Mistura de gaussianas

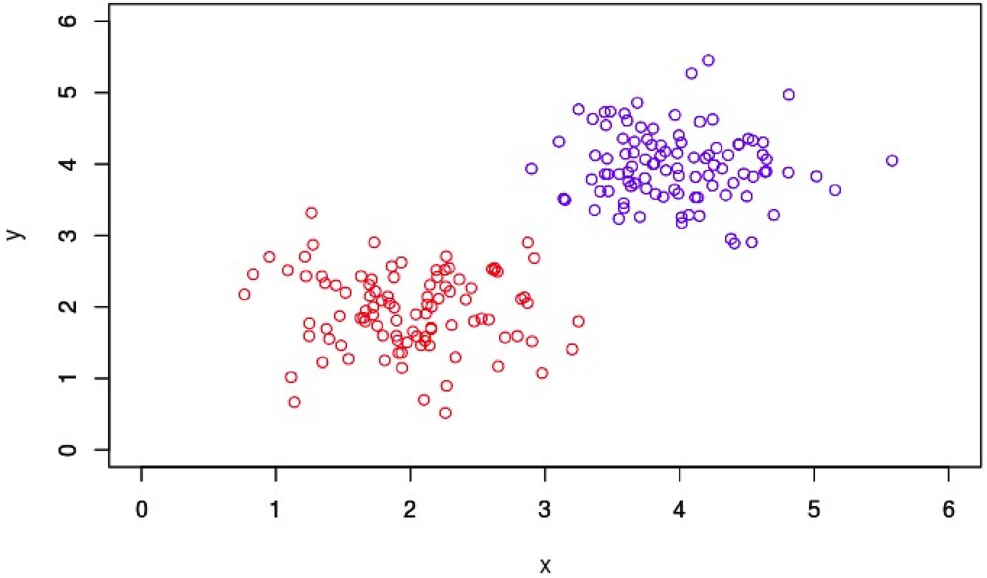
Qualidade de partições

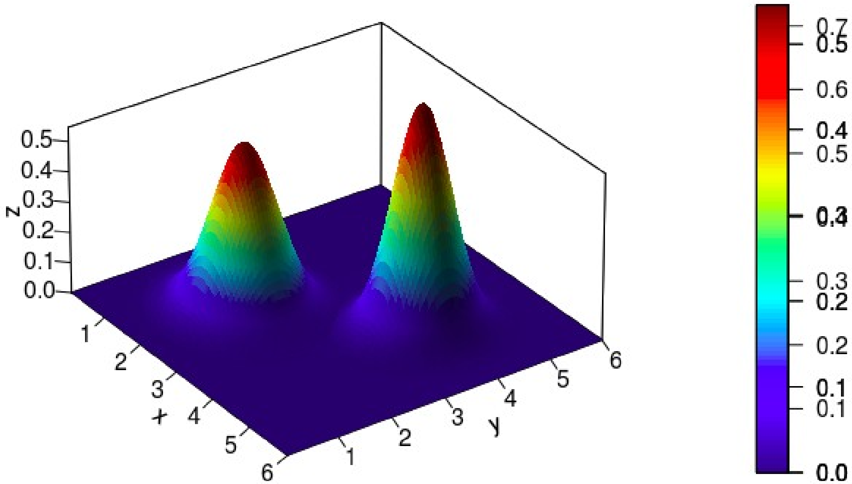
Prof. Frederico Coelho.

(Baseado no material de David Sontag)

Introdução

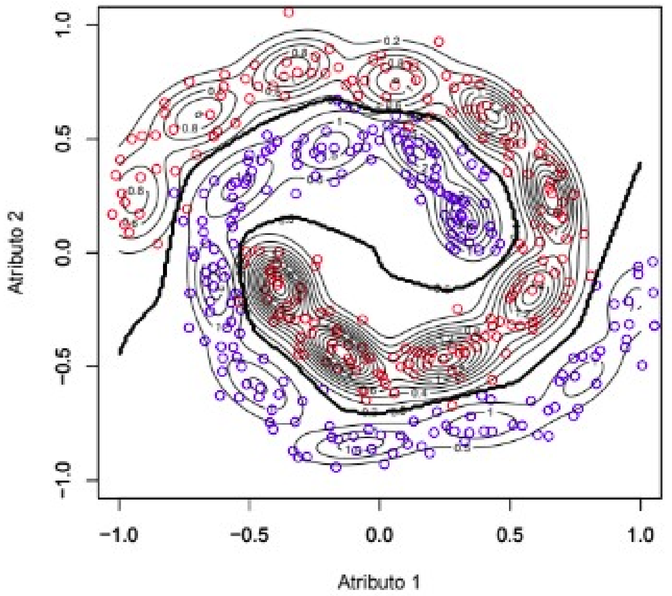
Em problemas 2D bem comportados é mais fácil visualizar e definir o número ideal de partições para a mistura de Gaussianas.





Introdução

Entretanto em problemas reais, na maioria das vezes não é possível definir o número de clusters de forma Visual.



- Uma forma seria avaliar o resultado do classificador em função do número de clusters;
- Definir uma forma de avaliar quantitativamente a qualidade da partição.

Índices de qualidade

- Índices Internos de qualidade de clusters;
  - Ball-Hall
  - C
  - Calinski-Harabasz
  - SD
  - Silhueta

- Índices externos de qualidade de clusters.
  - Czekanowski-Dice
  - Folkes-Mallows
  - Jaccard

Índice Calinski-Harabasz

$$\mathcal{I}_{\mathcal{CH}} = \frac{\frac{BGSS}{K-1}}{\frac{WGSSS}{N-K}} = \frac{N-K}{K-1}\frac{BGSS}{WGSS}$$

Dispersão entre clusters

$$BGSS = \sum_{k=1}^{K} n_k \|\mu^{\{k\}} - \mu\|^2$$

Dispersão intra-cluster k

$$WGSS^{\{k\}} = \sum_{i \in I_k} \|O_i^{\{k\}} - \mu^{\{k\}}\|^2.$$

$$WGSS = \sum_{k=0}^{K} WGSS^{\{k\}}$$

Índice Calinski-Harabasz

$$\mathcal{I}_{\mathcal{CH}} = \frac{\frac{BGSS}{K-1}}{\frac{WGSSS}{N-K}} = \frac{N-K}{K-1}\frac{BGSS}{WGSS}$$

Quanto maior este índice melhor a partição;

$$\mathcal{I}_{\mathcal{CH}} \propto \frac{\left\{\begin{array}{c} custer\ distance\ to \\ the\ data\ centroid \end{array}\right\}}{\left\{\begin{array}{c} sum\ of\ distances\ to \\ cluster\ centroid\ inside\ each\ cluster \end{array}\right\}}$$

Índice Silhueta

Definições:

Distância média intra-cluster onde O são amostras;

$$a(i) = \frac{1}{n_k - 1} \sum_{\substack{i' \in I_k \\ i' \neq i}} d(O_i, O_{i'})$$

Distância média de um ponto O a amostras de outros clusters;

$$\mathfrak{d}(O_i, C_{k'}) = \frac{1}{n_{k'}} \sum_{i' \in I_{k'}} d(Oi, O_{i'})$$

Mínima distância do conjunto anterior;

$$b(i) = \min_{k' \neq k} \mathfrak{d}(O_i, C_{k'})$$

Silhueta de cada ponto O_i;

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

Silhueta média de cada Cluster;

$$\mathfrak{s}_k = \frac{1}{n_k} \sum_{i \in I_k} s(i),$$

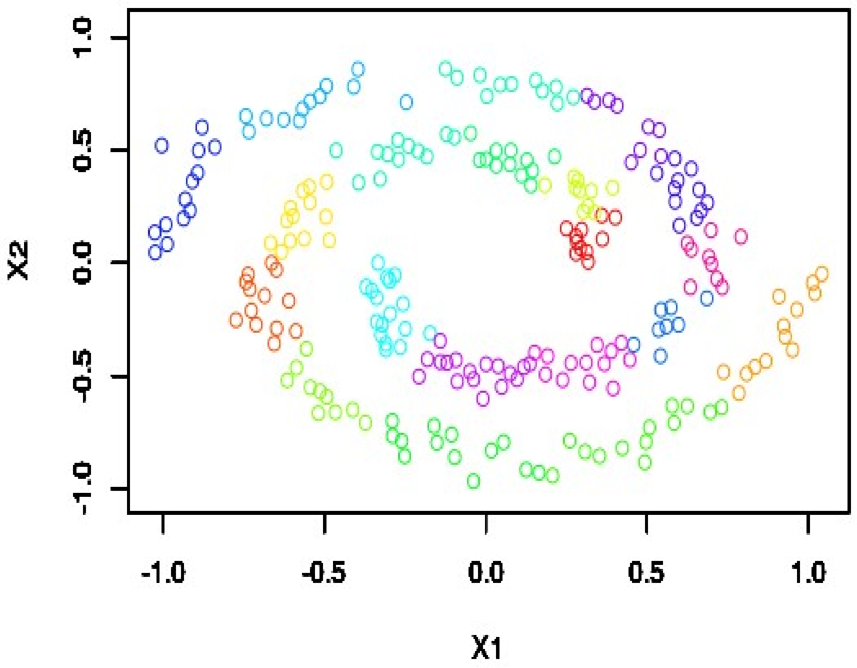
Índice Silhueta

$$\mathcal{I}_S = \frac{1}{K}\sum_{k=1}^{K} \mathfrak{s}_k$$

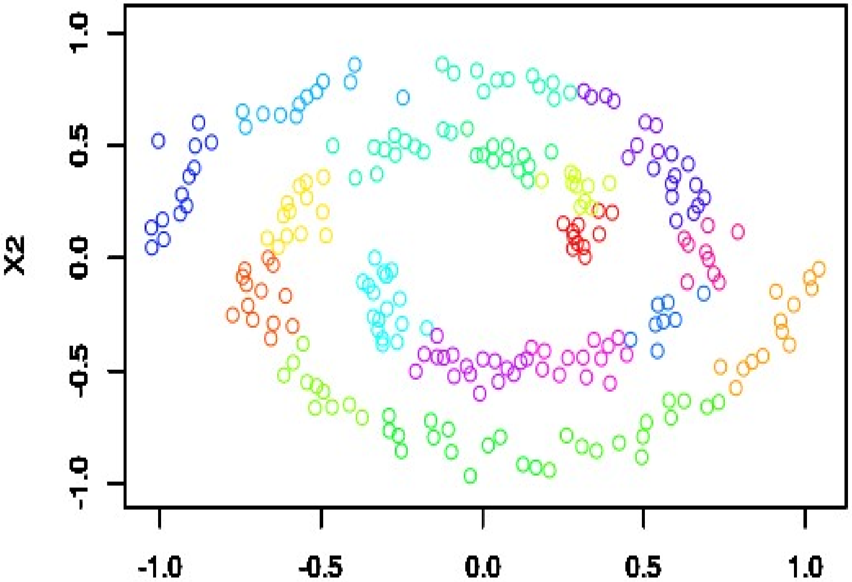
Quanto maior este índice melhor a partição;

$$I_S \propto \left\{\begin{matrix} average\ distance\ to\ point \\ inside\ other\ cluster \end{matrix}\right\} - \left\{\begin{matrix} average\ distance\ to \\ points\ inside\ cluster \end{matrix}\right\}$$

Espaço das Verossimilhanças



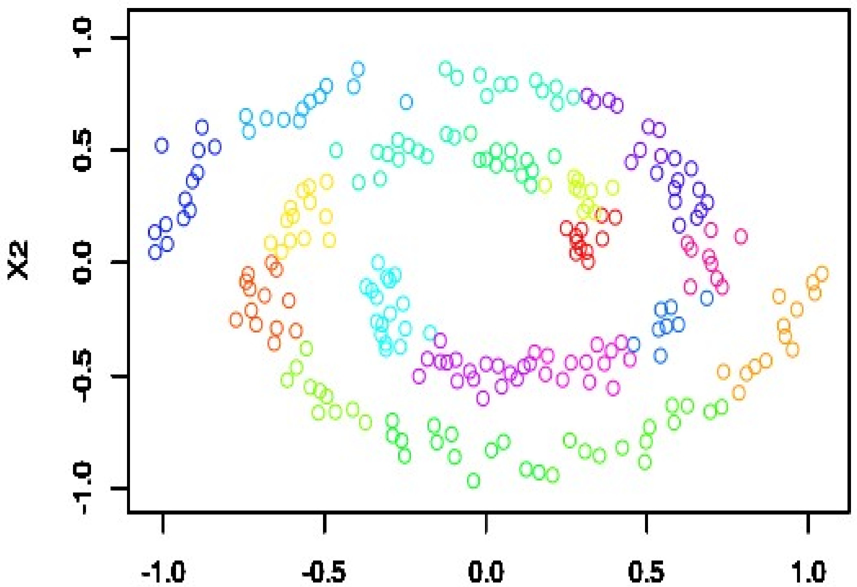
Espaço das Verossimilhanças

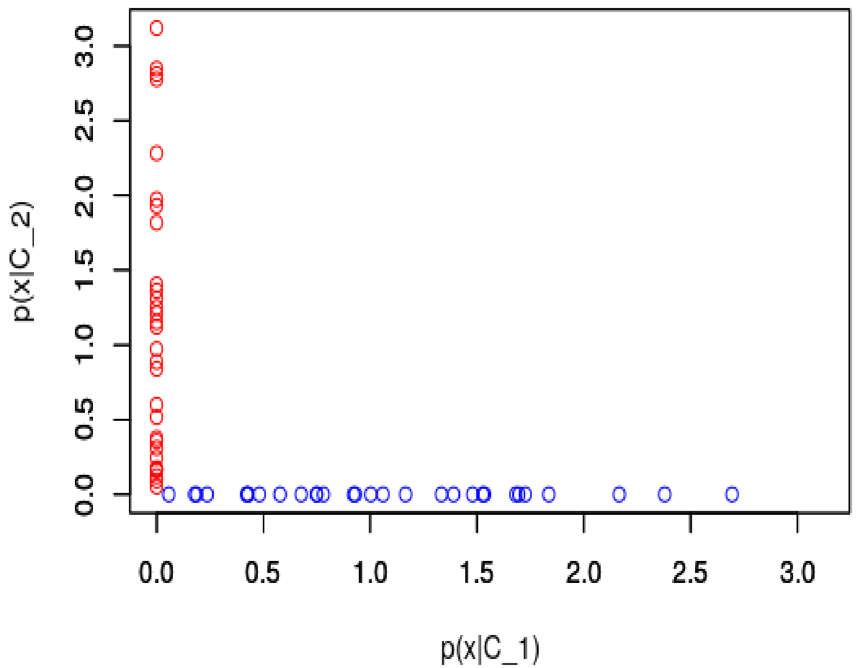


$$P(\mathbf{x}|S_1, \cdots, S_p) = \sum_{k=1}^{p} \pi_k \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu}_k)^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_k - \boldsymbol{\mu}_k)\right)$$

$P(x|C_1) = P(x|S_1,S_2,\ldots S_k)$
onde as partições S_1 a S_k pertencem a C_1

Espaço das Verossimilhanças

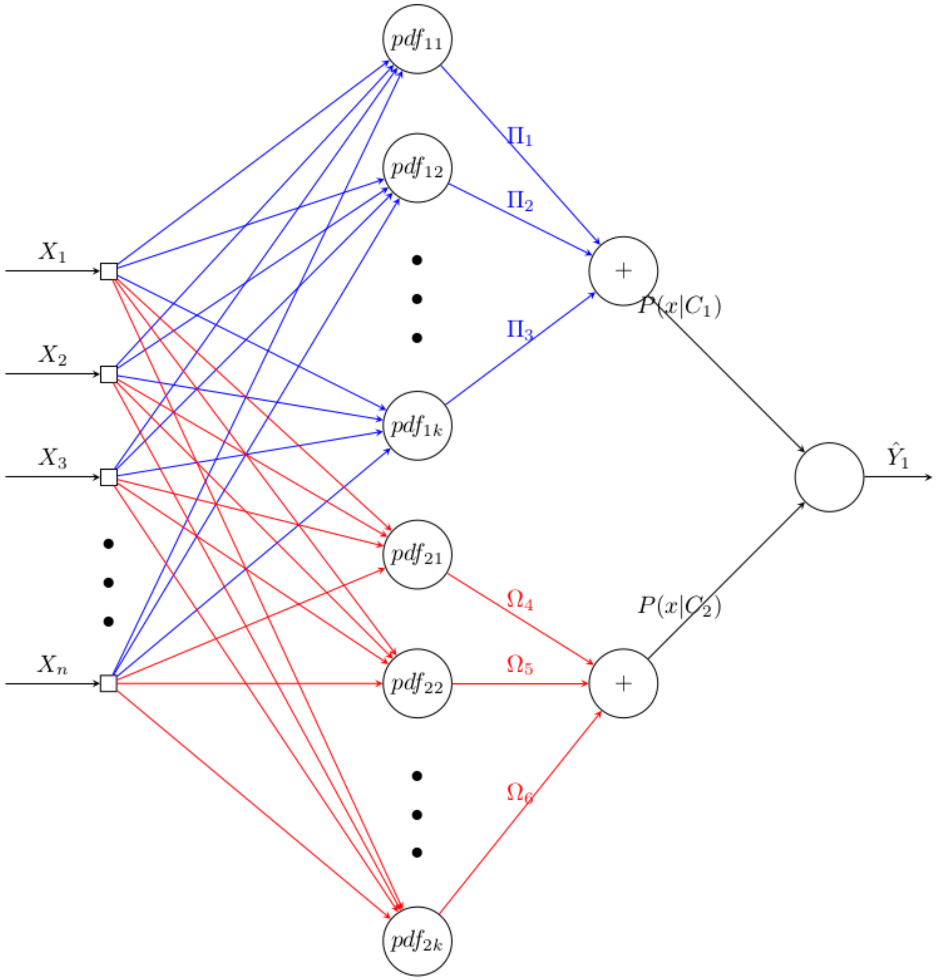


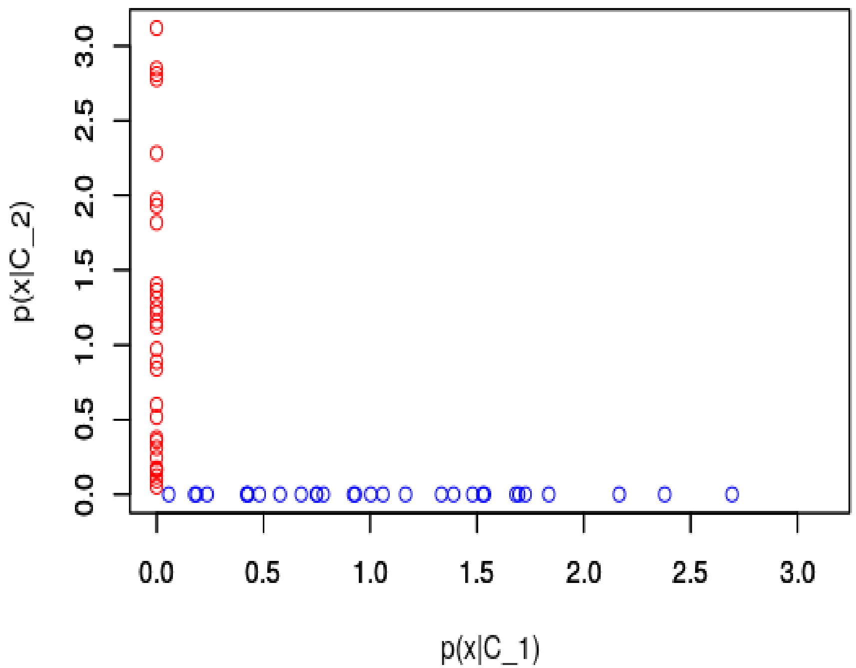


$$P(\mathbf{x}|S_1, \cdots, S_p) = \sum_{k=1}^{p} \pi_k \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}_k|}} \exp\left(-\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_k - \boldsymbol{\mu}_k)\right)$$

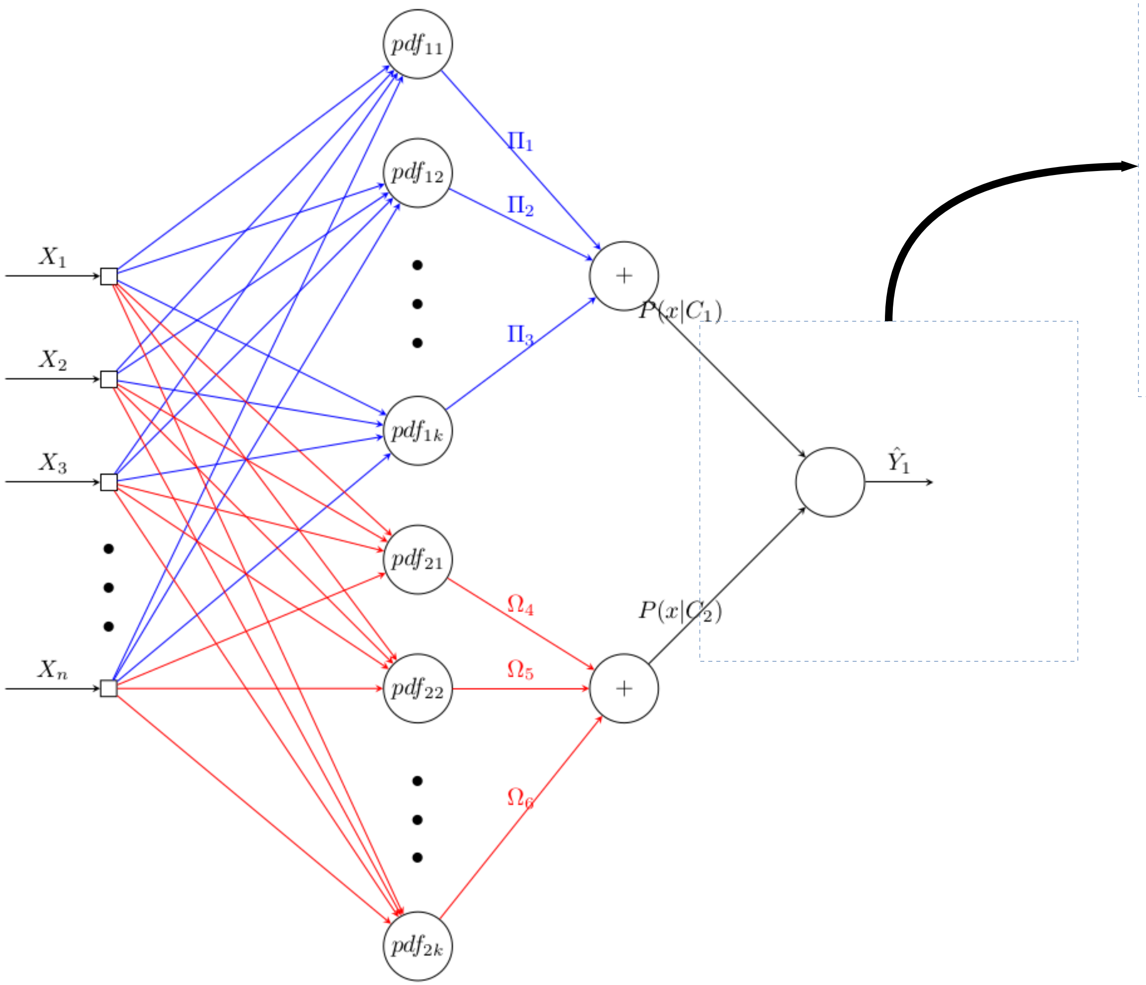
$P(x|C_1) = P(x|S_1,S_2,...S_k)$
onde as partições S_1 a S_k pertencem a C_1

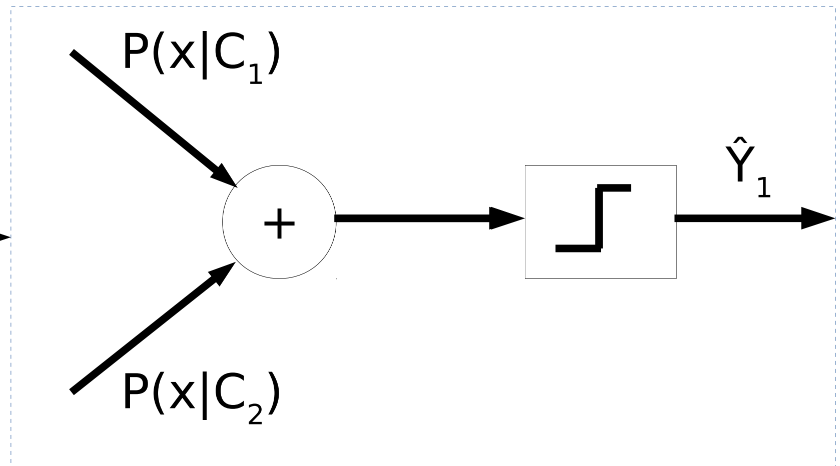
Espaço das Verossimilhanças





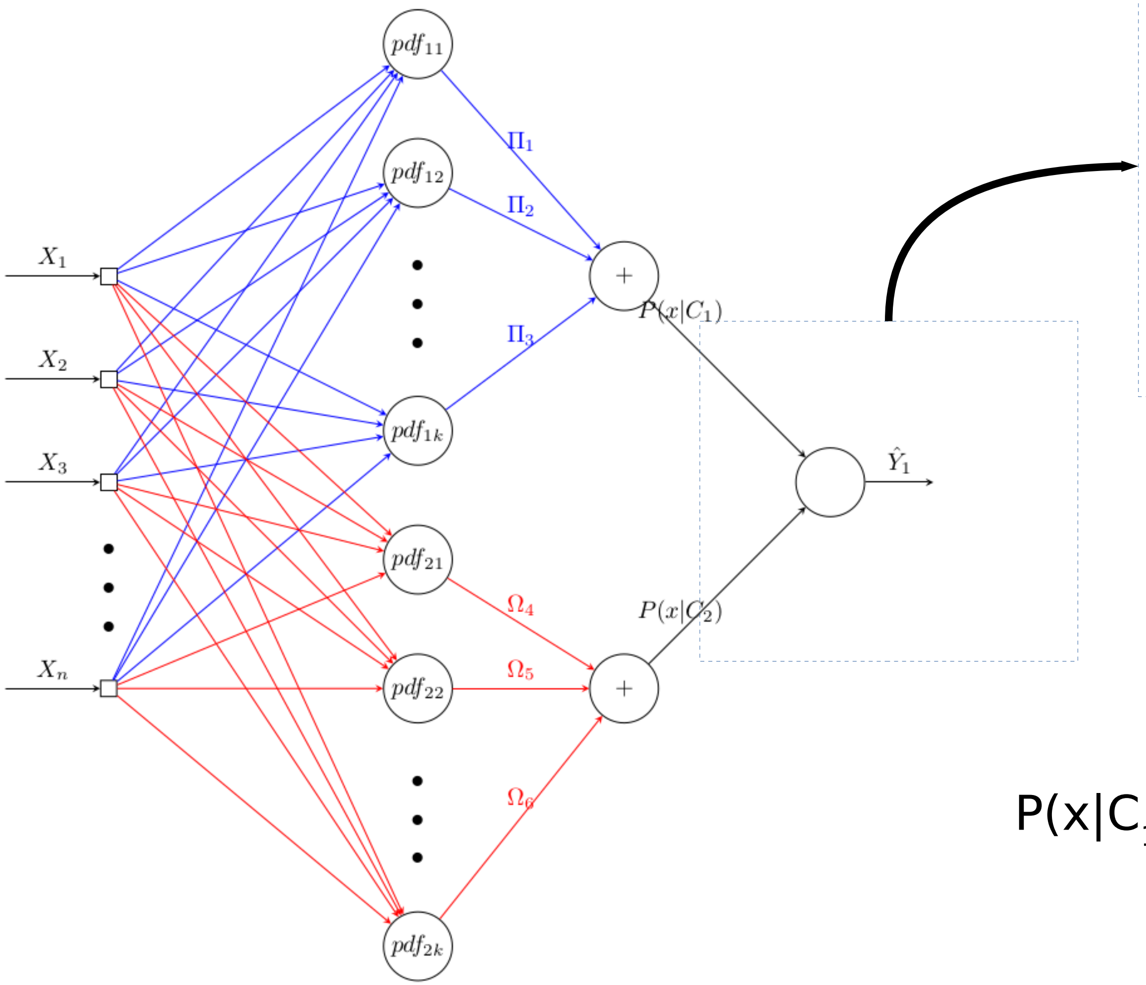
Espaço das Verossimilhanças

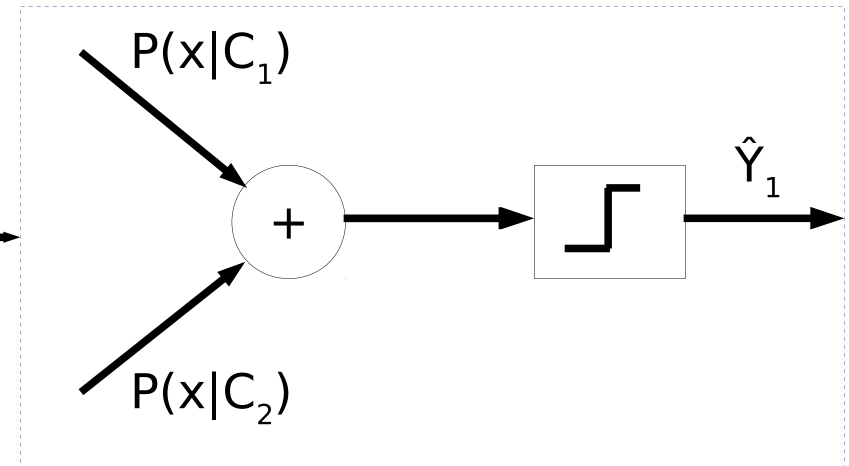




$$\frac{P(C_1|x)}{P(C_2|x)} = \frac{P(x|C_1)\,P(C_1)}{P(x|C_2)\,P(C_2)} \geq 1$$

Espaço das Verossimilhanças

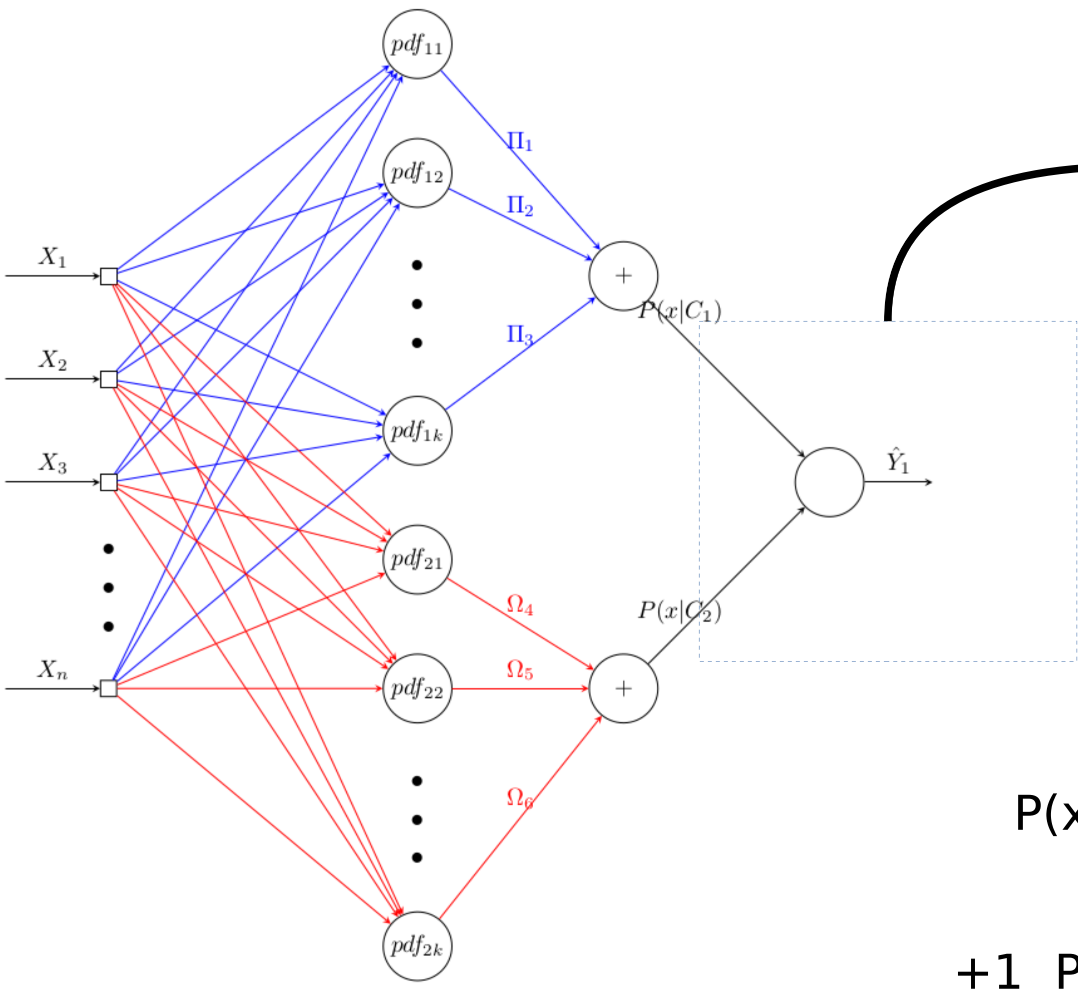


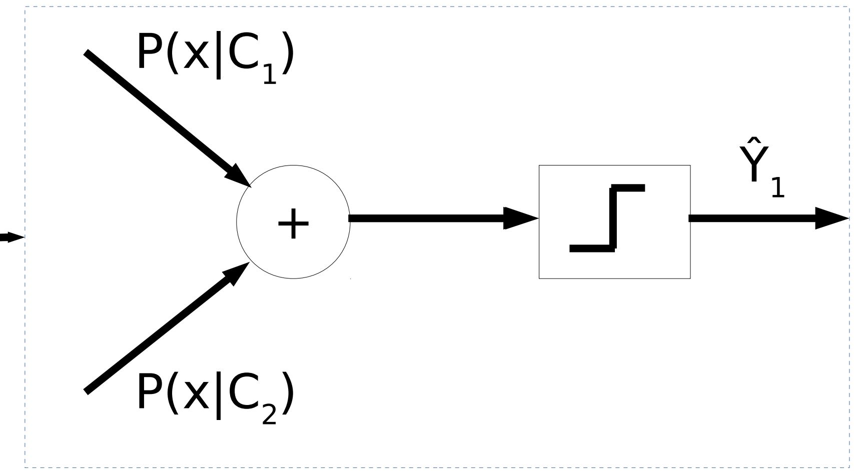


$$\frac{P(C_1|x)}{P(C_2|x)} = \frac{P(x|C_1)\,P(C_1)}{P(x|C_2)\,P(C_2)} \geq 1$$

$$P(x|C_1) - \frac{P(C_2)}{P(C_1)} P(x|C_2) \geq 0$$

Espaço das Verossimilhanças



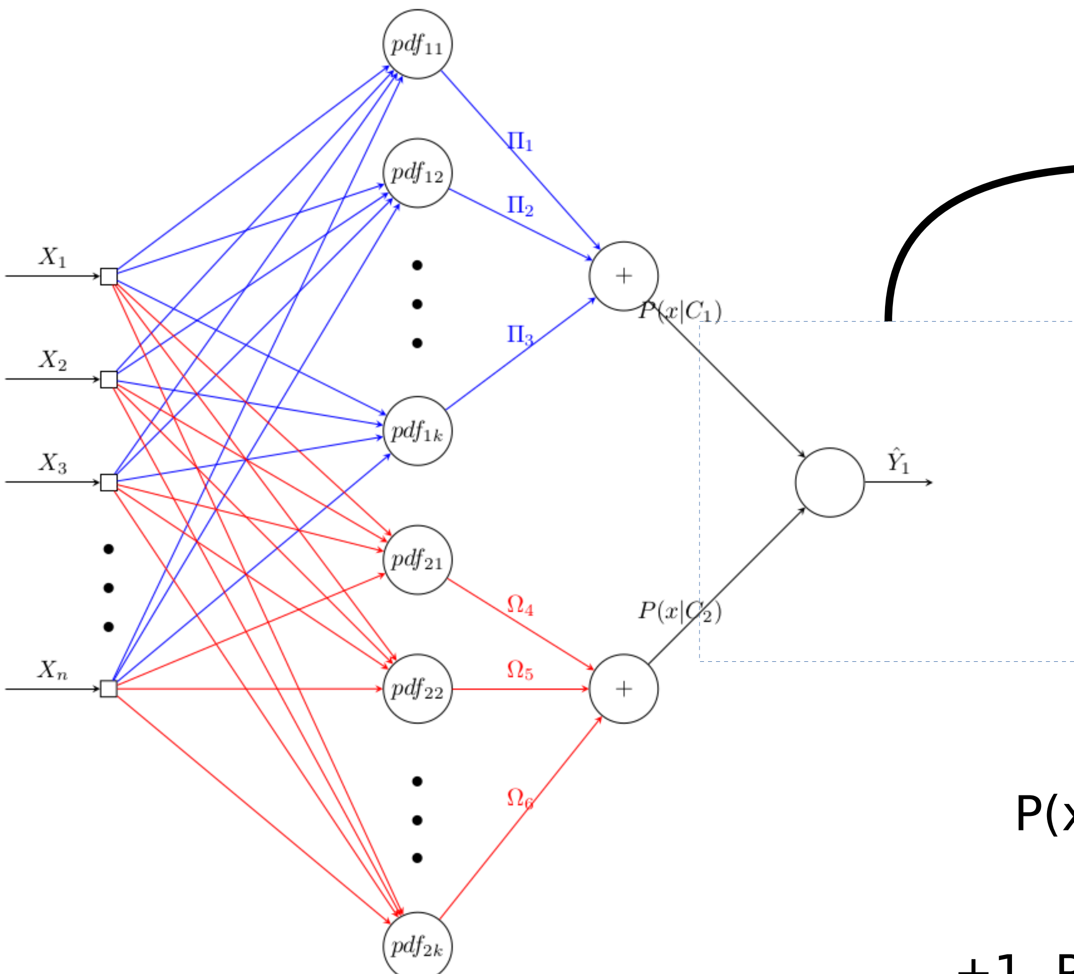


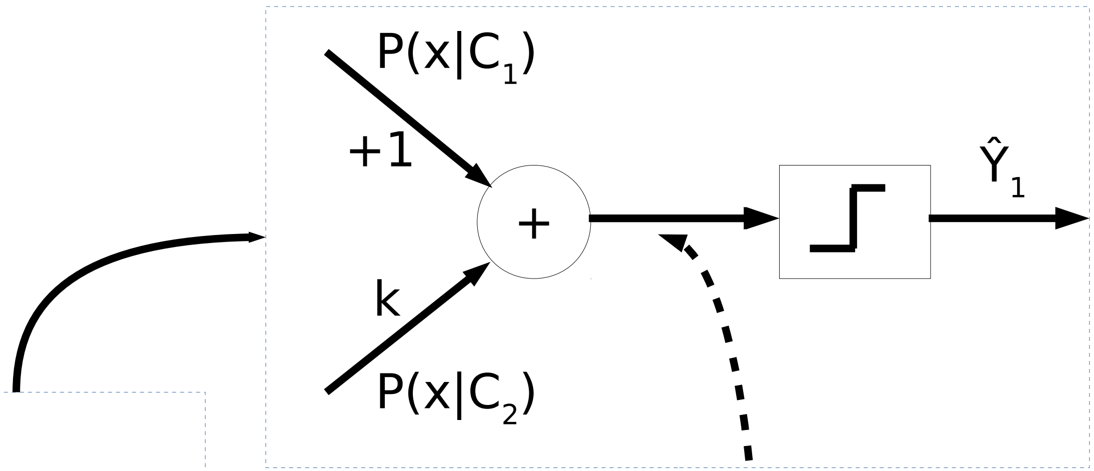
$$\frac{P(C_1|x)}{P(C_2|x)} = \frac{P(x|C_1)\,P(C_1)}{P(x|C_2)\,P(C_2)} \geq 1$$

$$P(x|C_1) - \frac{P(C_2)}{P(C_1)} P(x|C_2) \geq 0$$

$$+1 \;\; P(x|C_1) - k \;\; P(x|C_2) \geq 0$$

Espaço das Verossimilhanças



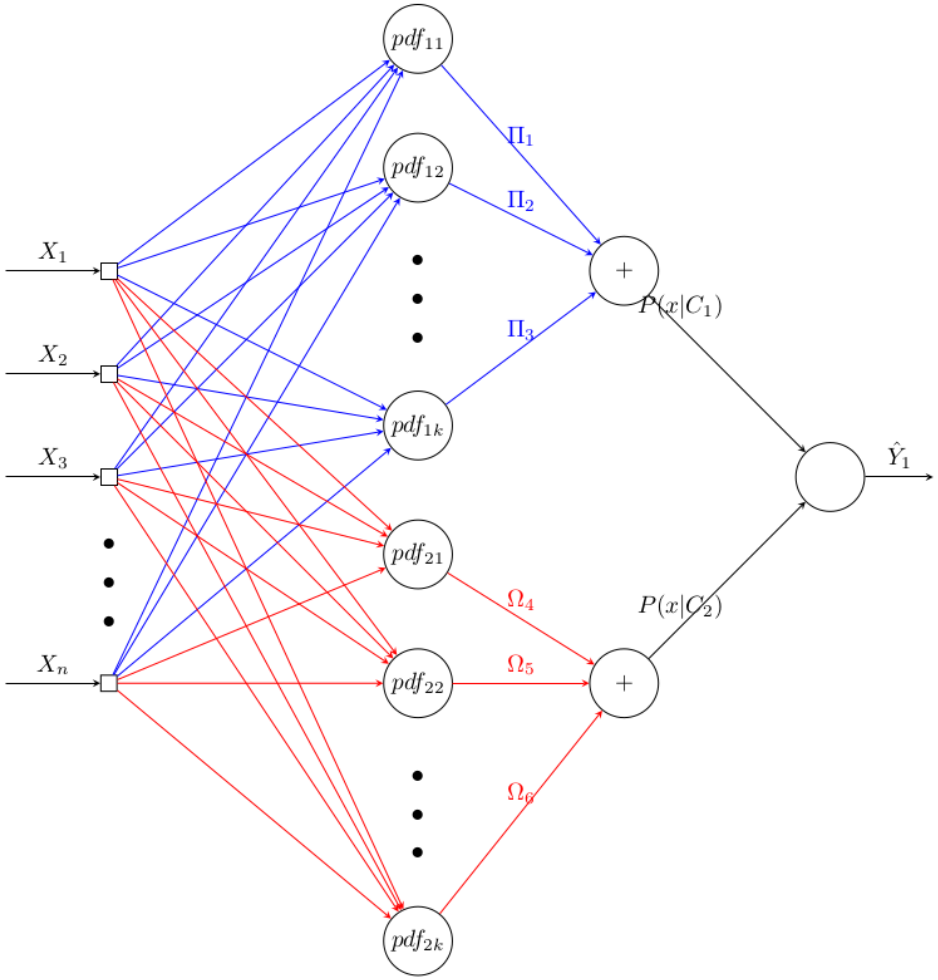


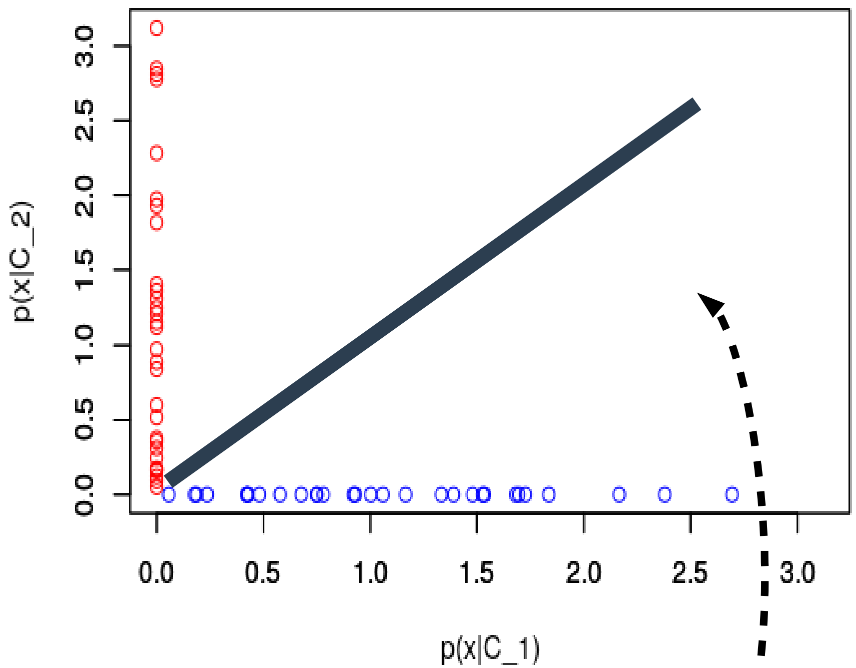
$$\frac{P(C_1|x)}{P(C_2|x)} = \frac{P(x|C_1)\,P(C_1)}{P(x|C_2)\,P(C_2)} \geq 1$$

$$P(x|C_1) - \frac{P(C_2)}{P(C_1)} P(x|C_2) \geq 0$$

$$+1\ P(x|C_1) - k\ P(x|C_2) \geq 0$$

Espaço das Verossimilhanças





$+1\ P(x|C_1) - k\ P(x|C_2) \geq 0$