Reconhecimento de Padrões

Estimativa de densidades

Caracterização do Problema

- Problema de classificação binária
- Objetivo: discriminar classes C_1 e C_2 a partir de um vetor de características

$$\mathbf{x} = [x_1, x_2, \cdots, x_n]$$

- Considere N amostras do conjunto

$$D = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$$

Caracterização do Problema

- Considere N amostras do conjunto

$$D = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$$

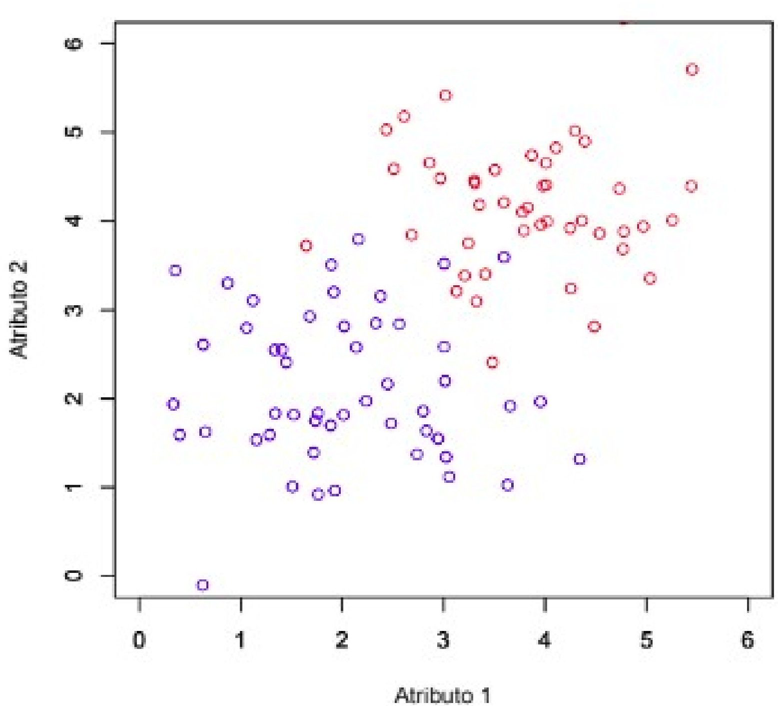
- Onde

$$y_i = -1 \;\forall\; \mathbf{x}_i \in C_1 \text{ e } y_i = +1 \;\forall\; \mathbf{x}_i \in C_2$$

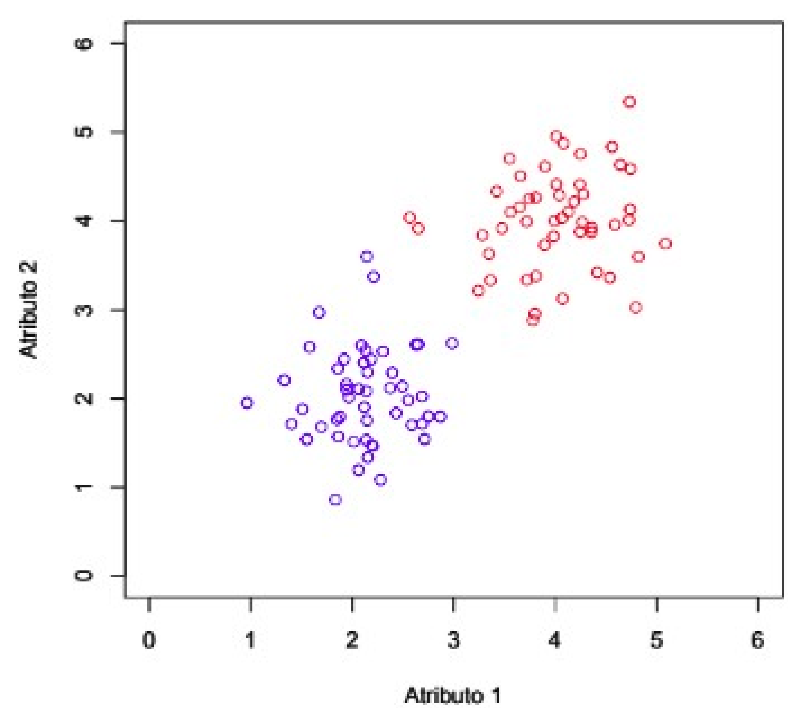
Caracterização do Problema

- O vetor X deve ser representativo do problema, discriminando corretamente C_1 e C_2
- Espera-se que um número suficiente de amostras permita a estimativa das funções de densidade de probabilidade que permitam a discriminação das duas classes. $P(x|C_1)$ e $P(x|C_2)$

Caracterização do Problema



(a) Exemplo de amostragem realizada a partir de atributos não representativos do problema.



(b) Exemplo de amostragem realizada a partir de atributos representativos do problema.

Caracterização do Problema

- Probabilidades a priori dos elementos das classes
- Exemplo: Uma doença atinge 20% da população. A probabilidade a priori de um cidadão apresentar a doença será de 20%, enquanto que a probabilidade a priori de não apresentar a doença será de 80%.

$$P(C_1) = 0.8 \text{ e } P(C_2) = 0.2$$

Caracterização do Problema

- Probabilidades a priori são baseadas no número de ocorrências de cada classe

$$P(C_1) = \frac{N_1}{N_1 + N_2}$$

$$P(C_2) = \frac{N_2}{N_1 + N_2}$$

Caracterização do Problema

- Classificar usando probabilidade a priori não considera nenhuma informação sobre a ocorrência de X.
- Com base na informação das distribuições dos atributos de X podemos estimar as PDFs condicionais (verossimilhanças) e então classificar efetivamente.
- As probabilidade a priori irão ponderar as verossimilhanças.

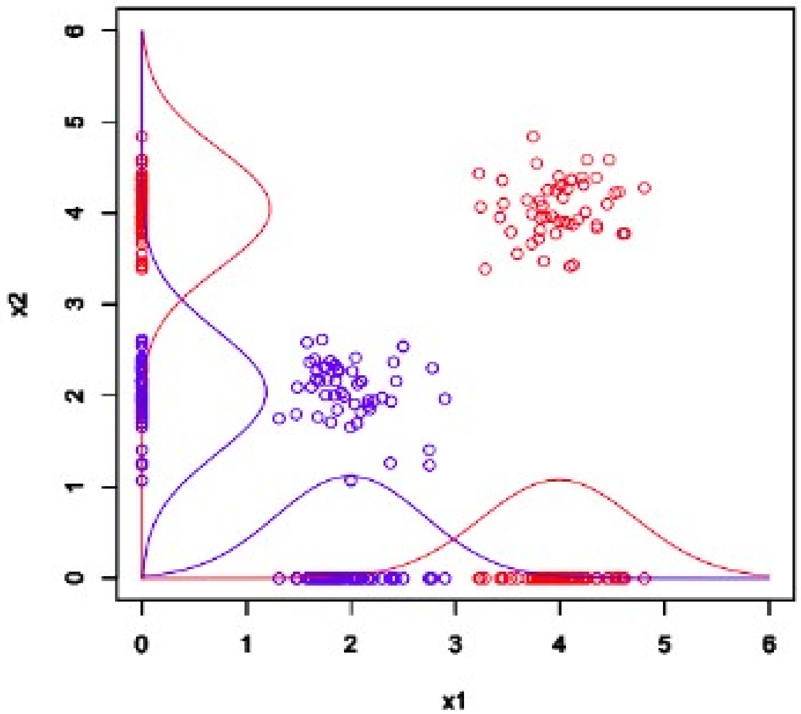
Estimando as PDFs

- Para atributos indepentendes (sem correlação entre eles)

  Verossimilhanças podem ser obtidas pelo produto das PDFs marginais

- PDFs marginais são obtidas individualmente para cada atributo de X

Estimando as PDFs



(a) Estimativa de pdfs marginais para duas classes C_1 e C_2. Densidades marginais são representadas por cores distintas sobre os eixos dos atributos x_1 e x_2.

$P(x_1|C_1)$ e $P(x_2|C_1)$ são duas distribuições marginais que compõem $P(x|C_1)$

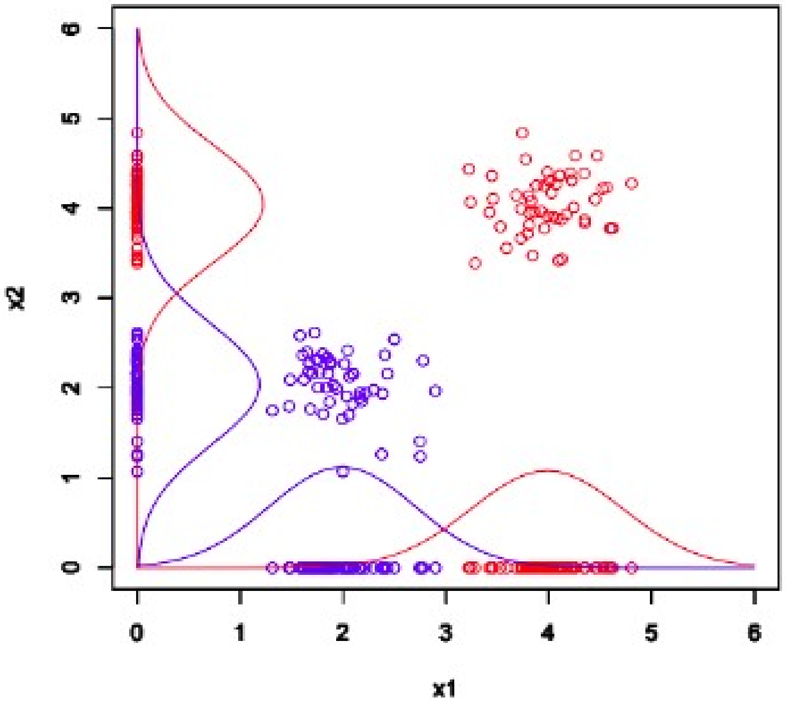
$P(x|C_1) = P((x_1,x_2)|C_1)$

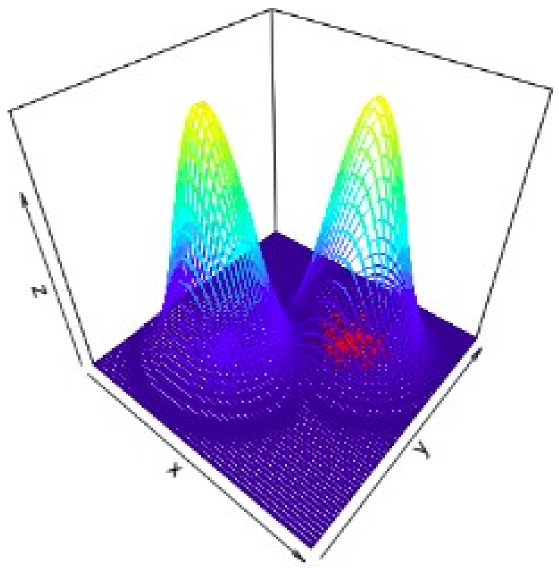
Pois $x = [x_1,x_2]$

Logo:

$P(x|C_1) = P(x_1|C_1)\ P(x_2|C_1)$
$P(x|C_2) = P(x_1|C_2)\ P(x_2|C_2)$

Estimando as PDFs





Estimando as PDFs

- As probabilidades marginais podem ser calculadas pela expressão de uma pdf normal univariada:

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

Estimando as PDFs

- Logo:

$$P(\mathbf{x}|C_1) = \frac{1}{2\pi\sigma_{11}\sigma_{12}} e^{-\left(\frac{(x_1-\mu_{11})^2}{2\sigma_{11}^2} + \frac{(x_2-\mu_{21})^2}{2\sigma_{21}^2}\right)}$$

$$P(\mathbf{x}|C_2) = \frac{1}{2\pi\sigma_{12}\sigma_{22}} e^{-\left(\frac{(x_1-\mu_{12})^2}{2\sigma_{12}^2} + \frac{(x_2-\mu_{22})^2}{2\sigma_{22}^2}\right)}$$

Estimando as PDFs

- Para atributos de x que não sejam independentes
- Considerar correlação ou covariância entre atributos

$$p(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left((\frac{x_1-\mu_1}{\sigma_1})^2 - (\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}) + (\frac{x_2-\mu_2}{\sigma_2})^2 \right)}$$

Onde p é o coeficiente de correlação linear entre x_1 e x_2

Multivariados

- Caso geral para vários atributos

$$p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n|\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

$$\mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \cdots & \rho_{1n}\sigma_1\sigma_n \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 & \cdots & \rho_{2n}\sigma_2\sigma_n \\ \vdots & \vdots & \cdots & \vdots \\ \rho_{n1}\sigma_n\sigma_1 & \rho_{n2}\sigma_n\sigma_2 & \cdots & \sigma_n^2 \end{bmatrix}$$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$