the taxi was blue. extensive testing shows that, under the dim lighting conditions, discrimination between blue and green is 75% reliable. Is it possible to calculate the most likely color for the taxi? (Hint: Distinguish carefully between the proposition that the taxi *is* blue and the proposition that it *appears* blue.) What about now, given that 9 out of 10 Baghdad taxis are green?
5. We briefly discussed in class a Bayesian network for modeling gene microarray data where we only considered possible clusters of genes and experimental conditions (or arrays). Extend this simple model to include transcriptional factors and transcriptional factor binding motifs. Explain what the nodes and links should be and how they should be connected, i.e., consider where conditional independence relationships may occur.