MSIT 423: Data Mining
Professor E.C. Malthouse
Final
Spring, 2018

- You may use a R, a pencil or pen, JWHT and your notes. **You may not talk to anyone about the exam.**
- The midterm is due by midnight CST on Wednesday, June 12. Write your answers in a single Word or PDF file and email to me before the due date at ecm@northwestern.edu.

1. I have a sample of people from the southern US and know the length of time spent watching different news channels: Fox News, CNN, MSNBC, broadcast news (ABC, NBC, CBS and local news), and public TV (PBS and BBC). The lengths of time spent watching the different stations were logged and then mean centered prior to analysis. The means, standard deviations and the solution to the $K = 6$ cluster solution are shown below:

```
> round(apply(news[ok,14:18], 2, mean), 4)
      Zfox       Zcnn     Zmsnbc Zbroadcast    Zpublic
         0          0          0          0          0
> apply(news[ok,14:18], 2, sd)
      Zfox       Zcnn     Zmsnbc Zbroadcast    Zpublic
  3.239593   2.991079   2.709933   1.750006   1.715756
> fit = kmeans(news[, c(8,10,11,9,12)], 6, iter.max=1000, nstart=20)
> summary(fit)

       n  Pct  Zfox  Zcnn Zmsnbc Zbroadcast Zpublic   RMSE
1  40267 0.11 -3.84  0.65  -1.43      -0.26   -0.37 1.6094
2  50958 0.14 -5.03 -4.06  -1.88      -1.61   -0.82 1.3890
3  77990 0.21  2.18  1.19  -1.85       0.39    0.00 1.5272
4  57936 0.16  0.60 -3.69  -1.68      -0.42   -0.53 1.5560
5  70310 0.19  2.89  3.25   3.44       0.86    1.27 1.5381
6  65463 0.18  0.04  1.12   2.34       0.40   -0.02 1.4919
  362924 1.00  0.00  0.00   0.00       0.00    0.00 1.5185
SSE =  4184120 ; SSB =  7716704
R-Squared =  0.6484177
Pseudo F =  133864.8
```

   (a) (12 points) Interpret the solution by writing <u>one sentence</u> about each cluster and giving it a name.

(b) (3 points) I chose not to standardize/scale the data (i.e., make the variables have standard deviation 1) prior to clustering. What would you expect to happen to the cluster solution if I had standardized them prior to clustering?

2. This problem uses the `breastcancer.csv` file on canvas. There are n=699 patients with tumors. The variable `y` equals 1 for those with cancer and 0 for those benign tumors. There are nine lab measures (`x1`–`x9`) and the goal of this problem is to build a machine classifier to associate the lab measures with outcome `y`.

   (a) Estimate a random forest model predicing y from the other variables. Report the out of bad estimate off prediction error. How many trees would you recommend having in your forest? Paste in the function call for your best model and give the output. Hint:

   ```
   bc = read.csv("breastcancer.csv")
   randomForest(factor(y)~., data=bc[,-1]) # -1 drops patient id
   ```

   (b) Estimate a bagging tree model. How many trees would you recommend? Paste in the function call for your best model and give the output.

   (c) Which model, random forests or bagged tree, would you suggest and why?

   (d) For your best model so far, generate a variable importance plot and tell what the four most important variables are.

   (e) For your best model so far, use the `partialPlot` function to study how `x3` is assocated with cancer. Write one sentence describing the plot.

   (f) For your best model so far, find AUC. Hint: if `fit` is the fitted object then use

   ```
   \library(pROC)
   plot.roc(bc$y, fit$votes[,2], print.auc=T)
   ```

   (g) Use logistic regression to predict `y` from all nine predictors without any transformations. Which variables are significant? Are the same as the ones from RF/bagging?

   (h) Estimate a ridge model and use `cv.glmnet` to find the best value of λ. Report the value of the best λ and AUC for the model. Hint:

   ```
   x = model.matrix(y ~ ., bc[,-1])
   fit.ridge = glmnet(x, bc$y, family="binomial", alpha=0)
   fit.cv = cv.glmnet(x, bc$y, family="binomial", alpha=0) # find optimal lambda
   phat = predict(fit.ridge, s=fit.cv$lambda.min, newx=x, type="resp")
   ```

(i) Estimate a lasso model and use `cv.glmnet` to find the best value of λ. Report the value of the best λ and AUC for the model. Also indicate which variables are dropped by lasso.

(j) Which model is best based on AUC?

3. This problem analyzes Fisher's Iris data, which should be available in R under the name `iris`. There are 50 observations on each of three species of irises (`Species`) giving a total of 150 observations. Four measurements were taken on each iris: the sepal length and width and the petal length and width, all measured in CM. Fisher's original application was supervised learning, where the four features were used to classify the (observed) species. We will apply unsupervised learning methods.

   (a) (3 points) Generate a scatterplot matrix encoding species as color as follows:

   ```
   plot(iris[, 1:4], col=iris$Species, pch=16)
   ```

   From now on, call the species black, red and green, consistent with the plot (black=setosa, red=versicolor and green=virginica). Comment briefly on the petal size of the blacks versus the others. Also comment on the sepals of the blacks relative to the others.

   (b) (2 points) Compute basic descriptives (mean, SD, min, max) of the four measurements.

   (c) (4 points) In the next parts you will use PCA to visualize the data. Before running PCA, discuss briefly the implications of standardizing the four variables versus not standardizing.

   (d) (3 points) Find PCs *without* standardizing. What fraction of variance (not SD) is accounted for by each of the first two PCs. Interpret them.

   (e) (3 points) Based on your inspection of the scatter plots, would you recommend K-means clustering? Explain briefly.