CS-482 Machine Learning

Feature Extraction and Dimensionality Reduction

Class Work

Given the following dataset on class grades on test1 and test 2, compute PCA with two components by computing
a) the covariance matrix A,
b) finding two eigen values using $|A-\lambda I|=0$,
c) computing eigenvectors by solving $Av = \lambda v$, for each λ
d) Determine the contribution of each eigenvalue to the principal component.
e) If there is one that is above 90% then use one component only, otherwise use two components.
f) Create the feature vector by using one or two eigen vectors.
g) Translate the original data (after standardization) to the new component by using the following transformation

$$\text{Final\_Data} = (\text{feature vector})^T * (\text{original\_data\_standardized})^T$$

Test1
9
10
10
10
9
10
8
9.5
9
10
9

Test 2
8
8.5
10

9.5
7.5
8.5
8
8.5
10
7.5
8

1. Given the matrix X whose rows represent different data points, you are asked to perform a k-means clustering on this dataset using the Euclidean distance as the distance function. Here k is chosen as 3. All data in X were plotted in Figure 1 below . The centers of 3 clusters were initialized as c_1= (6:2; 3:2) (red), c_2= (6:6; 3:7) (green), c_3= (6:5; 3:0) (blue).

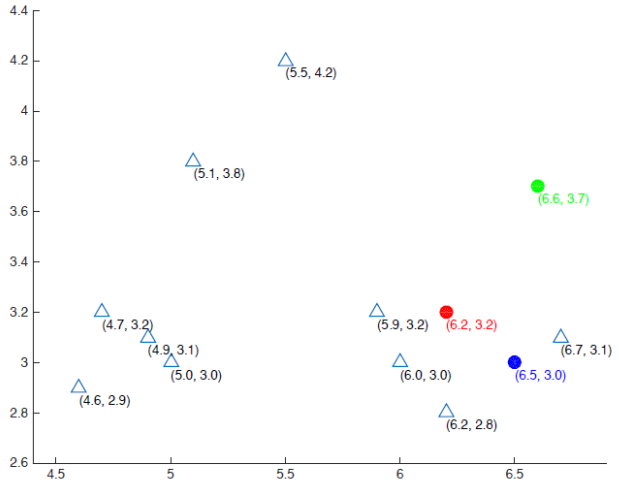

Figure 1: Scatter plot of datasets and the initialized centers of 3 clusters

$$\mathbf{X} = \begin{bmatrix} 5.9 & 3.2 \\ 4.6 & 2.9 \\ 6.2 & 2.8 \\ 4.7 & 3.2 \\ 5.5 & 4.2 \\ 5.0 & 3.0 \\ 4.9 & 3.1 \\ 6.7 & 3.1 \\ 5.1 & 3.8 \\ 6.0 & 3.0 \end{bmatrix}$$

1. What's the center of the first cluster (red) after one iteration? (Answer in the format of [x1, x2], round your results to three decimal places, same as problems 2 and 3)
2. What's the center of the second cluster (green) after two iteration?
3. What's the center of the third cluster (blue) when the clustering converges?