STAT460 – Homework 9
Due: April 1 at the start of class.

This homework set applies methods we have developed so far to a medical problem: gene expression in cancer.

A gene is a stretch of DNA inside the cell that tells the cell how to make a specifc protein. All cells in the body contain the same genes[1], but they do not always make the same proteins in the same quantities; the genes have different expression levels in different cell types, and cells can regulate gene expression levels in response to their environment. Different types of cells thus have different expression profles. Many diseases, including cancer, fundamentally involve breakdowns in the regulation of gene expression. The expression profle of cancer cells becomes abnormal, and different kinds of cancers have different expression profles.

Our data are gene expression measurements from cells drawn from 64 different tumors (from 64 different patients). In each case, a device called a microarray (or gene chip) measured the expression of each of 6830 distinct gene[2], essentially the logarithm of the chemical concentration of the gene's product. Thus, each record in the data set is a vector of length 6830.

The cells mostly come from known cancer types, so there are classes, in addition to the measurements of the expression levels. The classes are breast, CNS (central nervous system), colon, leukemia, melanoma, nsclc (non-small- cell lung cancer), ovarian, prostate, renal, K562A, K562B, MCF7A, MCF7D (those three are laboratory tumor cultures) and unknown.

The dataset and explanation are on Blackboard.

```
genesT = read.table("../data/nci.data")
genes = t(genesT) #This puts matrix in the usual nxp form

Y =
c('CNS','CNS','CNS','RENAL','BREAST','CNS','CNS','BREAST','NSCLC','NSCLC','RENAL','RENAL',
'RENAL','RENAL','RENAL','RENAL','RENAL','BREAST','NSCLC','RENAL','UNKNOWN','OVARIAN',
'MELANOMA','PROSTATE','OVARIAN','OVARIAN','OVARIAN','OVARIAN','OVARIAN','PROSTATE','NSCLC',
'NSCLC','NSCLC','LEUKEMIA','K562B-repro','K562A-repro','LEUKEMIA','LEUKEMIA','LEUKEMIA',
'LEUKEMIA','LEUKEMIA','COLON','COLON','COLON','COLON','COLON','COLON','COLON','MCF7A-repro',
'BREAST','MCF7D-repro','BREAST','NSCLC','NSCLC','NSCLC','MELANOMA','BREAST','BREAST',
'MELANOMA','MELANOMA','MELANOMA','MELANOMA','MELANOMA','MELANOMA')

#This part makes a color vector for plotting purposes
color = rep(0,length(Y))
tmp   = rainbow(length(levels(as.factor(Y))))
sweep = 1
for(lev in levels(as.factor(Y))){
  color[Y == lev] = tmp[sweep]
  sweep = sweep + 1
}
```

[1] Although, oddly, red-blood cells do not contain any DNA

[2] Strictly speaking, the RNA levels are measured, not the protein levels.

1. Visualization with PCA.

   Plot the first two PC scores of each cell. Label Label each cell by its type. What do we call these coordinates, in general. Ordinarily, this would be done with a biplot. Why won't this work very well in this case? *if you can't think of an answer, try to run it.*

   ```
   #Get the scores via some method
   plot(scores[,1:2],xlab='PC 1',ylab='PC 2',type='n')
   text(scores[,1:2],label=Y,col=color,cex=.5)
   ```

   (a) One tumor class (at least) forms a group in the projection. Say which, and explain your answer.

   (b) Identify a tumor class which does not form a compact cluster.

2. Redo the previous question but with the 1^{st} and 3^{rd} PC scores.

3. What percent of the total variance is contained in first 3 PCs? What follows isn't workable code, it is just a guide for you to follow. Refer to the lectures to fill in details.

   ```
   #Get the PCA decomposition of genes via prcomp
   #save summary( prcomp output ) to an object
   # Get the third row/ third column of that object
   ```

4. Using random forest, find the OOB misclassification rate. Looking at the OOB confusion matrix, which cancer type is predicted correctly most often? Which one(s) least often?

5. Suppose we have two (continuous valued) predictors X_1 and X_2. Draw an example (of your own invention) of a partition of these variables that could result from binary splitting. Your example should contain at least six regions. Draw a decision tree corresponding to this partition. Be sure to label all aspects of your figures, including the regions R_1 through R_6 and the split values for the tree (e.g. $X_1 < 10$).