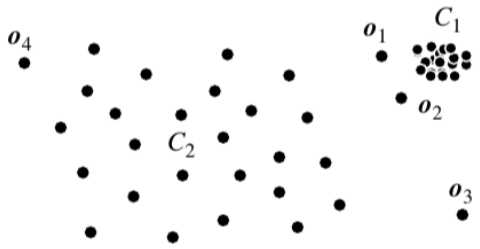
HW #3 Due: 4/11/2022

1. If we know the distributions of the samples are given below. Suppose that C_1 and C_2 are cluster centers with known respective covariance values (estimated from neighboring points on the plot). To detect outliers o_1 and o_2, of the Euclidean distance and the Mahalanobis distance, which one is better? Why?



2. Follow the numerical example in GMM and complete the computation of $\mu_2,\ \sigma_1^2, \sigma_2^2, \alpha_1,$ and α_2 in one step.
3. Use the three-urn example (on pp. 13 of the PPT file) to find the optimal state sequence (decoding problem) with the associated path probability if the observation sequence is "RBR." Use the number of red and blue balls in each urn to compute the emission probability for each state. The transition probability $A = \begin{bmatrix} \frac{2}{3} & \frac{1}{6} & \frac{1}{6} \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{bmatrix}$ and initial state distribution $\pi = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$.
4. Repeat problem 5 of HW 2, but use GMM with 2 mixtures instead. The GMM tools are supported in sklearn. Remember to use one model per class. Of the k-NN, Naïve Bayes, and GMM classifiers, which one has the best accuracy?
5. In this problem, you are asked to perform the wrapper-type feature selection using the Naïve Bayes classifier for cancer dataset (Breast Cancer Wisconsin (Original) Data Set, directly from the sklearn or downloading from https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29). To simplify the problem, we just want to keep 5 attributes out of 9 (hint: one attribute is useless. Which one is it?) To begin one experiment, randomly draw 60 % of the instances from each class for training, and 20% from each class for finding the best 5 attributes. Once the feature selection is complete, use the rest 20% for testing to obtain the accuracy. Repeat the selection 10 times to get the average accuracy. You need to deal with **missing attributes.** Compare the obtained accuracy with the same type of model trained with the full set of 9 features.