Problem Set 2 – Michael Lukacsko

Question 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?

Answer: To design an appropriate encoding for this language, a variable-length code where symbols with higher probabilities are assigned shorter codes, and symbols with lower probabilities are assigned longer codes can be used. One possible encoding for this language could be A=0, B=10, C=11. By using this encoding, a shorter code is assigned to the more frequent symbol A and longer code is assigned to the less frequent symbols B and C.

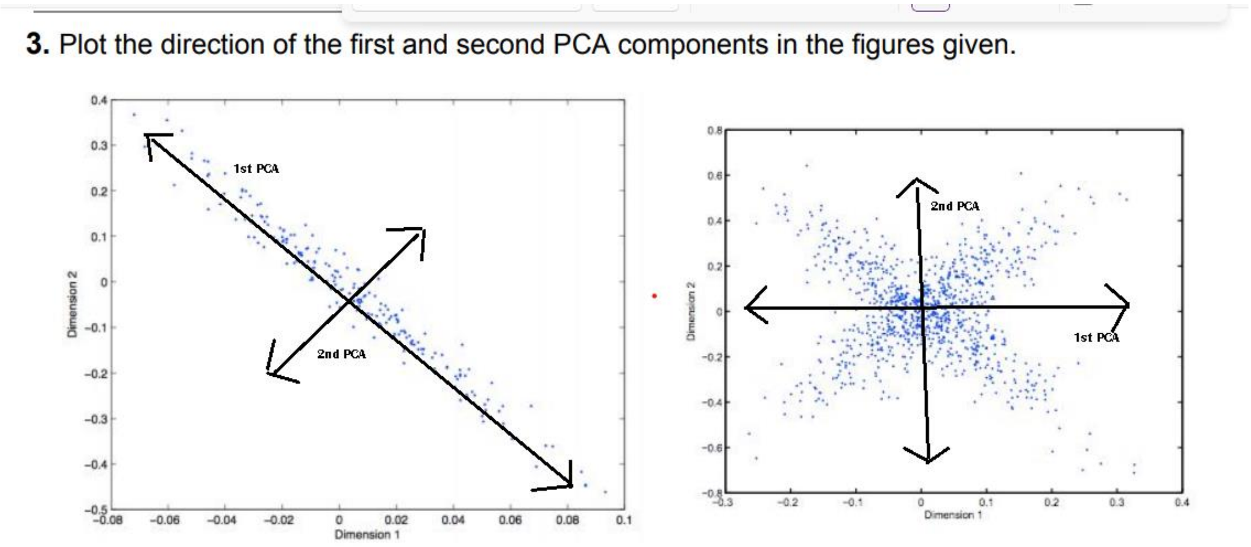
The entropy of the signal is calculated as follows:
H = - sum_i p_i log2(p_i) where p_i is the probability of symbol i. Using the probabilities provided,
= - (0.5 log2(0.5) + 0.25 log2(0.25) + 0.25 log2(0.25))
= - (0.5 * (-1) + 0.25 * (-2) + 0.25 * (-2))
= 1.5 bits
Hence, the entropy of the signal is 1.5 bits, which is the expected number of bits required to communicate one symbol using the encoding.

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

Answer: K-means is a clustering algorithm that groups data points into a number of clusters K by minimizing the sum of squared distances between each point and its cluster centroid. This can be considered as a special case of the EM algorithm, which is a more general algorithm for finding the best estimates of model parameters given some observed data. More specifically, K-means can be seen as a simplified version of the EM algorithm applied to a mixture of Gaussian densities model, where each Gaussian represents a cluster and its mean is the centroid of the corresponding cluster. The EM algorithm consists of two steps: the E step, which computes the expected probability of each data point belonging to each cluster given the current model parameters, and the M step, which updates the model parameters based on the expected probabilities computed in the E step. In the case of K-means, the E step corresponds to assigning each data point to its nearest centroid, and the M step corresponds to updating the centroids based on the assigned data points.

<u>Question 3</u>: Plot the direction of the first and second PCA components in the figures given.

<u>Answer</u>:



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