Universidade Federal do Ceará

Departamento de Computação

Aprendizagem de Máquina

César Lincoln Cavalcante Mattos

2020

Agenda

Classificadores estatísticos

Considere padrões de entrada $\boldsymbol{x}_i \in \mathbb{R}^D, i \in \{1, \cdots, N\}$, e respectivas classes $y_i \in \{C_1, C_2, \ldots C_K\}$.

Modelos discriminantes

Estimam parâmetros para as fronteiras de decisão entre classes a partir dos dados.

- **Regressão logística**: Aprendem a distribuição $p(y_i|\boldsymbol{x}_i)$ diretamente.

Classificadores estatísticos

Considere padrões de entrada $\boldsymbol{x}_i \in \mathbb{R}^D, i \in \{1, \cdots, N\}$, e respectivas classes $y_i \in \{C_1, C_2, \ldots C_K\}$.

Modelos discriminantes

Estimam parâmetros para as fronteiras de decisão entre classes a partir dos dados.

- **Regressão logística**: Aprendem a distribuição $p(y_i|\boldsymbol{x}_i)$ diretamente.

Modelos generativos

Modelam a distribuição das entradas associadas a cada classe.

- **Classificadores Bayesianos**: Consideram um modelo para $p(\boldsymbol{x}_i|y_i)$, definem uma priori $p(y_i)$ e aplicam a Regra de Bayes para obter $p(y_i|\boldsymbol{x}_i)$.

Agenda

Classificadores Bayesianos

- **Problema**: Dado um conjunto de características (atributos) x de um padrão, a qual classe o padrão pertence?

Classificadores Bayesianos

- **Problema**: Dado um conjunto de características (atributos) $\boldsymbol{x}$ de um padrão, a qual classe o padrão pertence?
- Pela Regra de Bayes:

$$p(y = C_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y = C_k)p(y = C_k)}{p(\boldsymbol{x})}, \quad k \in \{1, \cdots, K\}.$$

Classificadores Bayesianos

- **Problema**: Dado um conjunto de características (atributos) $\boldsymbol{x}$ de um padrão, a qual classe o padrão pertence?
- Pela Regra de Bayes:

$$p(y = C_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y = C_k)p(y = C_k)}{p(\boldsymbol{x})}, \quad k \in \{1, \cdots, K\}.$$

- A notação pode ser simplificada:

$$p(C_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_k)p(C_k)}{p(\boldsymbol{x})}, \quad k \in \{1, \cdots, K\}.$$

Classificadores Bayesianos

- **Problema**: Dado um conjunto de características (atributos) $\boldsymbol{x}$ de um padrão, a qual classe o padrão pertence?
- Pela Regra de Bayes:

$$p(y = C_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|y = C_k)p(y = C_k)}{p(\boldsymbol{x})}, \quad k \in \{1, \cdots, K\}.$$

- A notação pode ser simplificada:

$$p(C_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_k)p(C_k)}{p(\boldsymbol{x})}, \quad k \in \{1, \cdots, K\}.$$

- Formalmente, temos:

$$\text{posteriori} = \frac{\text{verossimilhança da classe} \times \text{priori}}{\text{evidência (ou verossimilhança marginal)}}$$

Classificadores Bayesianos

- Classificação binária (C_1 e C_2):

$$p(C_1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_1)p(C_1)}{p(\boldsymbol{x})} \propto p(\boldsymbol{x}|C_1)p(C_1)$$

$$p(C_2|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_2)p(C_2)}{p(\boldsymbol{x})} \propto p(\boldsymbol{x}|C_2)p(C_2)$$

- **Ideia**: Escolha a classe com maior probabilidade.

Classificadores Bayesianos

- Classificação binária (C_1 e C_2):

$$p(C_1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_1)p(C_1)}{p(\boldsymbol{x})} \propto p(\boldsymbol{x}|C_1)p(C_1)$$

$$p(C_2|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_2)p(C_2)}{p(\boldsymbol{x})} \propto p(\boldsymbol{x}|C_2)p(C_2)$$

- **Ideia**: Escolha a classe com maior probabilidade.
- **Problema**: Como calcular as distribuições acima?

Classificadores Bayesianos

- Classificação binária (C_1 e C_2):

$$p(C_1|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_1)p(C_1)}{p(\boldsymbol{x})} \propto p(\boldsymbol{x}|C_1)p(C_1)$$

$$p(C_2|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_2)p(C_2)}{p(\boldsymbol{x})} \propto p(\boldsymbol{x}|C_2)p(C_2)$$

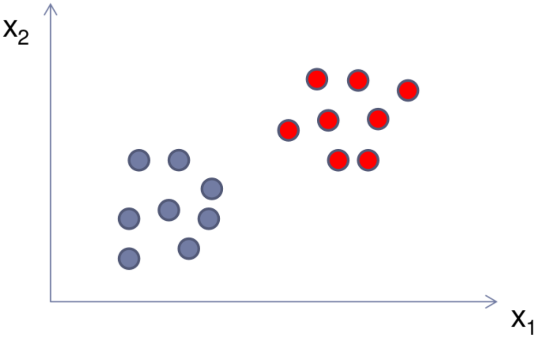
- **Ideia**: Escolha a classe com maior probabilidade.
- **Problema**: Como calcular as distribuições acima?
- **Ideia**: Estimar as probabilidades a partir do conjunto de treinamento.

Classificadores Bayesianos

- **Problema**: Como estimar as probabilidades $p(C_1)$ e $p(C_2)$?

Classificadores Bayesianos

- **Problema**: Como estimar as probabilidades $p(C_1)$ e $p(C_2)$?
- **Ideias**:
  - Considerar classes equiprováveis: $p(C_1) = p(C_2) = 0.5$
  - Proporcionais aos números de exemplos disponíveis.
  - Conhecidas pela natureza do problema.

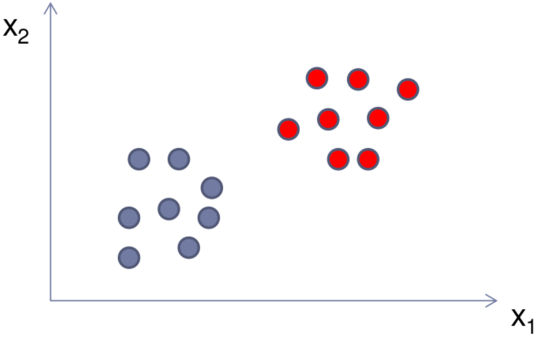


Classificadores Bayesianos

- **Problema**: Como estimar as probabilidades $p(\boldsymbol{x}|C_1)$ e $p(\boldsymbol{x}|C_2)$?

Classificadores Bayesianos

- **Problema**: Como estimar as probabilidades $p(\boldsymbol{x}|C_1)$ e $p(\boldsymbol{x}|C_2)$?
- **Ideia**: Considerar que os dados foram gerados de uma distribuição de probabilidade específica e estimar seus parâmetros.



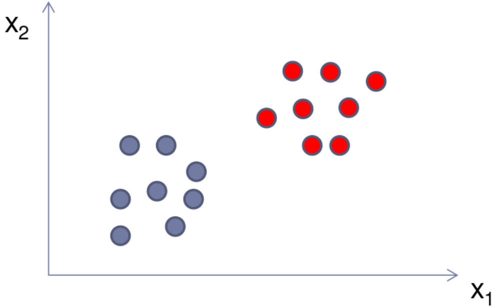
Análise de Discriminante Gaussiano

- Considerando distribuições Gaussianas, temos:

$$p(C_1) = p(C_2) = 0.5 \text{ ou } p(C_k) = \frac{N_k}{N}, \ \forall k \in \{1, 2\}$$

$$p(\boldsymbol{x}|C_1) = \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \text{ e } p(\boldsymbol{x}|C_2) = \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2).$$

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{N_k} \sum_{\boldsymbol{x}_i \in C_k} \boldsymbol{x}_i \text{ e } \hat{\boldsymbol{\Sigma}}_k = \frac{1}{N_k - 1} \sum_{\boldsymbol{x}_i \in C_k} (\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_i - \hat{\boldsymbol{\mu}}_k)^\top, \ \forall k \in \{1, 2\}.$$



Análise de Discriminante Gaussiano

- Classificação de um novo padrão $\boldsymbol{x}_*$:

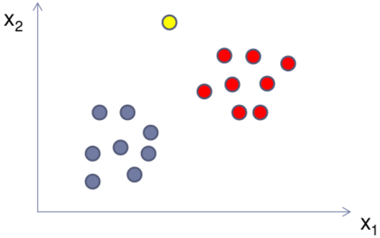
$$p(\boldsymbol{x}_*|C_k) = \frac{1}{|\boldsymbol{\Sigma}_k|^{1/2}(2\pi)^{D/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_* - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_* - \boldsymbol{\mu}_k)\right), \; k \in \{1, 2\}$$

- Escolha a classe mais provável:

$$p(C_k|\boldsymbol{x}_*) \propto p(\boldsymbol{x}_*|C_k)p(C_k), \; k \in \{1, 2\}$$

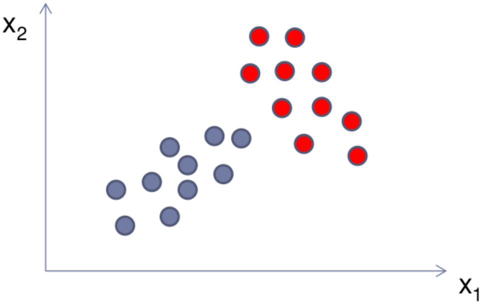
$$\log p(C_k|\boldsymbol{x}_*) \propto \log p(\boldsymbol{x}_*|C_k) + \log p(C_k), \; k \in \{1, 2\}$$

$$\log p(C_k|\boldsymbol{x}_*) \propto -\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\boldsymbol{x}_* - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_* - \boldsymbol{\mu}_k) + \log p(C_k)$$



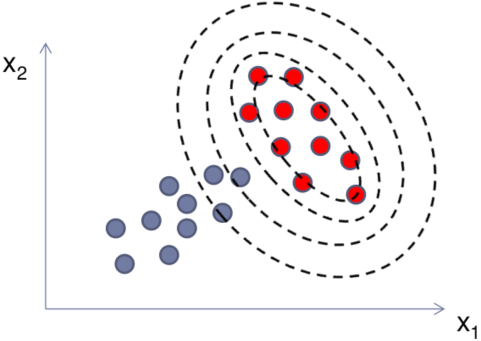
Análise de Discriminante Gaussiano

$$p(C_k|\boldsymbol{x}) \propto p(\boldsymbol{x}|C_k)p(C_k),\ k \in \{1, 2\}$$



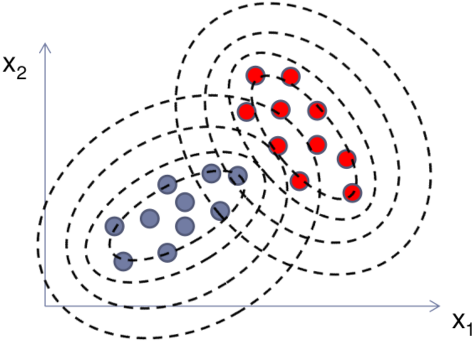
Análise de Discriminante Gaussiano

$$p(C_k|\boldsymbol{x}) \propto p(\boldsymbol{x}|C_k)p(C_k),\ k \in \{1, 2\}$$



Análise de Discriminante Gaussiano

$$p(C_k|\boldsymbol{x}) \propto p(\boldsymbol{x}|C_k)p(C_k),\ k \in \{1,2\}$$



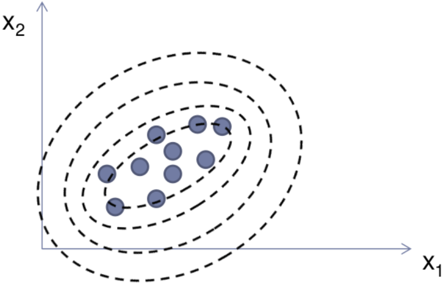
Análise de Discriminante Gaussiano

$$p(\boldsymbol{x}|C_k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),\ k \in \{1, 2\}$$

- Caso bidimensional ($D = 2$):

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

- Valores altos para a covariância σ_{12}:



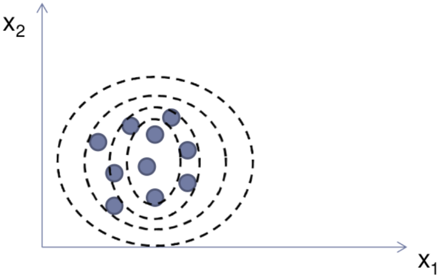
Análise de Discriminante Gaussiano

$$p(\boldsymbol{x}|C_k) = \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \ k \in \{1, 2\}$$

- Caso bidimensional ($D = 2$):

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

- Valores baixos (próximos de zero) para a covariância σ_{12}:

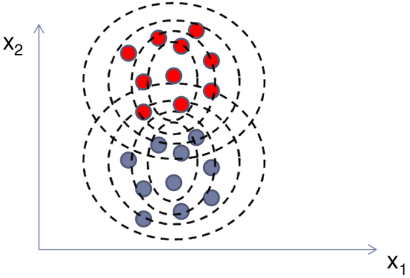


Classificadores Bayesianos

Naive Bayes

- Considera atributos independentes, dada a classe do padrão.
- Caso bidimensional ($D = 2$):

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} \left(\sigma_1^{(1)}\right)^2 & 0 \\ 0 & \left(\sigma_2^{(1)}\right)^2 \end{bmatrix} \text{ e } \boldsymbol{\Sigma}_2 = \begin{bmatrix} \left(\sigma_1^{(2)}\right)^2 & 0 \\ 0 & \left(\sigma_2^{(2)}\right)^2 \end{bmatrix}$$



Classificadores Bayesianos

Naive Bayes

- Dado $\boldsymbol{x} = [x_1, x_2, \ldots, x_D]^\top$, calcula as probabilidade das classes:

$$p(C_k|\boldsymbol{x}) \propto p(\boldsymbol{x}|C_k)p(C_k), \ \forall k.$$

- Considera atributos independentes dada a classe:

$$p(C_k|\boldsymbol{x}) \propto p(x_1|C_k)p(x_2|C_k)\cdots p(x_D|C_k)p(C_k), \ \forall k$$

$$p(C_k|\boldsymbol{x}) \propto p(C_k)\prod_{d=1}^{D} p(x_d|C_k), \ \forall k.$$

- Predição para um novo padrão $\boldsymbol{x}_*$:

$$\hat{y}_* = \arg\max_k p(C_k)\prod_{d=1}^{D} p(x_{*d}|C_k)$$

$$\hat{y}_* = \arg\max_k \left[\log p(C_k) + \sum_{d=1}^{D} \log p(x_{*d}|C_k)\right].$$

Classificadores Bayesianos

Naive Bayes Gaussiano

- Considera distribuições Gaussianas para $p(x_d|C_k)$:

$$p(C_k|\boldsymbol{x}) \propto p(C_k) \prod_{d=1}^{D} \mathcal{N}(x_d|\mu_{dk}, \sigma_{dk}^2), \ \forall k.$$

- Predição para um novo padrão $\boldsymbol{x}_*$:

$$\hat{y}_* = \arg\max_k \left[\log p(C_k) + \sum_{d=1}^{D} \log \mathcal{N}(x_{*d}|\mu_{dk}, \sigma_{dk}^2)\right]$$

$$\hat{y}_* = \arg\max_k \left[\log p(C_k) - \frac{D}{2}\log 2\pi\sigma_{dk}^2 - \frac{1}{2}\sum_{d=1}^{D} \frac{(x_{*d} - \mu_{dk})^2}{\sigma_{dk}^2}\right].$$

- **Observações**:
  - $\rightarrow$ $\hat{\mu}_{dk} = \frac{1}{N_k}\sum_{\boldsymbol{x}_i \in C_k} x_{id}$ e $\hat{\sigma}_{dk}^2 = \frac{1}{N_k - 1}\sum_{\boldsymbol{x}_i \in C_k}(x_{id} - \hat{\mu}_{dk})^2, \ \forall d, k.$
  - $\rightarrow$ Discriminante Gaussiano com matriz de covariância diagonal.

Resumo dos Classificadores Estatísticos

- **Análise de Discriminante Gaussiano**

$$\hat{y}_* = \arg\max_k \left[\log p(C_k) - \frac{1}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\boldsymbol{x}_* - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_* - \boldsymbol{\mu}_k) \right].$$

- **Naive Bayes**

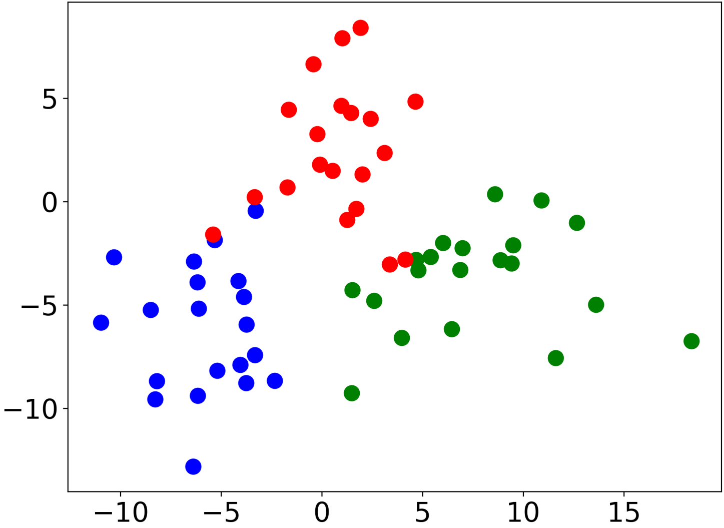
$$\hat{y}_* = \arg\max_k \left[\log p(C_k) + \sum_{d=1}^{D} \log p(x_{*d} | C_k) \right].$$

- **Naive Bayes Gaussiano**

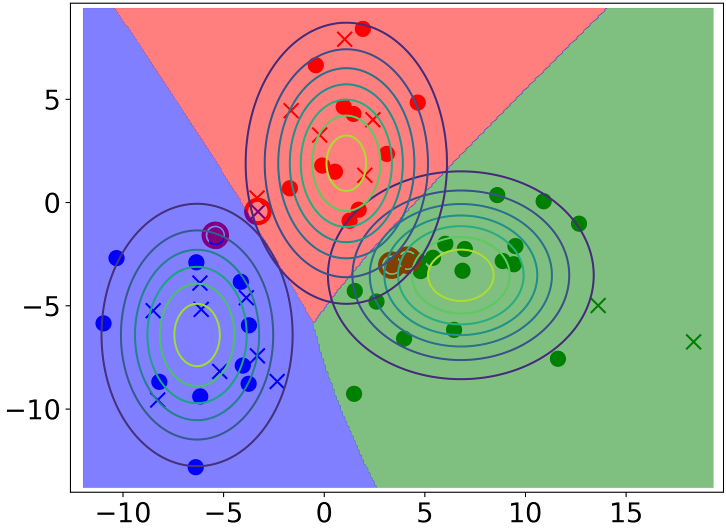
$$\hat{y}_* = \arg\max_k \left[\log p(C_k) - \frac{D}{2} \log 2\pi\sigma_{dk}^2 - \frac{1}{2} \sum_{d=1}^{D} \frac{(x_{*d} - \mu_{dk})^2}{\sigma_{dk}^2} \right].$$

- **Observação**: Os parâmetros de todas as distribuições podem ser estimados a partir dos dados (de treinamento) disponíveis.

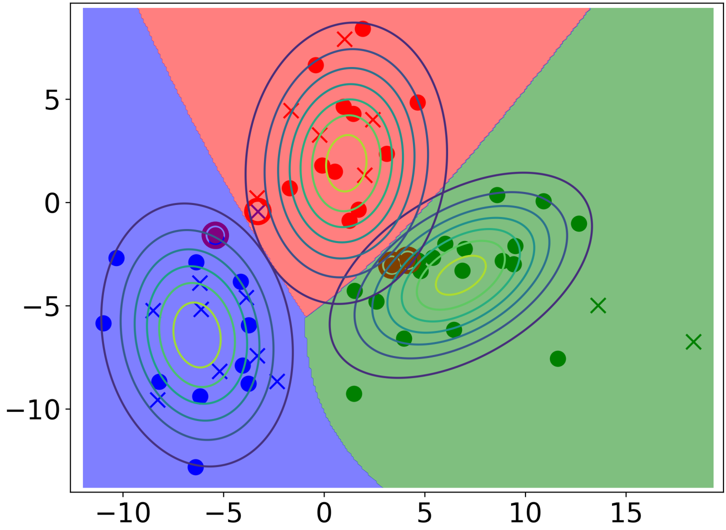
Classificadores Estatísticos



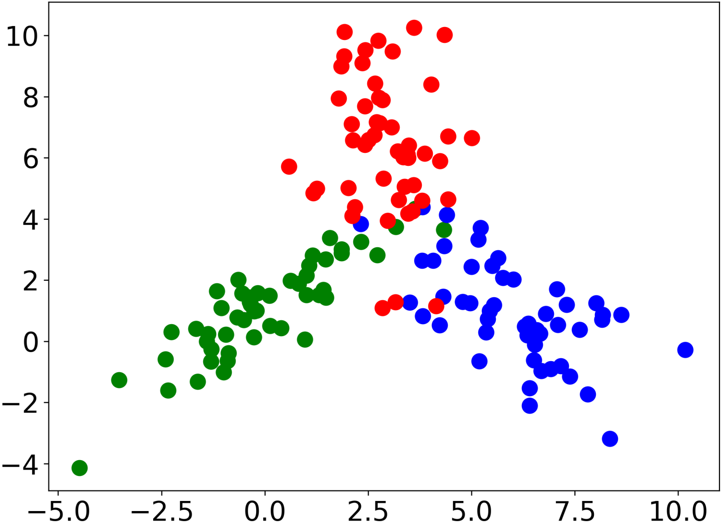
Naive Bayes Gaussiano



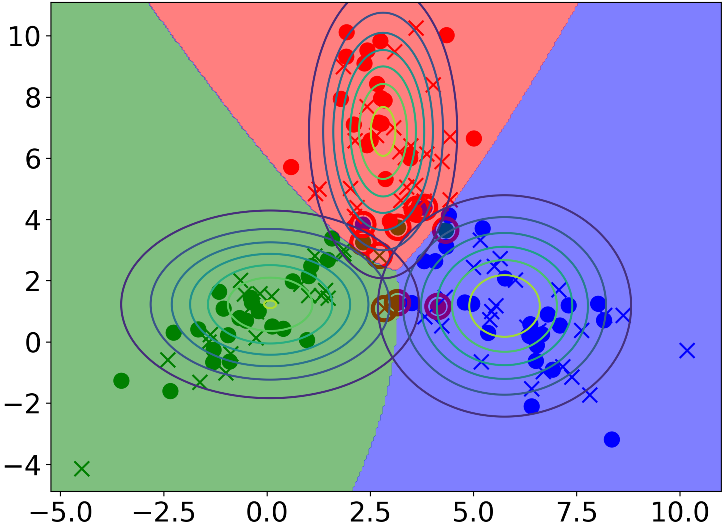
Análise de Discriminante Gaussiano



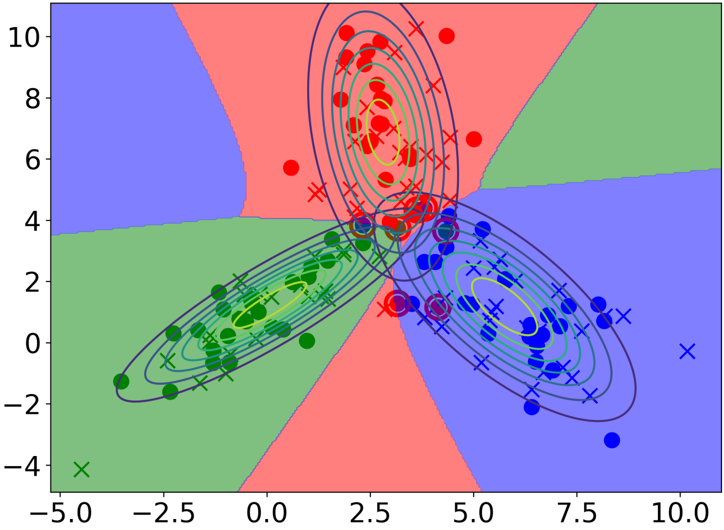
Classificadores Bayesianos



Naive Bayes Gaussiano



Análise de Discriminante Gaussiano



Classificadores Bayesianos

- Quais as vantagens/desvantagens de usar Regressão Logística (RL) ou Análise de Discriminante Gaussiano (ADG)?

Classificadores Bayesianos

- Quais as vantagens/desvantagens de usar Regressão Logística (RL) ou Análise de Discriminante Gaussiano (ADG)?
  - $\rightarrow$ ADG permite fronteiras de decisão não-lineares, dependendo das considerações feitas.
  - $\rightarrow$ ADG considera que as distribuições $p(\boldsymbol{x}|C_k)$ são Gaussianas, o que não necessariamente é verdade.
  - $\rightarrow$ ADG usualmente precisa de menos dados para obter uma boa solução.
  - $\rightarrow$ RL é mais robusta quando considerações incorretas são feitas.

Agenda

Tópicos adicionais

- Classificadores Naive Bayes não-Gaussianos ou mistos.
  - $\rightarrow$ Para classes $k \in \{1, \ldots, K\}$:

$$p(C_k|\boldsymbol{x}) \propto p(C_k) \prod_{d=1}^{D} p(x_d|C_k)$$

  - $\rightarrow$ Considerando, por exemplo, os d_1 primeiros atributos Gaussianos, os $d_2 - d_1$ seguintes binários (distribuição de Bernoulli) e os demais categóricos (distribuição multinoulli):

$$p(C_k|\boldsymbol{x}) \propto p(C_k) \prod_{d=1}^{d_1} \mathcal{N}(x_d|\mu_{dk}, \sigma_{dk}^2) \prod_{d=d_1+1}^{d_2} \mathrm{Ber}(x_d|q_{dk}) \prod_{d=d_2+1}^{D} \mathrm{Cat}(x_d|\boldsymbol{q}_{dk})$$

Agenda

Referências bibliográficas

- **Caps. 3 e 4** - MURPHY, Kevin P. **Machine learning: a probabilistic perspective**, 2012.