Stat521A Spring 2009: homework 6

In this final homework, you will compare various graphical and non-graphical density estimators on some discrete (categorical) data.
Download the latest version of PMTK from the URL below, bypassing the usual download mechanism: `http://www.cs.ubc.ca/~murphyk/pmtk/pmtk1.4.4.zip`. (The reason this is not yet public is that I have not tested that all the demos still work after various changes I have made; for the purposes of this homework, the code should be fine :)
Run the file `examples/discreteDensityEstimationShootout`. It compares performance and speed of the following density models: product of multinoullis, mixture of product of multinoullis, and a tree (learned using the Chow-Liu algorithm). It uses two datasets: a biological dataset (Sachs) with ternary variables, and a text dataset (newgroups) with binary variables (see Figure 1).
Do the following (as usual, turn in your code and plots).

1. Implement mixtures of trees using EM. See [MJ00] for the details; the core is the MixTree algorithm in their figure 6. More precisely, implement a PMTK class `DgmTreeTabularMix` with `fit` and `logprob` methods, so it can be plugged in to the shootout demo above. Optional: do MAP estimation instead of just MLE.

2. Try $K = 1, 2, 5, 10$ mixture components. ($K = 1$ should be the same as the existing tree code.) Print the boxplots generated by the shootout.

3. Which value of K would BIC choose? Which value of K would cross validation choose? (In the latter case, you have to do CV within each training fold, thus you will have two nested CVs.)

4. For the best K, plot the tree structures learned in each mixture component when applied to the newsgroup data. As an example of how to do this, run `examples/chowliuDemo`. You should get a figure like Figure 2.

References

[MJ00] M. Meila and M. I. Jordan. Learning with mixtures of trees. *J. of Machine Learning Research*, 1:1–48, 2000.

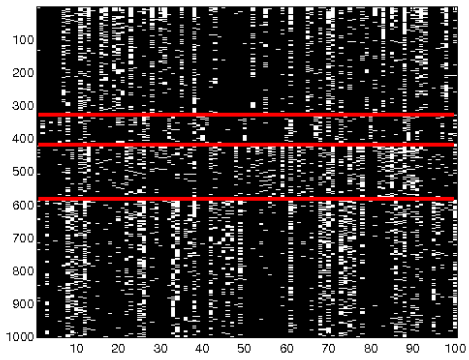

Figure 1: Subset of the newsgroups data. Each row is a document (represented as a bag-of-words bit vector), each column is a word. The red lines separate the 4 classes, which are, in descending order: comp, rec, sci, talk (these are the titles of USENET groups). We can see that there are subsets of words whose presence or absence is indicative of the class. Produced by `newsgroupsVisualize`.

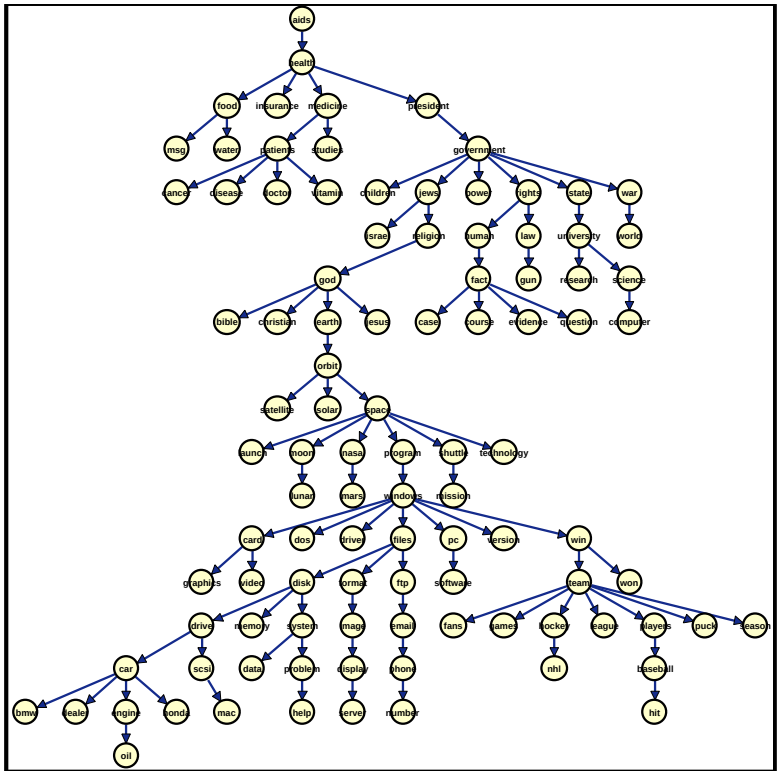

Figure 2: Chow-Liu tree learned from the newsgroups data (class labels are ignored). Note that the direction of the arrows is not important; node 1 (corresponding to the word "aids") was arbitrarily chosen as the root, and all edges flow outwards from there. Produced by `chowliuDemo`.