HW #6 Due: 5/28/2021

For problem 4 and 5, you can use whatever packages you are familiar with to complete these problems. If you have no preference, you may try sci-kit learn.

1. We have a dataset $S = \left\{\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}2\\2\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix}, \begin{bmatrix}1\\1\end{bmatrix}\right\}$. Follow the k-means algorithm to complete the assignment step and the update step in one iteration loop. Use k = 2 and initial conditions $\boldsymbol{\mu}_1 = \begin{bmatrix}-1\\-1\end{bmatrix}$ and $\boldsymbol{\mu}_2 = \begin{bmatrix}3\\3\end{bmatrix}$ in the computation.
2. Analytically compute the first principal component of the following data points:
$$\boldsymbol{x}_1 = \begin{bmatrix}1\\1\end{bmatrix},\ \boldsymbol{x}_2 = \begin{bmatrix}4\\4\end{bmatrix},\ \boldsymbol{x}_3 = \begin{bmatrix}5\\5\end{bmatrix}.$$
3. In the breast cancer dataset (used in HW 2), if the LDA is to be used for dimensionality reduction, what is the maximum number of feature dimensions after reduction? Why?
4. Use the breast cancer dataset (used in HW 2) to examine the accuracy vs number of features by PCA.
   a. How many components are necessary to ensure Pov(k) > 0.9?
   b. Set principal components from 1 to 9 and observe the change of accuracy. As usual, use 70/30 split and average 10 times to report the accuracy. When computing principal components, remember to use only training set. However, you also need to transform test samples to dimension-reduced space for testing. Use SVM with rbf (radical-basis function) kernel and default parameters as the classifier.
5. Use ICA for blind source separation with the accompanying audio files (org_1.wav & org_2.wav). Repeat the experiment with PCA and check if PCA can also separate the sources.