



UNIVERSIDADE
FEDERAL DO CEARÁ



Aprendizagem de Máquina

César Lincoln Cavalcante Mattos

2024

Agenda

- ① Classificadores estatísticos
- ② Classificadores Bayesianos
- ③ Tópicos adicionais
- ④ Referências

Classificadores estatísticos

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

Modelos discriminantes

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

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

Classificadores estatísticos

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

Modelos discriminantes

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

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

Modelos generativos

Modelam a distribuição das entradas associadas a cada classe.

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

Agenda

- ① Classificadores estatísticos
- ② Classificadores Bayesianos
- ③ Tópicos adicionais
- ④ Referências

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) \mathbf{x} de um padrão, a qual classe o padrão pertence?
- Pela Regra de Bayes:

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

Classificadores Bayesianos

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

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

- A notação pode ser simplificada:

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

Classificadores Bayesianos

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

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

- A notação pode ser simplificada:

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

- 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):

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

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

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

Classificadores Bayesianos

- Classificação binária (C_1 e C_2):

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

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

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

Classificadores Bayesianos

- Classificação binária (C_1 e C_2):

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

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

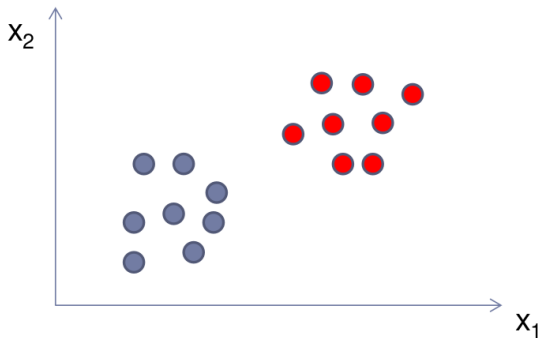
- **Ideia:** Escolha a classe com maior probabilidade.
- **Problema:** Como calcular as distribuições acima?
- **Ideia:** Estimar as probabilidades a partir do conjunto de treinamento.

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.

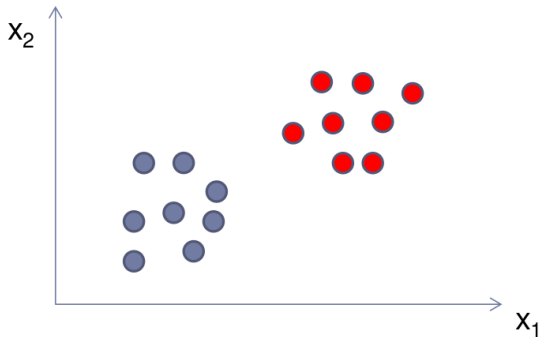


Classificadores Bayesianos

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

Classificadores Bayesianos

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



Análise de Discriminante Gaussiano

- Considerando distribuições Gaussianas, temos:

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

Análise de Discriminante Gaussiano

- Considerando distribuições Gaussianas, temos:

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

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

Análise de Discriminante Gaussiano

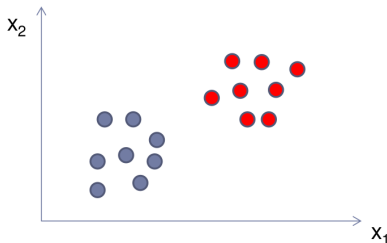
- Considerando distribuições Gaussianas, temos:

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

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

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{N_k} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i, \forall k \in \{1, 2\},$$

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



Análise de Discriminante Gaussiano

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

$$p(\mathbf{x}_* | C_k) = \mathcal{N}(\mathbf{x}_* | \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k), \quad k \in \{1, 2\}.$$

Análise de Discriminante Gaussiano

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

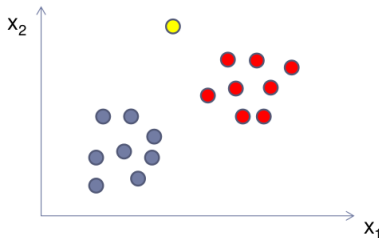
$$p(\mathbf{x}_* | C_k) = \mathcal{N}(\mathbf{x}_* | \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k), \quad k \in \{1, 2\}.$$

- Escolha a classe mais provável:

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

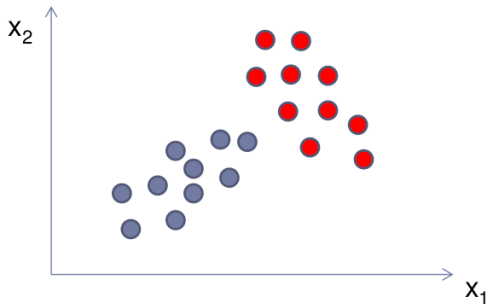
$$\log p(C_k | \mathbf{x}_*) \propto \log p(\mathbf{x}_* | C_k) + \log p(C_k), \quad k \in \{1, 2\},$$

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



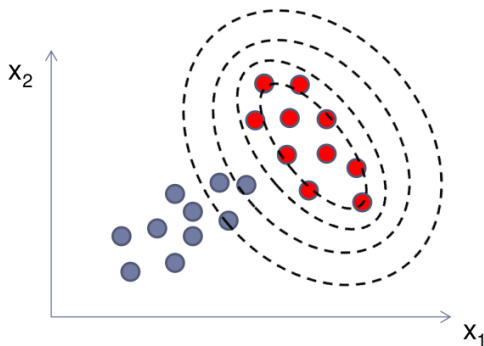
Análise de Discriminante Gaussiano

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



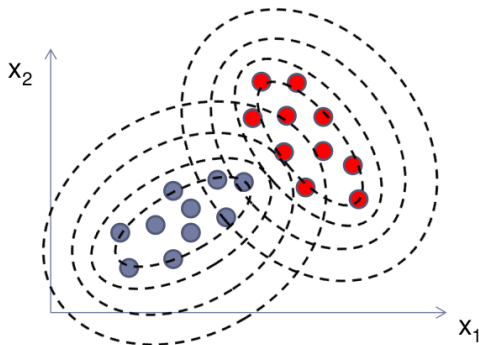
Análise de Discriminante Gaussiano

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



Análise de Discriminante Gaussiano

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



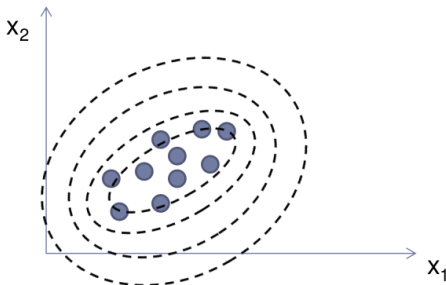
Análise de Discriminante Gaussiano

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

- Caso bidimensional ($D = 2$):

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{bmatrix}$$

- Valores altos para a covariância $\sigma_{1,2}$:



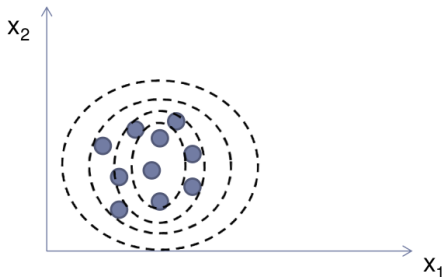
Análise de Discriminante Gaussiano

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

- Caso bidimensional ($D = 2$):

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{bmatrix}$$

- Valores baixos (próximos de zero) para a covariância $\sigma_{1,2}$:

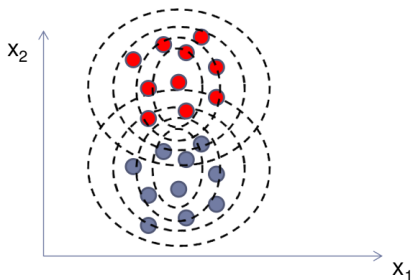


Classificadores Bayesianos

Naive Bayes

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

$$\Sigma_k = \begin{bmatrix} \sigma_{1k}^2 & 0 \\ 0 & \sigma_{2k}^2 \end{bmatrix}$$



Classificadores Bayesianos

Naive Bayes

- Dado $\mathbf{x} = [x_1, x_2, \dots, x_D]^\top$, calcula a probabilidade das classes:

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

Classificadores Bayesianos

Naive Bayes

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

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

- Considera atributos independentes dada a classe:

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

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

Classificadores Bayesianos

Naive Bayes

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

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

- Considera atributos independentes dada a classe:

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

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

- Predição para um novo padrão \mathbf{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 $p(x_d | C_k)$:

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

Classificadores Bayesianos

Naive Bayes Gaussiano

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

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

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

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

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

- **Observações:**

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

Resumo dos Classificadores Estatísticos

- **Análise de Discriminante Gaussiano**

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

- **Naive Bayes**

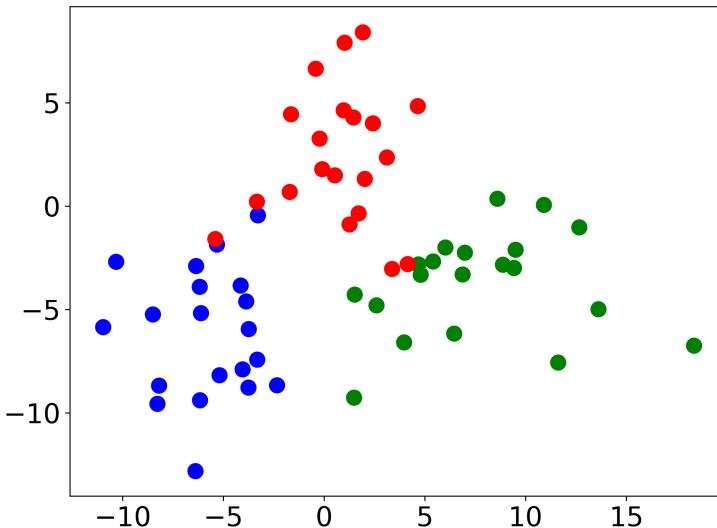
$$\hat{y}_* = \arg \max_k \left[\log p(C_k) + \sum_{d=1}^D \log p(x_{*d} | C_k) \right].$$

- **Naive Bayes Gaussiano**

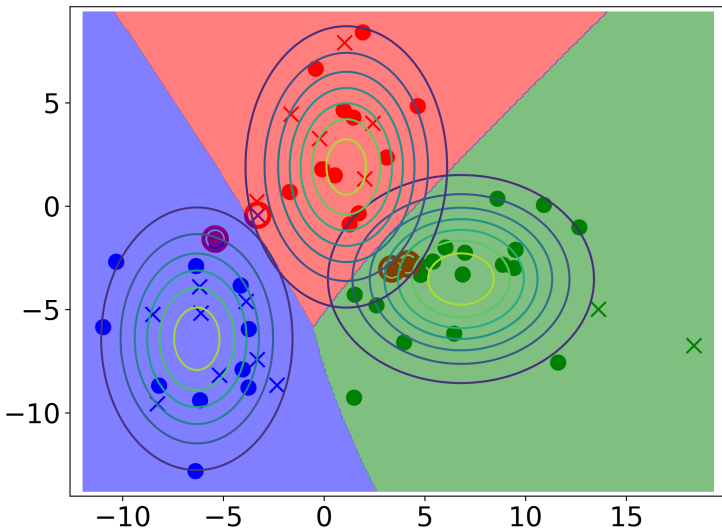
$$\hat{y}_* = \arg \max_k \left[\log p(C_k) - \frac{1}{2} \sum_{d=1}^D \log 2\pi \sigma_{dk}^2 - \frac{1}{2} \sum_{d=1}^D \frac{(x_{*d} - \mu_{dk})^2}{\sigma_{dk}^2} \right].$$

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

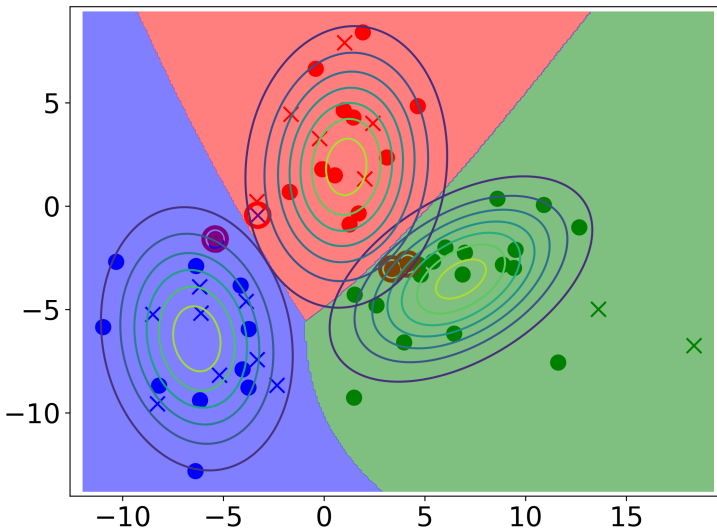
Classificadores Estatísticos



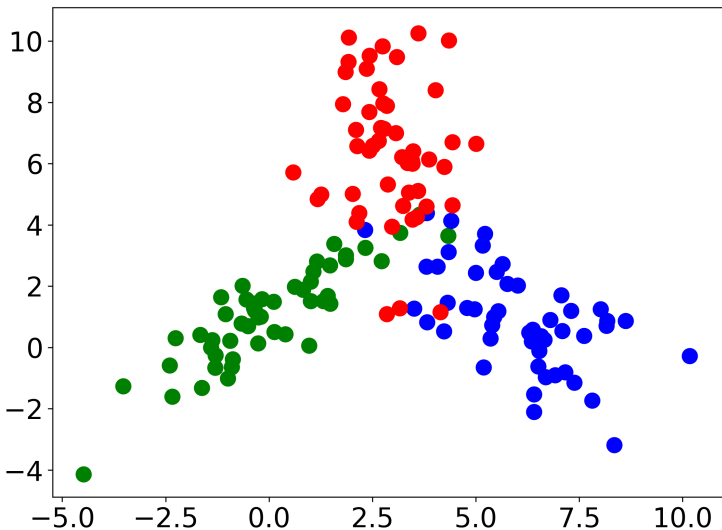
Naive Bayes Gaussiano



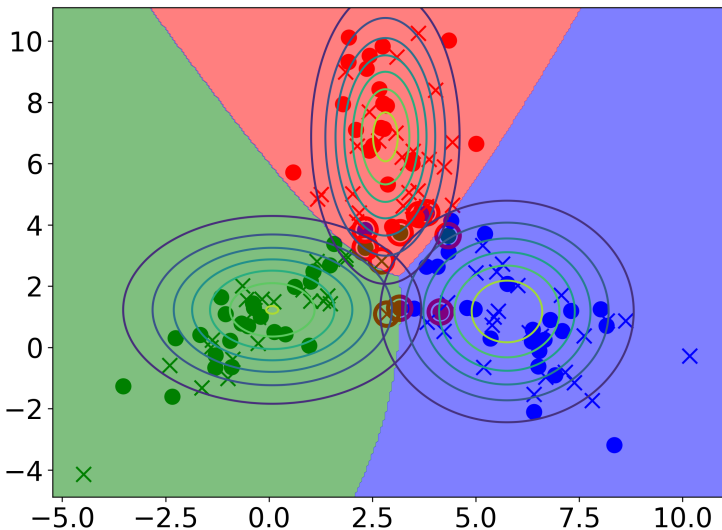
Análise de Discriminante Gaussiano



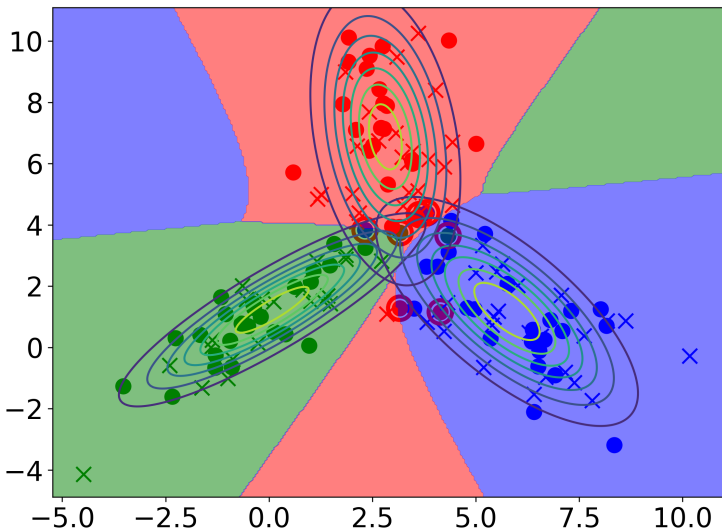
Classificadores Bayesianos



Naive Bayes Gaussiano



Análise de Discriminante Gaussiano

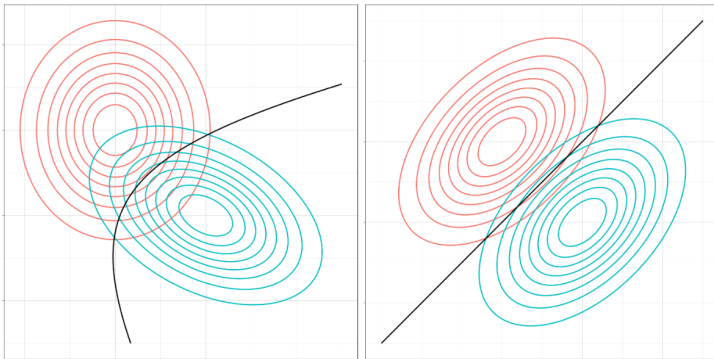


Resumo dos Classificadores Estatísticos

- O que impacta o formato das fronteiras de decisão?

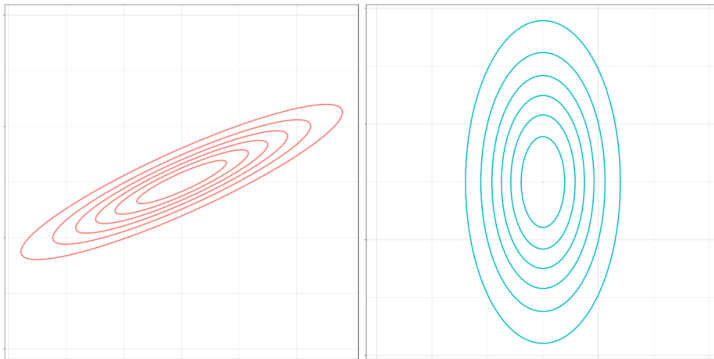
Resumo dos Classificadores Estatísticos

- O que impacta o formato das fronteiras de decisão?
- Matrizes de covariância das classes.
 - Se $\Sigma_j \neq \Sigma_k, \forall j, k$, fronteira quadrática.
 - Se $\Sigma_k = \Sigma, \forall k$, contrário, fronteira linear.



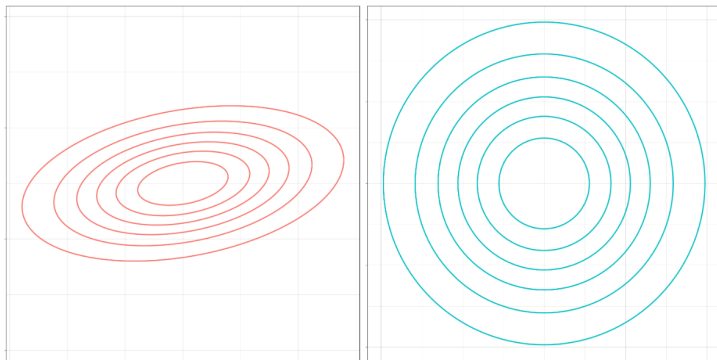
Resumo dos Classificadores Estatísticos

- O que impacta o formato das fronteiras de decisão?
- Correlação entre atributos.
 - Se atributos são correlacionados, contornos elípticos inclinados.
 - Se $\Sigma_k, \forall k$, é diagonal, contornos elípticos retos.



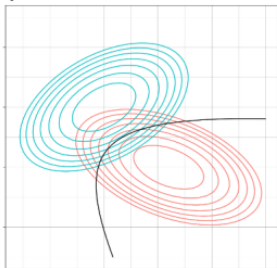
Resumo dos Classificadores Estatísticos

- O que impacta o formato das fronteiras de decisão?
- Correlação entre atributos.
 - Se atributos são correlacionados, contornos elípticos inclinados.
 - Se $\Sigma_k = \sigma_k^2 \mathbf{I}, \forall k$, contornos esféricos.

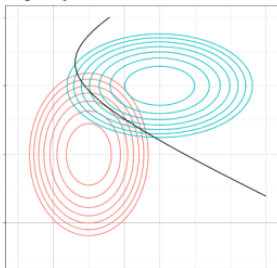


Resumo dos Classificadores Estatísticos

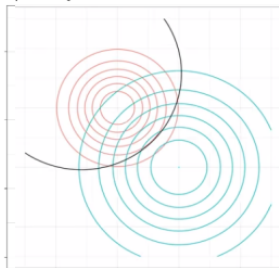
QDA



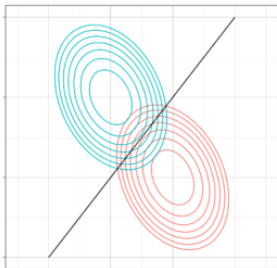
Diagonal QDA



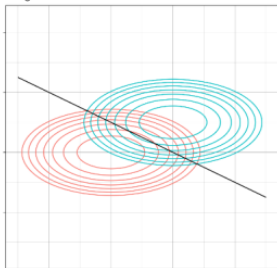
Spherical QDA



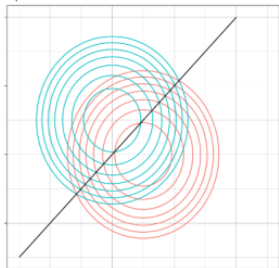
LDA



Diagonal LDA



Spherical LDA



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)?
 - ADG permite fronteiras de decisão não-lineares, dependendo das considerações feitas.
 - ADG considera que as distribuições $p(\mathbf{x}|C_k)$ são Gaussianas, o que não necessariamente é verdade.
 - ADG usualmente precisa de menos dados para obter uma boa solução.
 - ADG permite incluir ou remover classes sem precisar retreinar o modelo por completo.
 - NB permite incluir ou remover atributos sem precisar retreinar o modelo por completo.
 - RL é mais robusta quando considerações incorretas são feitas.

Agenda

- ① Classificadores estatísticos
- ② Classificadores Bayesianos
- ③ Tópicos adicionais
- ④ Referências

Tópicos adicionais

- Classificadores Naive Bayes não-Gaussianos ou mistos.

→ Para classes $k \in \{1, \dots, K\}$:

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

→ 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):

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

- Inferência Bayesiana: adicionar prioris para $p(C_k)$ (e.g. Dirichlet) e os parâmetros das distribuições (e.g. Gaussiana, gamma inversa, beta, Dirichlet).

Agenda

- ① Classificadores estatísticos
- ② Classificadores Bayesianos
- ③ Tópicos adicionais
- ④ Referências

Referências bibliográficas

- **Caps. 3 e 4 - MURPHY, Kevin P. Machine learning: a probabilistic perspective, 2012.**
- **Cap. 9 - MURPHY, Kevin P. Probabilistic Machine Learning: An Introduction, 2021.**
- **Caps. 1 e 8* - BISHOP, Christopher M. Pattern recognition and machine learning, 2006.**