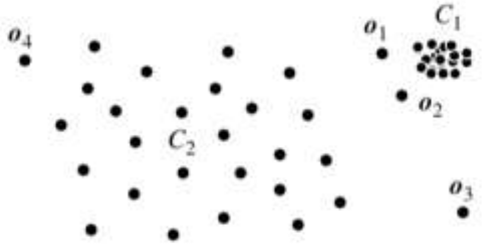
HW #3 Due: 4/12/2023

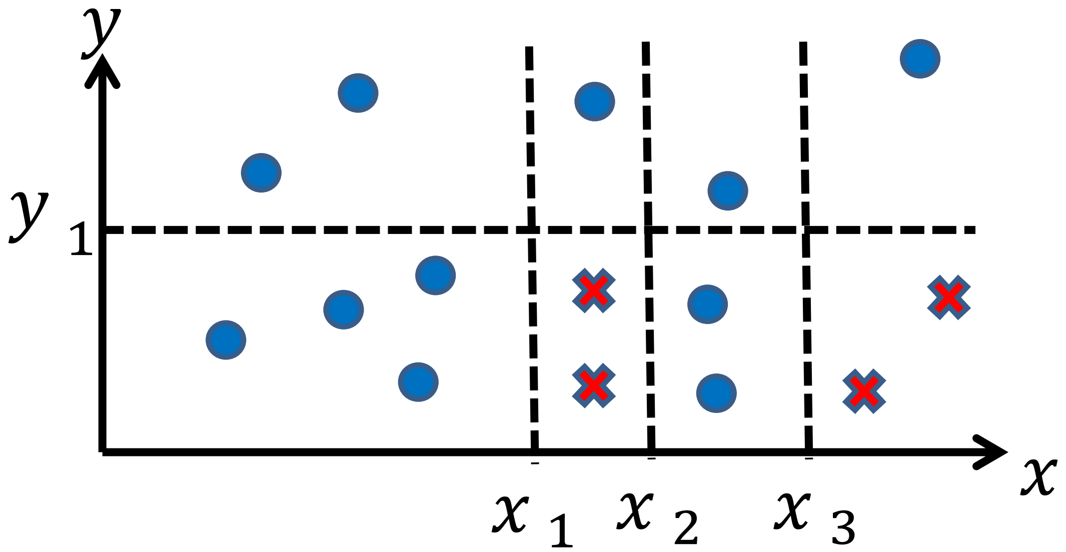
1. If we know the distributions of the samples are given below. Suppose that C_1 and C_2 are cluster centers with known respective covariance values (estimated from neighboring points on the plot). To detect outliers o_1 and o_2, of the Euclidean distance and the Mahalanobis distance, which one is better?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

$\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$



$(x-\mu)^T\Sigma^{-1}(x-\mu)$

2. Plot a decision tree for the following data points. You just need to use one “>” or “<” in a vertex.



3. Follow the numerical example in GMM and complete the computation of μ_2, $\sigma_1^2, \sigma_2^2, \alpha_1$, and α_2 in one step.
4. Repeat the classification of the Iris dataset, but use GMM with 2 mixtures instead. The GMM tools are supported in sklearn. Remember to use one model per class. Use the typical 70/30 train/test split.
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 3 attributes out of 9. To begin one experiment, randomly draw 60 % of the instances from each class for

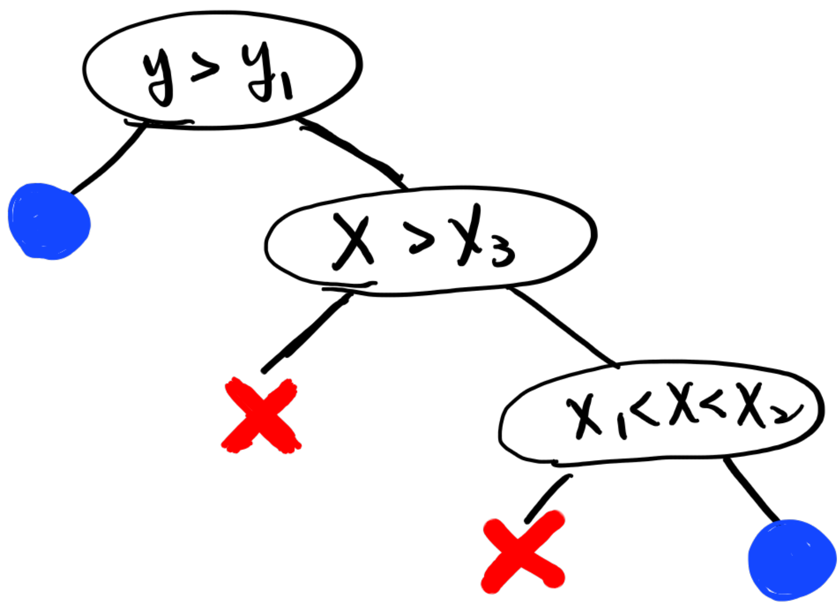
training, and 20% from each class for finding the best 3 attributes. Once the feature selection is complete, use the rest 20% for testing to obtain the accuracy. Repeat the selection 10 times to get the average accuracy. Compare the obtained accuracy with the same type of model trained with the full set of 9 features.

1. Mahalanobis' Method is better due to the formula:

$$(x-\mu)^T \Sigma^{-1} (x-\mu)$$

consider the all distributed points of the same group, so it find outliers more objectly.

2. Decision Tree of the graph.



3. one-dimensional data

$[0.9, 0.7, 1.2, 2.4, 1.8]$

suppose

$\mu_1 = 1, \mu_2 = 2$

$\sigma_1^2 = \sigma_2^2 = 1$

$\alpha_1 = \alpha_2 = 0.5$

Start from $\beta_j(x) = p(j|x) = \frac{\alpha_j g_j(x)}{\sum_{k=1}^{2} \alpha_k g_k(x)}$

$\because g(x)$

(pdf of normal distribution) (univariate Gaussian)

$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

Put our assumption in $g_1(x)$, $g_2(x)$.

$g_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-1}{1}\right)^2}$

$g_2(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-2}{1}\right)^2}$

For $x = 0.9, 0.7, 1.2, 2.4, 1.8$.

we have

$g_1(x) = 0.3970 \quad 0.3814 \quad 0.3910 \quad 0.1497 \quad 0.2897$

$g_2(x) = 0.2179 \quad 0.1714 \quad 0.2897 \quad 0.3683 \quad 0.3910$

than is to compute $\beta_j(x) = P(j|x)$

$$= \frac{\alpha_j g_j(x)}{\sum_{k=1}^{2} \alpha_k g_k(x)}$$

and $\alpha_1 = \alpha_2 = 0.5$

therefore, $\beta_1(x) = \frac{g_1(x)}{g_1(x)+g_2(x)}$

$$\beta_2(x) = \frac{g_2(x)}{g_1(x)+g_2(x)}.$$

$\beta_1(x) =$ 0.6457 0.6900 0.5744 0.2891 0.4256

$\beta_2(x) =$ 0.3543 0.3100 0.4256 0.7109 0.5744

$x =$ 0.9 0.7 1.2 2.4 1.8

update method of

$$\mu_j = \frac{\sum_{i=0}^{n} \beta_j x_i}{\sum_{i=0}^{n} \beta_j}$$

$$\sigma_j = \frac{\sum_{i=0}^{n} \beta_j (x_i - \mu_j)^2}{\sum_{i=0}^{n} \beta_j}$$

$$\alpha_j = \frac{1}{n} \sum_{i=0}^{n} \beta_j$$

So the new

$$\mu_1 = \frac{3.2131}{2.6247} = 1.2242$$

$$\mu_2 = \frac{3.7866}{2.3753} = 1.5941$$

$$\sigma_1^2 = \frac{0.7985}{2.6247} = 0.3042$$

$$\sigma_2^2 = \frac{0.9706}{2.3753} = 0.4086$$

$$\alpha_1 = \frac{2.6247}{5} = 0.5249$$

$$\alpha_2 = \frac{2.3753}{5} = 0.4751$$