1. Exercises

1.1 A simple example: two Gaussians

Question 1 *Obtain a line to classify the data by using what you know about the distributions of the data. In which sense is it optimal?*

As is mentioned, the artificial dataset consists of two classes of Gaussians, and they are denoted by $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ where $\mu_1 = (1,1)^T$, $\mu_2 = (-1,-1)^T$ and $\Sigma_1 = \Sigma_2 = \mathrm{diag}(1,1)$. The two classes are shown in Figure 1.

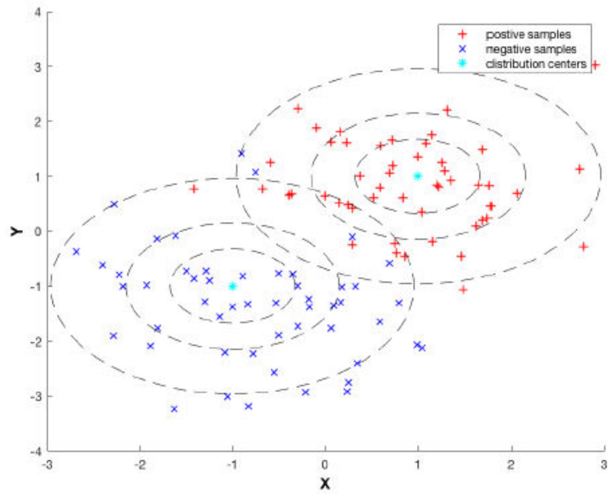

Figure 1: The dataset of the two classes of Gaussians with black dash lines denoted the confidence intervals of 50%, 75% and 95%.

Since we have already known the ground-truth distributions of the two classes, we can have a basic idea of classifying the new data $x \in \mathbb{R}^2$ via Euclidean distance. To be specific, if $\|x-\mu_1\|_2 < \|x-\mu_2\|_2$, then x belongs to class positive, while if $\|x-\mu_1\|_2 \geq \|x-\mu_2\|_2$, then we have x belongs to class negative. Then, the linear classifier can be a mid-perpendicular line connecting the two distribution centers, which is shown in Figure 2.

This linear classifier is optimal in the sense of Bayes decision rule. Recall that the Bayes rules give that $\mathrm{P}(C_i|x) \propto \mathrm{P}(x|C_i)\mathrm{P}(C_i)$, $i = 1,2$, $x \in \mathbb{R}^2$, then under the assumption that $\mathrm{P}(x|C_i)$ subjects to a normal distribution, e.g. $\mathcal{N}(\mu_i, \Sigma_i)$, we have

$$\log \mathrm{P}(C_i|x) \propto -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2}\log|\Sigma_i| + \log \mathrm{P}(C_i), \quad i = 1,2. \tag{1}$$

Since $\Sigma_1 = \Sigma_2 = \mathrm{diag}(1,1)$, we can further simplify (1) to

$$\log \mathrm{P}(C_i|x) \propto -\frac{1}{2}\|x-\mu_i\|_2^2 + \log \mathrm{P}(C_i), \quad i = 1,2.$$

If the prior class probabilities are equal, we consider the decision rule $d_i = -\frac{1}{2}\|x-\mu_i\|_2^2$ instead of $\log \mathrm{P}(C_i|x)$. More specifically, if $d_1 > d_2$, then we have x belongs to the class positive, and x belongs to the class negative for $d_1 \leq d_2$. This exactly coincides with how we derive the decision rule based on the Euclidean distance. This completes the explanation.

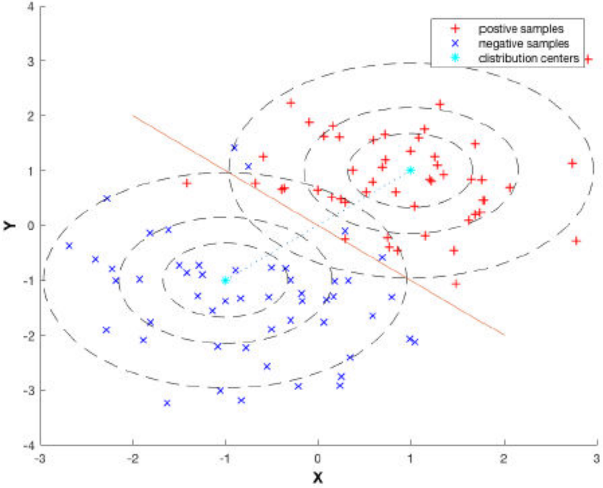

Figure 2: The linear classifier represented by the red line separates the two classes, which is $y = -x$.