B565: Homework 1

1. Consider the pairs plot of the "iris" data discussed in class. If we want to predict the iris species using only one of the petal and sepal measurements, based on this plot which one should it be? Explain your choice clearly.

2. This problem explores and idea known as "the wisdom of crowds." A clear jar contains some unknown number of beans. 100 randomly chosen people inspect the jar and make estimates, $x_1, \ldots, x_{100}$ of the true number of beans, x_{true}, with no knowledge of the others' estimates. The sample average, $\bar{x} = \frac{1}{100}\sum_{i=1}^{100} x_i$, is computed.

   (a) Consider modeling $x_1, \ldots, x_{100}$ as a random sample from some distribution with mean μ and variance σ^2. Is this a reasonable model? Justify your answer clearly.

   (b) What is the approximate distribution of $\bar{x}$?

   (c) What is the approximate probability that $\bar{x}$ misses μ by more than 2 beans, assuming $\sigma = 10$?

   (d) What can you say about the actual error, $\bar{x} - x_{\text{true}}$. Make clear any assumptions you are making.

3. Consider the trees data set which you can import to R by

```
> data(‘‘trees’’)
```

This data set measures girth (circumference), height, and volume on a small collection of trees. Treat the data set as a matrix, using

```
> X = as.matrix(trees)
```

The features (colummns) can be centered to have mean 0 using

```
> X = scale(X,center=TRUE,scale=FALSE)
```

you may wish to use a pairs plot visualize these data.

   (a) Suppose we have data $(x_1, y_1), \ldots, (x_n, y_n)$. The *sample correlation* between two variables x and y, $r(x, y)$ is defined to be

$$r(x,y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

   It can be shown than $|r(x, y)| \leq 1$. Informally, the sample correlation measures the strength of the linear relationship between two variables. That is

   - $r(x, y) = 1$ means y can be predicted perfectly as $y_i = ax_i + b$ with $a > 0$
   - $r(x, y) = -1$ means y can be predicted perfectly as $y_i = ax_i + b$ with $a < 0$
   - $r(x, y) = 0$ means there is no formula $y_i = ax_i + b$ that predicts y better than simply choosing $a = 0$ (not using a).

   In the usual case where $|r(x, y)|$ is not exactly 0 or 1 it means that one variable can be partly predicted from the other using a linear model. More precisely, $r^2(x, y)$ is the percent of squared error explained by the best linear model. That is

$$r^2(x,y) = 1 - \frac{\sum_i (y_i - ax_i + b)^2}{\sum_i (y_i - \bar{y})^2}$$

   where a, b are the optimal line coefficients.

   Approximate all 9 possible correlations using the pairs plot.

   (b) Suppose we consider each pair, (v_1, v_2), of distinct variables and consider the usefulness of v_1 for making predications about v_2. Rank the pairs of variables from best to worst in terms of making these predictions, indicating ties where you believe they exist. Justify your answers.

   (c) **Using matrix calculations** compute the sample covariance matrix. Do not use an R package to do this.

4. In the “curse of dimensionality” R example done in class we let X be a d-dimensional vector of “runif” variables. This means the components of X are independent Unif(0,1) random variables. The R example computes the distance of X to the edge of the hypercube by

$$D = \frac{1}{2} - \max_{i=1\ldots d} |\frac{1}{2} - X_i|$$

   (a) The definition of independence says that for $X_1, \ldots, X_d$,

$$P(X_1 \in A_1, \ldots, X_d \in A_d) = \prod_{i=1}^{d} P(X_i \in A_i)$$

   where $X_1 \in A_1, \ldots, X_d \in A_d$ means $X_1 \in A_1$ *and* $X_2 \in A_2$ *and* $\ldots$, $X_d \in A_d$. What is the probability that a point is at least a away from an edge? That is what is $P(D > a)$ for $0 < a < \frac{1}{2}$.

   (b) Given any distance to edge δ, find d so that the probability of X lying within δ of the edge of the hypercube is at least .9. That is, find d so that $P(D < \delta) \geq .9$.

5. This problem from *Tan-Steinbach-Kumar, Ch. 2.* You are given a set of points S in Euclidean space, as well as the distance of each point in S to a point x. (It doesn't matter if $x \in S$).

   (a) If the goal is to find all points within a specified distance, ϵ, of point y, $y \neq x$, explain how you could use the triangle inequality and the already calculated distances to x to potentially reduce the number of distance calculations necessary? *Hint* The triangle inequality, $d(x, z) \leq d(x, y) + d(y, z)$ can be rewritten as $d(x, y) \geq d(x, z) - d(y, z)$.

   (b) In general, how would the distance between x and y affect the number of distance calculations?

   (c) Suppose that you can find a small subset of points S' from the original data set, such that every point in the data set is within a specified distance ϵ of at least one point in S', and that you also have the pairwise distance matrix for S'. Describe a technique that uses this information to compute, with a minumum of distance calculations, the set of all points within a distance of β of a specified point from the data set.