HW #4 Due: 5/10/2023

1. Suppose that we have a dataset containing samples $[x_1 \quad x_2]^T$ from jointly Gaussian random variables with $\mu = \begin{bmatrix} 2 \\ 4 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. Is it possible to reduce these two attributes into one by using PCA? Use $\text{PoV}(1) > 0.9$ as the criteria to answer this question. Can this question be answered without computing eigenvalues? If so, how?
2. If we want to use the original ICA algorithm for dimensionality reduction, can we directly pick independent components with larger energy? Explain.
3. You are asked to use the LDA to reduce the dimensionality of the breast cancer dataset before classification. If only "benign" or "malignant" is to be determined, what is the highest dimension of features after LDA reduction?
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 the random forest classifier with default parameters from sklearn for this problem.
5. Use the factor analysis to reduce the feature dimension from 4 to 3 for Iris data set. As usual, take 70% of the samples as training set to perform FA. Use 5-NN to classify the test set and then report the average accuracy after 10 trials. For simplicity, you may assume $\Psi = 0$ and use the pseudo inverse solution.