Visualization and Learning Structure: Problem Session 7/5/16

We're looking at some variables from a real data set about mammographic masses (tumors in breast tissue). The data set has only been altered to remove missing values. **Remember to look at today's code to help you with the below problems.**

M. Elter, R. Schulz-Wendtland, and T. Wittenberg (2007). *The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process*, Medical Physics 34(11), p.4164-4172

This data set is used to develop classification models for whether or not patients have a malignant mass without invasive biopsies. We have some mammography results for 816 patients whose lesions were later determined to be malignant or benign.

BIRADS is an assessment: 1 = definitely benign; 5 = highly suggestive of malignancy. *Age*: patient's age in years. *Shape*: 1 = round, 2 = oval, 3 = lobular, 4 = irregular. *Margin*: 1 = circumscribed, 2 = microlobulated, 3 = obscured, 4 = ill-defined, 5 = spiculated. *Density*: 1 = high, 2 = iso, 3 = low, 4 = fat-containing.
Severity gives the classification of the mass: 1 = malignant, 0 = benign.

Download *mammogram.txt* from the website (`http://www.stat.cmu.edu/~rnugent/PCMI2016`) and read it into R: `mammogram<-read.table("mammogram.txt")`.

1. The *Shape*, *Margin*, and *Density* variables are all ordered, as is *BIRADS*.
   Use sunflower plots to examine the relationship between the physician's assessment (*BIRADS*) and each of the three lesion descriptions. Describe the trends (if any).
   Which of the three seem to have the strongest relationship with *BIRADS*?

2. Create a contour plot (default *nlevels*) of a two-dimensional kernel density estimate (default bandwidth) for *Age* and *Margin*. (`library(MASS); kde2d`)
   Add the observations to the graph using `points() (pch = 16)`, color-coded by the severity of the masses: malignant = green; benign = red.
   You can create a vector of colors by using ifelse():
   `col.vec<-ifelse(Severity==1,"green","red")`; then use `col=col.vec`
   (a) How many modes does the density estimate have? Given the integer-valued variables, why do we still see "circular/oval" modes in the density estimate?
   Can you characterize this relationship by whether the mass was malignant?
   (b) Experiment with the bandwidths to find a density estimate that better reflects the "striped" nature of the data (include your graph). What do you choose?
   Which dimension dictates whether or not there are density estimate valleys?

3. We're interested in the relationship between *Age* and *BIRADS* with respect to *Severity* but are concerned that the ordinal *BIRADS* variable will make it difficult to see any possible trend. Jitter both *Age*, *BIRADS* by 0.25 (`help(jitter)`), and create a density estimate of the jittered variables using the default parameters.

   Create a heat map of this density estimate. Add the observations to the heat map using `points()` (`col=1`) where the patients are pch-coded according to severity of the mass: malignant = x (`pch=4`); benign = o (`pch=1`). Can use`pch.vec<-ifelse(Severity==1,4,1)`, then `pch=pch.vec`.

   (a) Describe any high frequency areas.
   Do they correspond more commonly to malignant or benign masses?

   (b) Is there a relationship between *Age* and *BIRADS*? Is it dependent on *Severity*?

   (c) What is the maximum amount that we could jitter each variable without changing the structure of the data?

4. Experiment with using `image()` (for example) to visualize matrices you're creating in your wavelet work. In particular, experiment with visualizing "cross-sections" or "level sets" of your matrix by using something like the following:

```
yourmatrix<-read.csv("yourmatrix.csv")
##or read.table("yourmatrix.txt") depending on how you stored it
image(yourmatrix)
levelset<-ifelse(yourmatrix>threshold,1,0)
##you pick the threshold number
image(levelset)
```