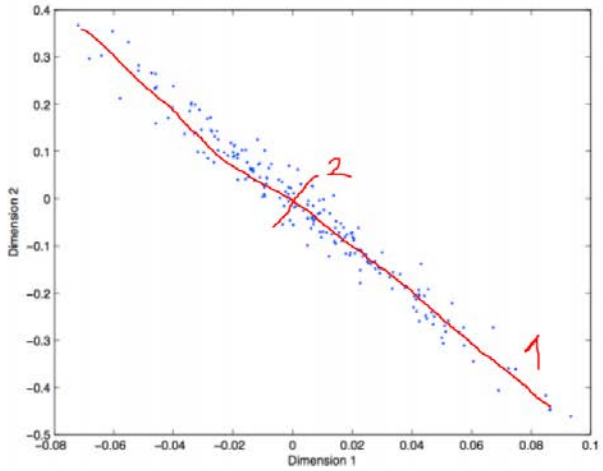
1. You have to communicate a signal in a language that has 3 symbols A, B and C. The probability of observing A is 50% while that of observing B and C is 25% each. Design an appropriate encoding for this language. What is the entropy of this signal in bits?

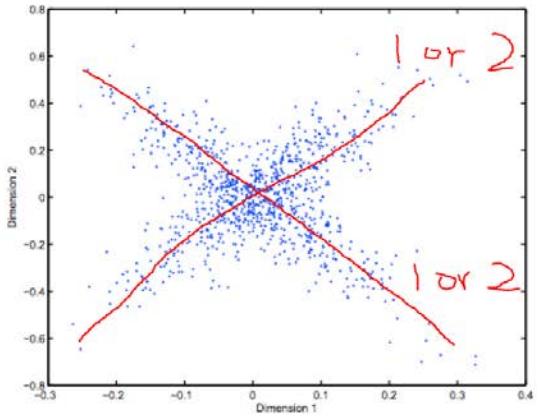
The bit string for A is 0, for B is 10 and for C is 11. The entropy of this coding is – 0.5*log2(0.5) – 0.25*log2(0.25) – 0.25*log2(0.25) = 1.5

2. Show that the Kmeans procedure can be viewed as a special case of the EM algorithm applied to an appropriate mixture of Gaussian densities model.

K means can be seen as a EM algorithm that require the Gaussian densities to be spherical for all features, and in the maximization step a mean of data points was used to update the centroid parameter. The centroid location is the only parameter to be estimated, no covariance or slope for any features.

3. Plot the direction of the first and second PCA components in the figures given.




4. Which clustering method(s) is most likely to produce the following results at k = 2?

a. Hierarchical clustering will perform best for this data set. Because the distances between points in two clusters have a very sparse distribution. The K means method will have a hard time find the right location of centers, and it's also very hard to assign the cluster for the points that are very close but came from different classes. EM will also suffer from the same problem as K means. And the single link hierarchical clustering can identify the two classes perfectly since there will always be a very close point pair within the two classes. Maybe the other two hierarchical clustering can also find optimal classes, they are much slower than single link clustering.

b. Both EM and K means can perform very well for this data set. Because it looks like we only need to estimate two centroid locations for both methods. Given the complexity of Gaussian distribution and posterior probability, I think K means will converge faster than EM. All the hierarchical clustering methods will have problem distinguish the point that are very close but comes from two classes.

c. EM will have the best performance for this situation. Because there are points that are very close to each other but comes from different classes. The hierarchical clustering can't distinguish those points. Also, the centroid setting means K means can't use centroid locations to cluster the two different classes. Though EM also won't work very well for the overlapped points, it can distinguish the outside ones.