CS 613 - Machine Learning

Assignment 1 - Dimensionality Reduction & Clustering

Alex Lapinski

Fall 2016

10/01/2016

Part 1 - Answers to Theory Questions

1. Why do we like to use quadratic error functions (say over a 4th degree polynomial function) (2pts)?

   The primary reason for using a quadratic error function is that there can only be one maximum or minum.

2. Consider the following data:

$$\begin{bmatrix} -2 & 1 \\ -5 & -4 \\ -3 & 1 \\ 0 & 3 \\ -8 & 11 \\ -2 & 5 \\ 1 & 0 \\ 5 & -1 \\ -1 & -3 \\ 6 & 1 \end{bmatrix}$$

(a) Find the principle components of the data (you must show the math, including how you compute the eigenvectors and eigenvalues). Make sure you standardize the data first and that your principle components are normalized to be unit length (5pts).

$Mean_{col1} = \mu_1 = (-2 + -5 + -3 + 0 + -8 + -2 + 1 + 5 + -1 + 6)/10 = -0.9$
$Mean_{col2} = \mu_2 = (1 + -4 + 1 + 3 + 11 + 5 + 0 + -1 + -3 + 1)/10 = 1.4$
$Mean = \mu = \begin{bmatrix} -0.9 & 1.4 \end{bmatrix}$

$$CenteredData = RawData - \begin{bmatrix} -0.9 & 1.4 \end{bmatrix} CenteredData = \begin{bmatrix} -1.1 & -0.4 \\ -4.1 & -5.4 \\ -2.1 & -0.4 \\ -0.9 & 1.6 \\ -7.1 & 9.6 \\ -1.1 & 3.6 \\ 1.9 & -1.4 \\ 5.9 & -2.4 \\ -0.1 & -4.4 \\ 6.9 & -0.4 \end{bmatrix}$$

$\sigma_1 = \sqrt{(-1.1^2 + -4.1^2 + -2.1^2 + -0.9^2 + -7.1^2 + -1.1^2 + 1.9^2 + 5.9^2 + -0.1^2 + 6.9^2)/(10-1)}$
$\sigma_1 = \sqrt{160.9/9} = \sqrt{17.88} = 4.23$
$\sigma_2 = \sqrt{(-0.4^2 + -5.4^2 + -0.4^2 + 1.6^2 + 9.6^2 + 3.6^2 + -1.4^2 + -2.4^2 + -4.4^2 + -0.4^2)/(10-1)}$
$\sigma_2 = \sqrt{164.4/9} = \sqrt{18.27} = 4.27$
$StandardDeviation = \sigma = \begin{bmatrix} 4.23 & 4.27 \end{bmatrix}$

$$StandardizedData = CenteredData/\sigma = \begin{bmatrix} -0.26 & -0.09 \\ -0.97 & -1.26 \\ -0.50 & -0.09 \\ 0.21 & 0.37 \\ -1.68 & 2.25 \\ -0.26 & 0.84 \\ 0.45 & -0.33 \\ 1.39 & -0.56 \\ -0.02 & -1.03 \\ 1.63 & -0.09 \end{bmatrix}$$

Now that the data has been standardized by centering it and dividing it by the standard deviation for each column, we can compute the covariance matrix Σ.
$cov = \Sigma = E[(X_i - \mu_i)(X_j - \mu_j)]$
$\mu_1 = (-0.26 + -0.97 + -0.5 + 0.21 + -1.68 + -0.26 + 0.45 + 1.39 + -0.02 + 1.63)/10$
$\mu_1 = -0.01/10 = -0.001$
$\mu_2 = (-0.09 + -1.26 + -0.09 + 0.37 + 2.25 + 0.84 + -0.33 + -0.36 + -1.03 + -0.09)/10$
$\mu_2 = 0.01/10 = 0.001$

$$\Sigma = cov = \begin{bmatrix} cov_{11} & cov_{12} \\ cov_{21} & cov_{22} \end{bmatrix}$$
$cov_{11} = ((-0.26+0.001)^2 + (-0.97+0.001)^2 + (-0.5+0.001)^2 + (0.21+0.001)^2 + (-1.68+0.001)^2 + (-0.26+0.001)^2 + (0.45+0.001)^2 + (1.39+0.001)^2 + (-0.02+0.001)^2 + (1.63+0.001)^2)/N - 1$
$cov_{11} = (-0.0671 + 0.939 + 0.249 + 0.0445 + 2.819 + 0.0671 + 0.2034 + 1.9349 + 0.004 + 2.6602)/10 - 1$
$cov_{11} = (8.946/9) = 0.9983$

$cov_{12} = cov_{21} = ((-0.26 + 0.001)(-0.09 - 0.001) + (-0.97 + 0.001)(-1.26 - 0.001) + (-0.5+0.001)(-0.09-0.001)+(0.21+0.001)(0.37-0.001)+(-1.68+0.001)(2.25-0.001)+(-0.26 + 0.001)(0.84 - 0.001) + (0.45 + 0.001)(-0.33 - 0.001) + (1.39 + 0.001)(-0.36 - 0.001) + (-0.02 + 0.001)(-1.03 - 0.001) + (1.63 + 0.001)(-0.09 - 0.001))/N - 1$

$cov_{12} = cov_{21} = (0.0236+1.2219+0.0454+0.0779+-3.776+-0.217+-0.149+-0.780+0.02+-0.148)/10 - 1$
$cov_{12} = cov_{21} = (-3.681/9) = -0.409$

$cov_{22} = ((-0.09-0.001)^2 + (-1.26-0.001)^2 + (-0.09-0.001)^2 + (0.37-0.001)^2 + (2.25-0.001)^2 + (0.84-0.001)^2 + (-0.33-0.001)^2 + (-0.36-0.001)^2 + (-1.03-0.001)^2 + (-0.09-0.001)^2)/N - 1 cov_{22} = (0.01 + 1.59 + 0.14 + 5.06 + 0.7 + 0.11 + 0.31 + 1.06 + 0.01)/10 - 1$
$cov_{22} = (9.0/9) = 1.0$

$$\Sigma = \begin{bmatrix} cov_{11} & cov_{12} \\ cov_{21} & cov_{22} \end{bmatrix} = \begin{bmatrix} 0.9983 & -0.409 \\ -0.409 & 1.0 \end{bmatrix}$$

Now that we have the Covariance matrix, we plug it into the equation $\Sigma w = \alpha w$ and compute the eigen values and eigen vectors, where the eigen values will be α and the eigen vectors will be the vector w.

We'll set the equation $|\Sigma - \lambda I|$ equal to zero, since real eigen values only exist if this is equal to zero.

$$0 = |\Sigma - \lambda I|$$

$$0 = |\begin{bmatrix} 0.9983 & -0.409 \\ -0.409 & 1.0 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}|$$

$$0 = |\begin{bmatrix} 0.9983 - \lambda & -0.409 - 0 \\ -0.409 - 0 & 1.0 - \lambda \end{bmatrix}|$$

$$0 = (0.9983 - \lambda)(1.0 - \lambda) - (-0.409)(-0.409)$$

$$0 = 0.9983 - 1.9983\lambda + \lambda^2 - 0.167$$
$$0 = \lambda^2 - 1.9983\lambda + 0.8313$$
$$\lambda = \frac{1.9983 \pm \sqrt{-1.9983^2 - 4*1*0.8313}}{2*1}$$
$$\lambda = \frac{1.9983 \pm \sqrt{0.82}}{2}$$

Now we can plug one of the values of lambda into the equation to get our eigen vectors.

$$\left[\begin{bmatrix} 0.9983 & -0.409 \\ -0.409 & 1 \end{bmatrix} - \begin{bmatrix} \frac{1.9983+\sqrt{0.82}}{2} & 0 \\ 0 & \frac{1.9983+\sqrt{0.82}}{2} \end{bmatrix}\right] \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$
$$\left[\begin{bmatrix} 0.9983 & -0.409 \\ -0.409 & 1 \end{bmatrix} - \begin{bmatrix} 1.45 & 0 \\ 0 & 1.45 \end{bmatrix}\right] \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$
$$\begin{bmatrix} 0.9983 - 1.45 & -0.409 - 0 \\ -0.409 - 0 & 1 - 1.45 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

We then write out one of the equations to solve for y. $-0.4517x + -0.409y = 0$
$y = \frac{0.4517x}{0.409}$
We can then place any value for x, say 1 and get a vector.
$$e_1 = \begin{bmatrix} 1 \\ \frac{0.4517}{0.409} \end{bmatrix}$$
Then we'll standardize this value. $\frac{1}{2.21} \begin{bmatrix} 1 \\ 1.10 \end{bmatrix} = \begin{bmatrix} 0.453 \\ 0.498 \end{bmatrix}$

The Final Principal component is (The eigen vector associated with the highest eigen value):
$$\begin{bmatrix} 0.453 \\ 0.498 \end{bmatrix}$$

(b) Project the data onto the principal component corresponding to the largest eigenvalue found in the previous part (3pts).

$StandardizedData * ProjectionMatrix = 1DArray$

$$\begin{bmatrix} -0.26 & -0.09 \\ -0.97 & -1.26 \\ -0.50 & -0.09 \\ 0.21 & 0.37 \\ -1.68 & 2.25 \\ -0.26 & 0.84 \\ 0.45 & -0.33 \\ 1.39 & -0.56 \\ -0.02 & -1.03 \\ 1.63 & -0.09 \end{bmatrix} \begin{bmatrix} 0.453 \\ 0.498 \end{bmatrix} = \begin{bmatrix} -0.16 \\ -1.07 \\ -0.27 \\ 0.28 \\ 0.36 \\ 0.30 \\ 0.04 \\ 0.35 \\ -0.52 \\ 0.69 \end{bmatrix}$$

3. Consider the following data:

$$\text{Class 1} = \begin{bmatrix} -2 & 1 \\ -5 & -4 \\ -3 & 1 \\ 0 & 3 \\ -8 & 11 \end{bmatrix}, \text{Class 2} = \begin{bmatrix} -2 & 5 \\ 1 & 0 \\ 5 & -1 \\ -1 & -3 \\ 6 & 1 \end{bmatrix}$$

(a) Compute the information gain for each feature. You could standardize the data overall, although it won't make a difference. (5pts).

$Feature_1$ = column 1
$Feature_2$ = column 2

p = number of samples in class 1 = 5 total
n = number of samples in class 2 = 5 total

Initial Entropy = $H(\frac{5}{10}, \frac{5}{10}) = -\frac{5}{10}log_2(\frac{5}{10}) + -\frac{5}{10}log_2(\frac{5}{10}) = 1$

$remainder(Feature_1) = \frac{1+1}{5+5}(-\frac{1}{5}log_2(\frac{1}{5})-\frac{1}{5}log_2(\frac{1}{5})+\frac{1+0}{5+5}(-\frac{1}{5}log_2(\frac{1}{5})-\frac{0}{5}log_2(\frac{0}{5})+\frac{1+0}{5+5}(-\frac{1}{5}log_2(\frac{1}{5})-\frac{0}{5}log_2(\frac{0}{5})+\frac{1+0}{5+5}(-\frac{1}{5}log_2(\frac{1}{5})-\frac{0}{5}log_2(\frac{0}{5})+\frac{0+1}{5+5}(-\frac{0}{5}log_2(\frac{0}{5})-\frac{1}{5}log_2(\frac{1}{5})+\frac{0+1}{5+5}(-\frac{0}{5}log_2(\frac{0}{5}) - \frac{1}{5}log_2(\frac{1}{5}) + \frac{0+1}{5+5}(-\frac{0}{5}log_2(\frac{0}{5}) - \frac{1}{5}log_2(\frac{1}{5}) + \frac{0+1}{5+5}(-\frac{0}{5}log_2(\frac{0}{5}) - \frac{1}{5}log_2(\frac{1}{5})$

$remainder(Feature_1) = \frac{2}{10}(0.322 - 0.322) + \frac{1}{10}(0.322 - 0) + \frac{1}{10}(0.322 - 0) + \frac{1}{10}(0.322 - 0) + \frac{1}{10}(0.322 - 0) + \frac{1}{10}(0 - 0.322) + \frac{1}{10}(0 - 0.322) + \frac{1}{10}(0 - 0.322) + \frac{1}{10}(0 - 0.322)$

$remainder(Feature_1) = \frac{2}{10}(0) + \frac{4}{10}(0.322) + \frac{4}{10}(-0.322) = 0$

$IG(Feature_1) = 1.0 - 0 = 1.0(Noinformationgain)$

$remainder(Feature_2) = \frac{2+1}{5+5}(-\frac{2}{5}log_2(\frac{2}{5})-\frac{1}{5}log_2(\frac{1}{5}))+\frac{1+0}{5+5}(-\frac{1}{5}log_2(\frac{1}{5})-\frac{0}{5}log_2(\frac{0}{5}))+\frac{1+0}{5+5}(-\frac{1}{5}log_2(\frac{1}{5})-\frac{0}{5}log_2(\frac{0}{5})) + \frac{1+0}{5+5}(-\frac{1}{5}log_2(\frac{1}{5}) - \frac{0}{5}log_2(\frac{0}{5})) + \frac{0+1}{5+5}(-\frac{0}{5}log_2(\frac{0}{5}) - \frac{1}{5}log_2(\frac{1}{5})) + \frac{0+1}{5+5}(-\frac{0}{5}log_2(\frac{0}{5}) - \frac{1}{5}log_2(\frac{1}{5})) + \frac{0+1}{5+5}(-\frac{0}{5}log_2(\frac{0}{5}) - \frac{1}{5}log_2(\frac{1}{5})) + \frac{0+1}{5+5}(-\frac{0}{5}log_2(\frac{0}{5}) - \frac{1}{5}log_2(\frac{1}{5}))$

$remainder(Feature_2) = \frac{3}{10}(0.23 - 0.322) + \frac{1}{10}(0.322 - 0) + \frac{1}{10}(0.322 - 0) + \frac{1}{10}(0.322 - 0) + \frac{1}{10}(0 - 0.322) + \frac{1}{10}(0 - 0.322) + \frac{1}{10}(0 - 0.322) + \frac{1}{10}(0 - 0.322)$

$remainder(Feature_2) = \frac{3}{10}(-0.092) + \frac{7}{10}(0.322) = -0.0276 + 0.2254 = 0.1978$
$IG(Feature_2) = 1.0 - 0.1978 = 0.8022$

(b) Which feature is more discriminating based on results in part a (1pt)?

We would get a higher information gain with $Feature_2$ (Column 2). Thus $Feature_2$ is more discriminating.

(c) Using LDA, find the direction of projection (you must show the math). Normalize this vector to be unit length.
Note: You don't not have to standardize the data since your computations should take into account the mean and standard deviations of the classes separately. (5pts).

(d) Project the data onto the principal component found in the previous part (3pts).

$$StandardizedClass_1 = (\begin{bmatrix} -2 & 1 \\ -5 & -4 \\ -3 & 1 \\ 0 & 3 \\ -8 & 11 \end{bmatrix} - \begin{bmatrix} -0.9 & 1.4 \end{bmatrix}) / \begin{bmatrix} 4.23 & 4.27 \end{bmatrix} = \begin{bmatrix} -0.26 & -0.09 \\ -0.97 & -1.26 \\ -0.50 & -0.09 \\ 0.21 & 0.37 \\ -1.68 & 2.25 \end{bmatrix}$$

$$StandardizedClass_2 = (\begin{bmatrix} -2 & 5 \\ 1 & 0 \\ 5 & -1 \\ -1 & -3 \\ 6 & 1 \end{bmatrix} - \begin{bmatrix} -0.9 & 1.4 \end{bmatrix}) / \begin{bmatrix} 4.23 & 4.27 \end{bmatrix} = \begin{bmatrix} -0.26 & 0.84 \\ 0.45 & -0.33 \\ 1.39 & -0.56 \\ 0.21 & 0.37 \\ -1.68 & 2.25 \end{bmatrix}$$

$$Class_1 = \begin{bmatrix} -0.26 & -0.09 \\ -0.97 & -1.26 \\ -0.50 & -0.09 \\ 0.21 & 0.37 \\ -1.68 & 2.25 \end{bmatrix} \begin{bmatrix} 0.453 \\ 0.498 \end{bmatrix} = \begin{bmatrix} -0.16 \\ -1.07 \\ -0.27 \\ 0.28 \\ 0.36 \end{bmatrix}$$

$$Class_2 = \begin{bmatrix} -0.26 & 0.84 \\ 0.45 & -0.33 \\ 1.39 & -0.56 \\ 0.21 & 0.37 \\ -1.68 & 2.25 \end{bmatrix} \begin{bmatrix} 0.453 \\ 0.498 \end{bmatrix} = \begin{bmatrix} 0.30 \\ 0.04 \\ 0.35 \\ -0.52 \\ 0.69 \end{bmatrix}$$

(e) Does the projection you performed in the previous part seem to provide good class separation? Why or why not (1pt)?

Yes, this does provide a very good separation. The main reason behind this is that Class 1 ends up being the top half of the projection from the previous question. Then Class 2 is at the bottom of the projection from the earlier question.
This provides a clean separation of the projection.

Part 2 - PCA Result

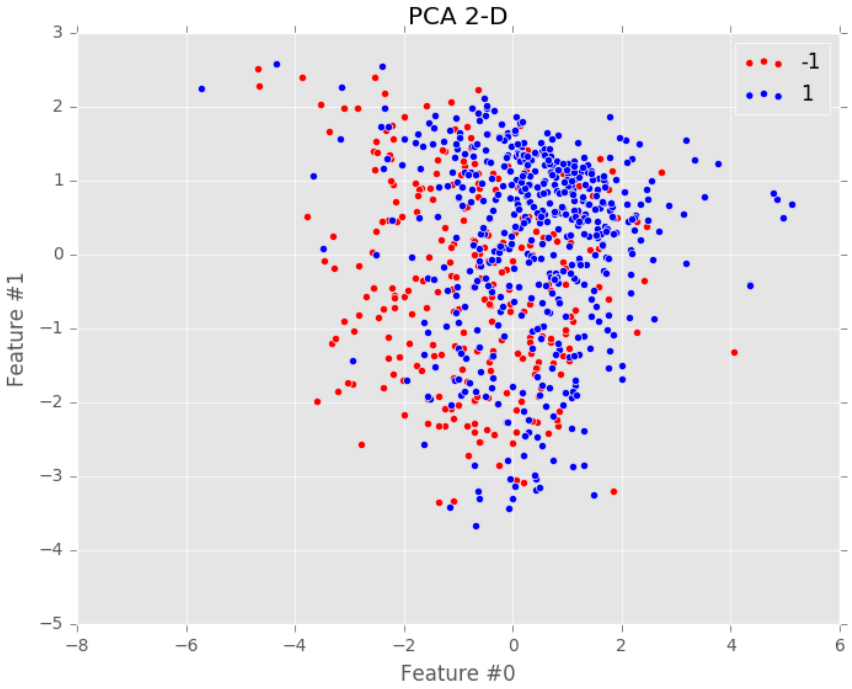

Figure 1: PCA 2D Feature 0 by Feature 1

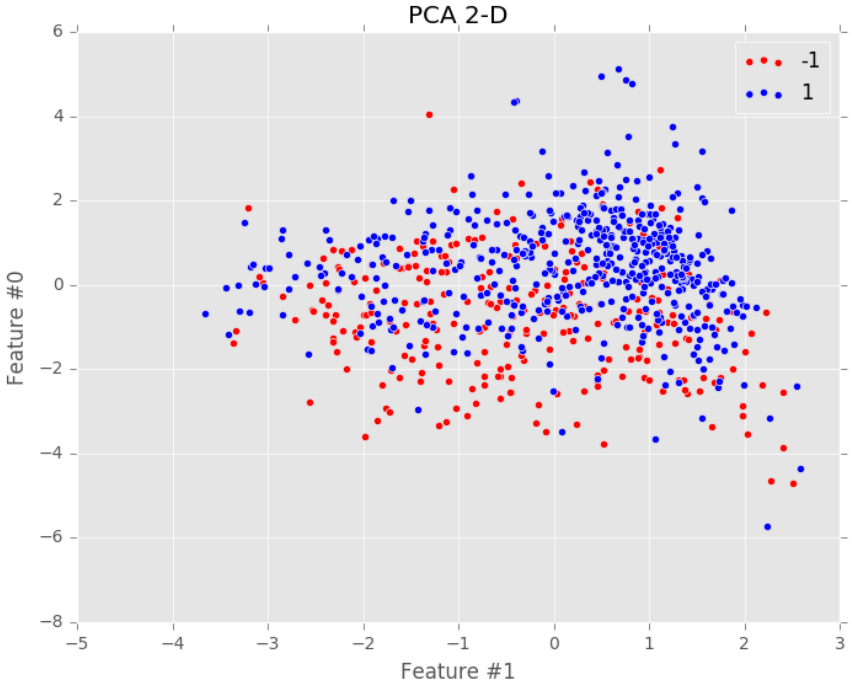

Figure 2: PCA 2D Feature 1 by Feature 0

Part 3 - Visualization of k-means

Initial Setup

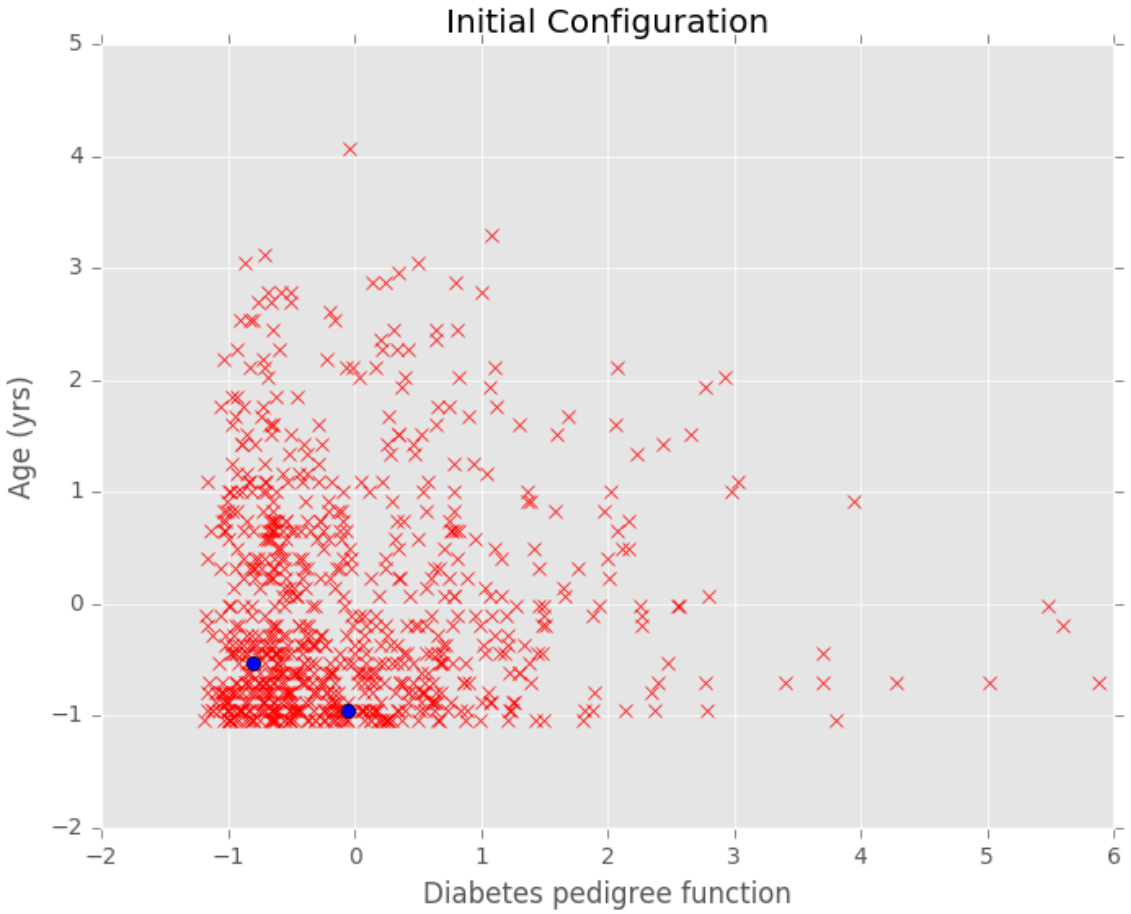

Figure 3: K-Means Initial Configuration

Initial Cluster Assignment

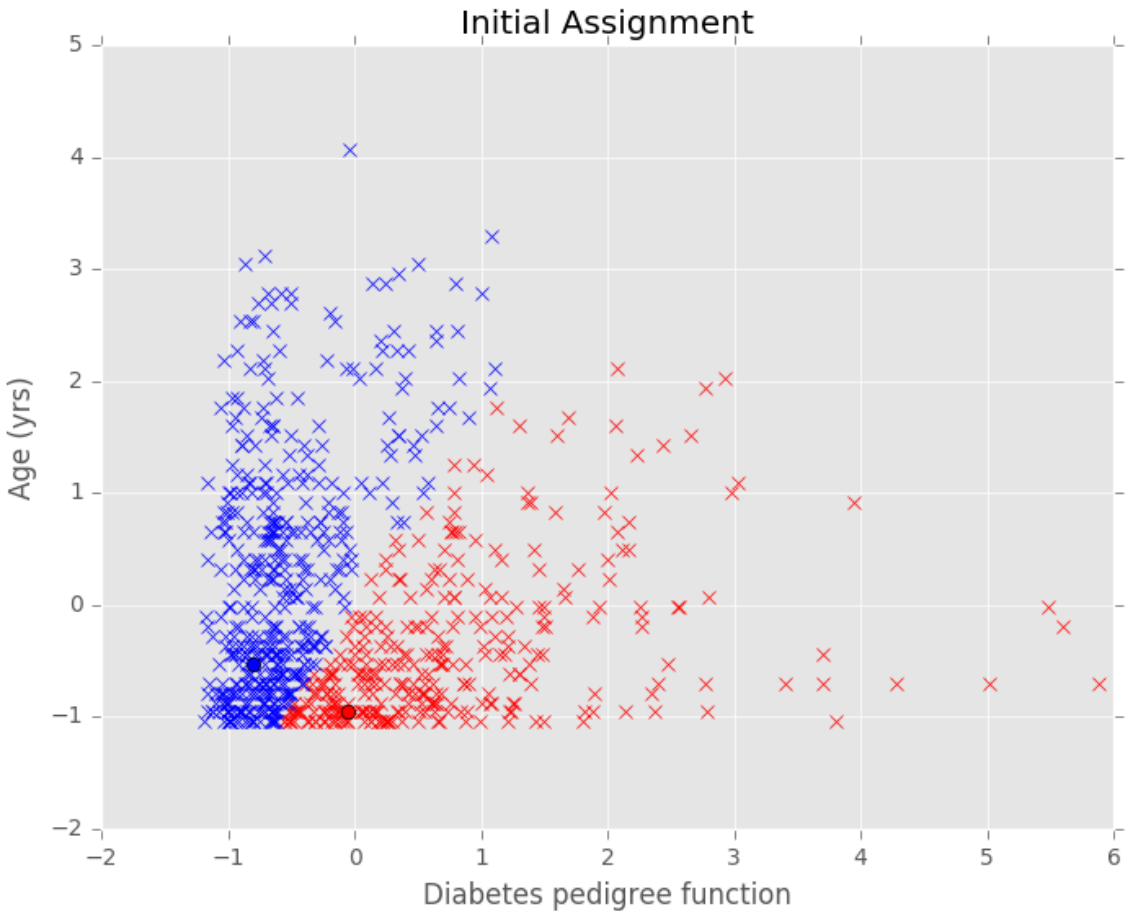

Figure 4: K-Means Initial Assignment

Final Cluster Assignment

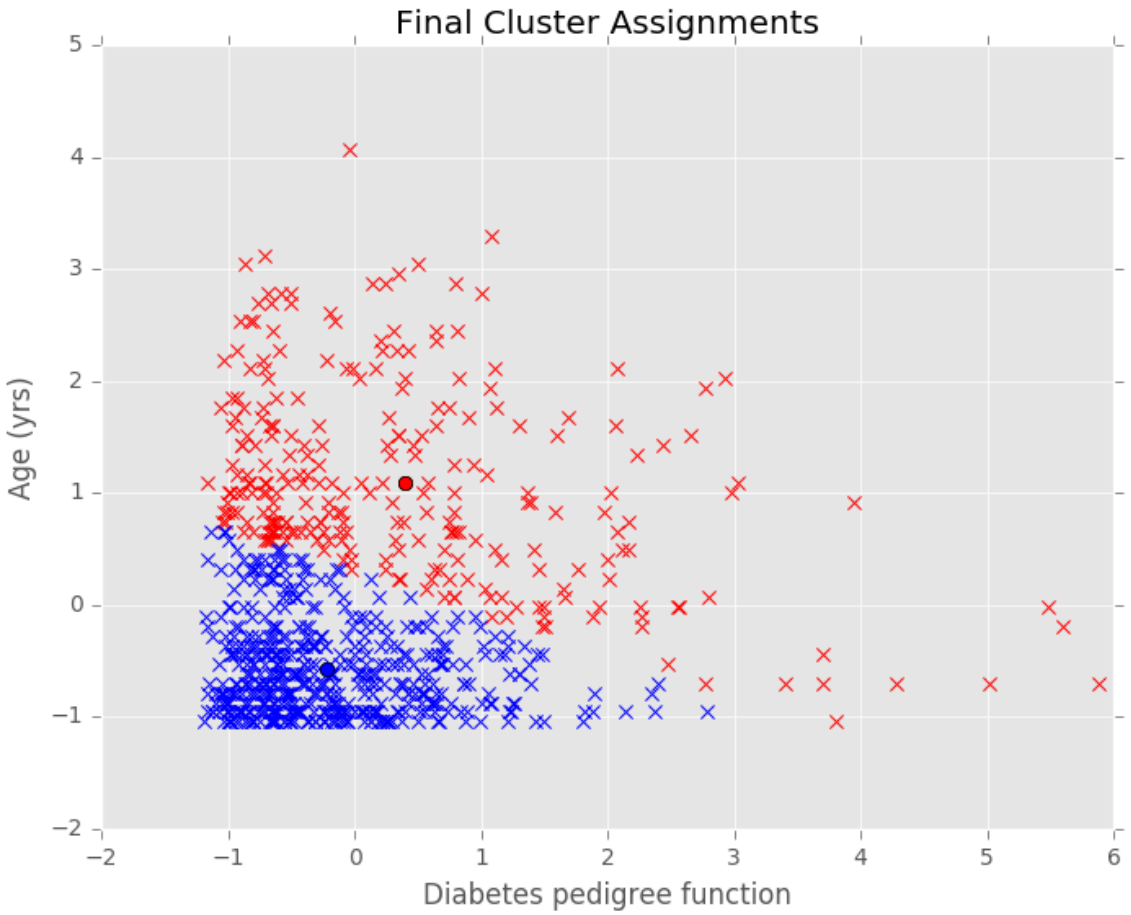

Figure 5: K-Means Final Assignment

Results

The K-Means implementation took 41 iterations to run to completion.

I ran about 3 different trials, all seemed to take 41 iterations.