



UNIVERSIDADE
FEDERAL DO CEARÁ



Aprendizagem de Máquina

César Lincoln Cavalcante Mattos

2024

Agenda

① Avaliação de modelos

Ajuste e avaliação de modelos

Avaliação de modelos de regressão

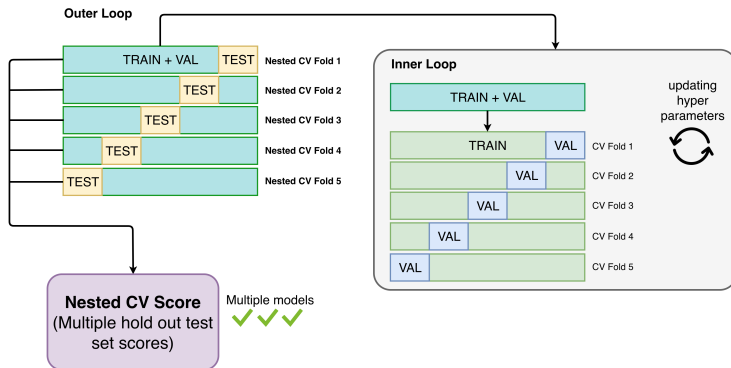
Avaliação de modelos de classificação

② Tópicos adicionais sobre avaliação de modelos

③ Referências

Ajuste e avaliação de modelos

- O ajuste de hiperparâmetros deve ser considerado na avaliação.
- Uma abordagem possível é a **validação cruzada aninhada** (*nested cross-validation*).



- A escolha final do loop externo deve ser reajustada no loop interno usando todos os dados, resultando na solução final.

Avaliação de modelos de regressão

- Sendo $\hat{y}_i = f(\mathbf{x}_i)$ a i -ésima predição referente à entrada \mathbf{x}_i e y_i a saída real, temos:

→ **RMSE (root mean squared error):** $\sqrt{\frac{1}{N} \sum_i (y_i - \hat{y}_i)^2}$;

→ **MAE (mean absolute error):** $\frac{1}{N} \sum_i |y_i - \hat{y}_i|$;

→ **MRE (mean relative error):** $\frac{1}{N} \sum_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|$.

Avaliação de modelos de regressão

- Sendo $\hat{y}_i = f(\mathbf{x}_i)$ a i -ésima predição referente à entrada \mathbf{x}_i e y_i a saída real, temos:
 - **RMSE (root mean squared error)**: $\sqrt{\frac{1}{N} \sum_i (y_i - \hat{y}_i)^2}$;
 - **MAE (mean absolute error)**: $\frac{1}{N} \sum_i |y_i - \hat{y}_i|$;
 - **MRE (mean relative error)**: $\frac{1}{N} \sum_i \left| \frac{y_i - \hat{y}_i}{y_i} \right|$.

Observações

- O RMSE é proporcional ao quadrado dos erros, sendo mais afetado por dados discrepantes.
- O MAE é diretamente proporcional aos erros absolutos, sendo mais fácil de compreender.
- O MRE é invariante à magnitude dos dados, o que pode facilitar a interpretação.

Avaliação de modelos de regressão

- Considere uma perda L_i calculada para o i -ésimo exemplo.
- O **erro padrão** (*standard error*) pode ser computado por:

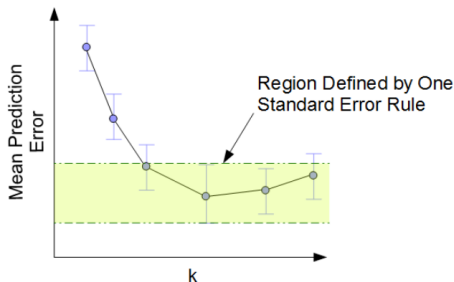
$$SE = \frac{\hat{\sigma}}{\sqrt{N}}, \quad \hat{\sigma}^2 = \frac{1}{N-1} \sum_i^N (L_i - \bar{L})^2, \quad \bar{L} = \frac{1}{N} \sum_i^N L_i.$$

Avaliação de modelos de regressão

- Considere uma perda L_i calculada para o i -ésimo exemplo.
- O **erro padrão** (*standard error*) pode ser computado por:

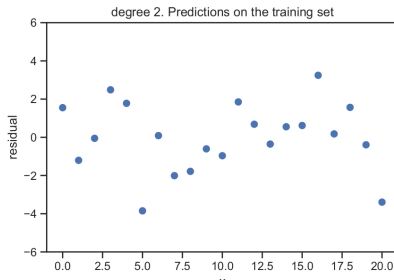
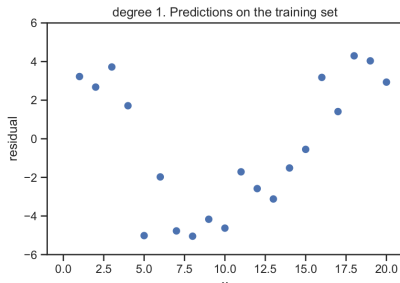
$$SE = \frac{\hat{\sigma}}{\sqrt{N}}, \quad \hat{\sigma}^2 = \frac{1}{N-1} \sum_i (L_i - \bar{L})^2, \quad \bar{L} = \frac{1}{N} \sum_i L_i.$$

- **One standard error rule:** Na avaliação de múltiplos modelos, escolha o mais simples (e.g. menos parâmetros) que esteja no máximo a um erro padrão de diferença do melhor modelo.



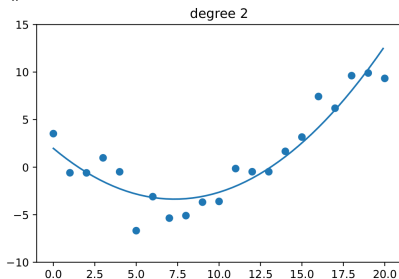
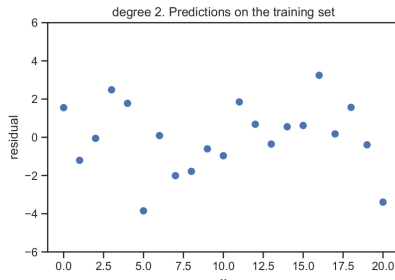
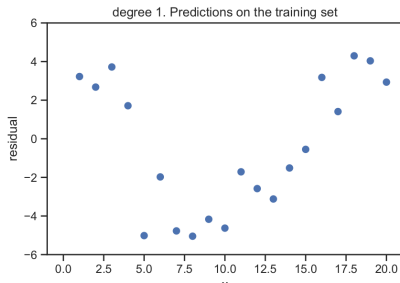
Análise de resíduos

- Resíduos são os erros obtidos no conjunto de treinamento.
- A análise gráfica dos resíduos pode ajudar a avaliar o modelo.



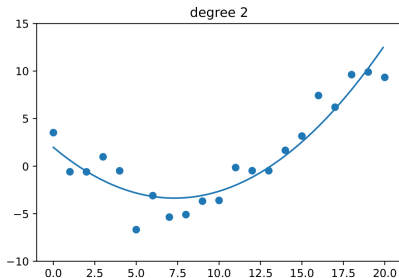
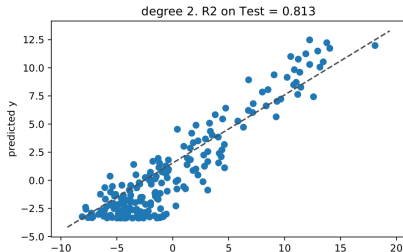
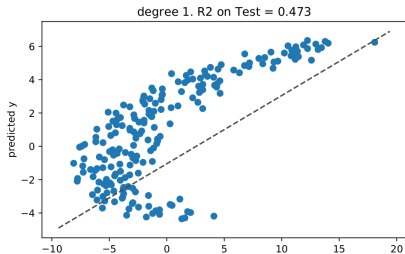
Análise de resíduos

- Resíduos são os erros obtidos no conjunto de treinamento.
- A análise gráfica dos resíduos pode ajudar a avaliar o modelo.



Análise de resíduos

- Podemos também analisar diretamente os valores esperados versus os preditos.



Análise de resíduos

- A qualidade do ajuste também pode ser medida pelo coeficiente de determinação R^2 :

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \quad \bar{y} = \frac{1}{N} \sum_i y_i.$$

- O numerador é a soma residual dos quadrados (*residual sum of squares*).
- O denominador é a soma total dos quadrados (*total sum of squares*).
- Em geral, $0 \leq R^2 \leq 1$. Valores altos indicam maior redução na variância em relação a uma predição constante $\hat{y}_i = \bar{y}$.

Análise de resíduos

- A qualidade do ajuste também pode ser medida pelo coeficiente de determinação R^2 :

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \quad \bar{y} = \frac{1}{N} \sum_i y_i.$$

- O numerador é a soma residual dos quadrados (*residual sum of squares*).
- O denominador é a soma total dos quadrados (*total sum of squares*).
- Em geral, $0 \leq R^2 \leq 1$. Valores altos indicam maior redução na variância em relação a uma predição constante $\hat{y}_i = \bar{y}$.
- **Importante:** O valor de R^2 não é indicado para avaliar modelos não-lineares nos parâmetros!

Avaliação de classificadores

- Considere uma classificação binária, com classes 0 e 1.
- Seja y_* a classe correta e \hat{y} a classe predita.
- A função de perda zero-um é definida por:

$$l_{01} = \mathbb{I}(y_* \neq \hat{y}).$$

- A perda esperada *a posteriori* pode ser escrita por:

$$p(y_* \neq \hat{y} | \mathbf{x}) = 1 - p(y_* = \hat{y} | \mathbf{x}).$$

- A classificação que minimiza a perda esperada será:

$$\hat{y} = \arg \max_y p(y | \mathbf{x}),$$

que equivale à solução MAP (*maximum a posteriori*).

Avaliação de classificadores

- **Problema:** Suponha que o custo de errar a classe 0 (c_{01}) seja diferente de errar a classe 1 (c_{10}).

Avaliação de classificadores

- **Problema:** Suponha que o custo de errar a classe 0 (c_{01}) seja diferente de errar a classe 1 (c_{10}).
- **Ideia:** Vamos considerar os custos na tomada de decisão.
- Escolhemos a classe 1 se:

$$p(y = 1|\mathbf{x})c_{10} > p(y = 0|\mathbf{x})c_{01}$$

$$p(y = 1|\mathbf{x})c_{10} > [1 - p(y = 1|\mathbf{x})]c_{01}$$

$$p(y = 1|\mathbf{x}) > \frac{c_{01}}{c_{01} + c_{10}}.$$

Avaliação de classificadores

- **Problema:** Suponha que o custo de errar a classe 0 (c_{01}) seja diferente de errar a classe 1 (c_{10}).
- **Ideia:** Vamos considerar os custos na tomada de decisão.
- Escolhemos a classe 1 se:

$$\begin{aligned}p(y = 1|\mathbf{x})c_{10} &> p(y = 0|\mathbf{x})c_{01} \\p(y = 1|\mathbf{x})c_{10} &> [1 - p(y = 1|\mathbf{x})]c_{01} \\p(y = 1|\mathbf{x}) &> \frac{c_{01}}{c_{01} + c_{10}}.\end{aligned}$$

- Caso $c_{10} = Rc_{01}$, temos:

$$p(y = 1|\mathbf{x}) > \frac{1}{R + 1}.$$

Avaliação de classificadores

- **Problema:** Seria possível o modelo não retornar nenhuma classe durante uma predição?

Avaliação de classificadores

- **Problema:** Seria possível o modelo não retornar nenhuma classe durante uma predição?
- **Ideia:** Podemos considerar uma “opção de rejeição”:

$$y_* = \begin{cases} C_*, & \text{se } p(y_* = C_*|\mathbf{x}) > \tau_*, C_* = \arg \max_C p(y_* = C|\mathbf{x}) \\ \text{rejeita}, & \text{caso contrário} \end{cases}$$

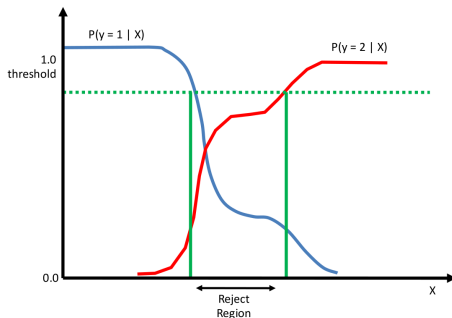
em que $\frac{1}{K} < \tau_* = 1 - \frac{\tau_r}{\tau_e} < 1$ é um limiar calculado a partir do custo da ação de rejeição τ_r e o custo do erro de classificação τ_e .

Avaliação de classificadores

- **Problema:** Seria possível o modelo não retornar nenhuma classe durante uma predição?
- **Ideia:** Podemos considerar uma “opção de rejeição”:

$$y_* = \begin{cases} C_*, & \text{se } p(y_* = C_* | \mathbf{x}) > \tau_*, C_* = \arg \max_C p(y_* = C | \mathbf{x}) \\ \text{rejeita,} & \text{caso contrário} \end{cases}$$

em que $\frac{1}{K} < \tau_* = 1 - \frac{\tau_r}{\tau_e} < 1$ é um limiar calculado a partir do custo da ação de rejeição τ_r e o custo do erro de classificação τ_e .

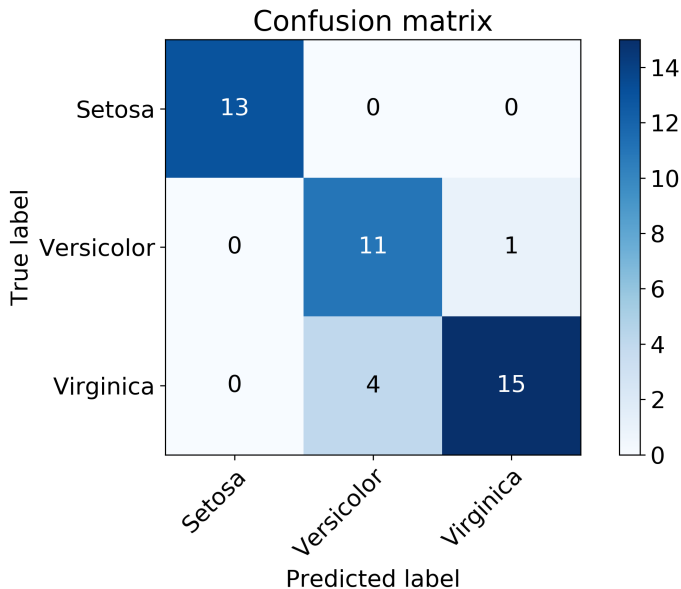


Avaliação de classificadores

Matriz de confusão

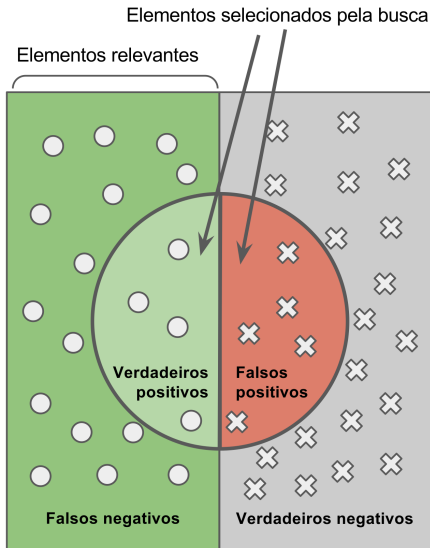
- Matriz que **sumariza os acertos e erros** de um classificador.
- Normalmente as classes (rótulos) verdadeiras são colocadas no eixo vertical, enquanto as classes preditas ficam no eixo horizontal.
- Os elementos na diagonal principal da matriz correspondem aos acertos do classificador.
- Os demais elementos correspondem aos erros do classificador.
- Classificadores obtidos por **algoritmos diferentes podem obter erros diferentes**, mesmo que a taxa de erro total seja semelhante.

Matriz de confusão - Arvore de decisão - Flores íris



Classificação binária (“positivo” ou “negativo”)

Precisão e Revocação



$$\text{Precisão} = \frac{\text{Verdadeiros positivos}}{\text{Verdadeiros positivos} + \text{Falsos positivos}}$$

"Quantos elementos selecionados são relevantes?"

$$\text{Revocação} = \frac{\text{Verdadeiros positivos}}{\text{Verdadeiros positivos} + \text{Falsos negativos}}$$

"Quantos elementos relevantes foram selecionados?"

Avaliação de classificadores binários

- **Revocação (sensibilidade, recall ou taxa de verdadeiros positivos):**

$$\text{revocação} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos negativos}}$$

Avaliação de classificadores binários

- **Revocação (sensibilidade, recall ou taxa de verdadeiros positivos):**

$$\text{revocação} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos negativos}}$$

- **Precisão (precision ou valor preditivo positivo):**

$$\text{precisão} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos positivos}}$$

Avaliação de classificadores binários

- **Revocação (sensibilidade, recall ou taxa de verdadeiros positivos):**

$$\text{revocação} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos negativos}}$$

- **Precisão (precision ou valor preditivo positivo):**

$$\text{precisão} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos positivos}}$$

- **F_1 score (F-score ou F-measure):**

$$F_1 = \left(\frac{\text{revocação}^{-1} + \text{precisão}^{-1}}{2} \right)^{-1} = 2 \frac{\text{revocação} \times \text{precisão}}{\text{revocação} + \text{precisão}} \in [0, 1]$$

Avaliação de classificadores binários

- **Revocação (sensibilidade, recall ou taxa de verdadeiros positivos):**

$$\text{revocação} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos negativos}}$$

- **Precisão (precision ou valor preditivo positivo):**

$$\text{precisão} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos positivos}}$$

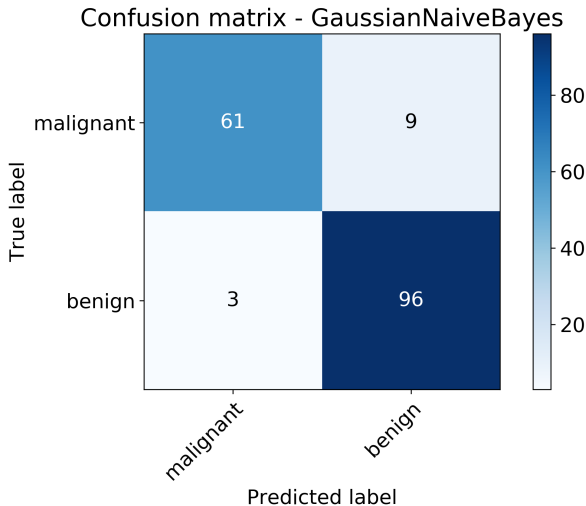
- **F_1 score (F-score ou F-measure):**

$$F_1 = \left(\frac{\text{revocação}^{-1} + \text{precisão}^{-1}}{2} \right)^{-1} = 2 \frac{\text{revocação} \times \text{precisão}}{\text{revocação} + \text{precisão}} \in [0, 1]$$

- **F_β score (revocação β vezes mais importante que a precisão):**

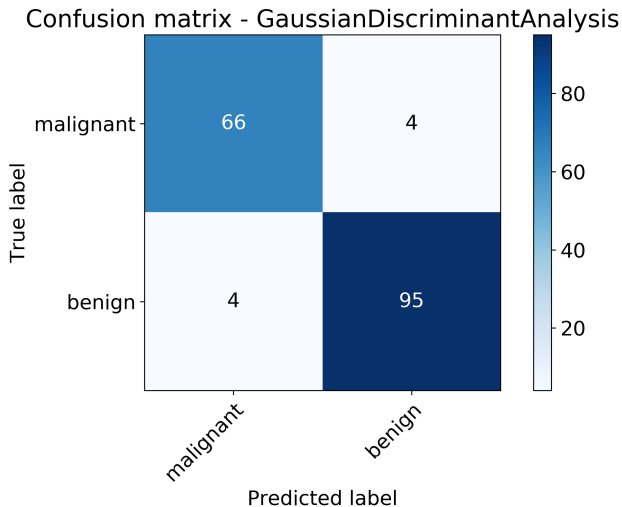
$$F_\beta = (1 + \beta^2) \frac{\text{revocação} \times \text{precisão}}{\text{revocação} + \beta^2 \text{precisão}} \in [0, 1]$$

Naive Bayes Gaussiano - Breast Cancer



$$\text{revocação} = \frac{61}{61+9} \approx 0.8714, \quad \text{precisão} = \frac{61}{61+3} \approx 0.9531, \quad F_1 \approx 0.9104$$

Discriminante Gaussiano - Breast Cancer



$$\text{revocação} = \frac{66}{66+4} \approx 0.9429, \quad \text{precisão} = \frac{66}{66+4} \approx 0.9429, \quad F_1 \approx 0.9429$$

Curva ROC (receiver operating characteristic)

- Em classificadores binários, temos $\hat{y}_* = \begin{cases} 1, & \text{se } p(\hat{y}_*|\mathbf{x}_*) \geq \tau, \\ 0, & \text{caso contrário.} \end{cases}$

Curva ROC (receiver operating characteristic)

- Em classificadores binários, temos $\hat{y}_* = \begin{cases} 1, & \text{se } p(\hat{y}_*|\mathbf{x}_*) \geq \tau, \\ 0, & \text{caso contrário.} \end{cases}$
- Apesar do valor $\tau = 0.5$ ser usual, podemos usar $\tau \in [0, 1]$.
- A **curva ROC** é obtida para diversos valores de $\tau \in [0, 1]$.

$$\text{taxa de falsos positivos (FPR)} = \frac{\text{falsos positivos}}{\text{falsos positivos} + \text{verdadeiros negativos}} \text{ e}$$

$$\text{taxa de verdadeiros positivos (TPR)} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos negativos}}$$

Curva ROC (receiver operating characteristic)

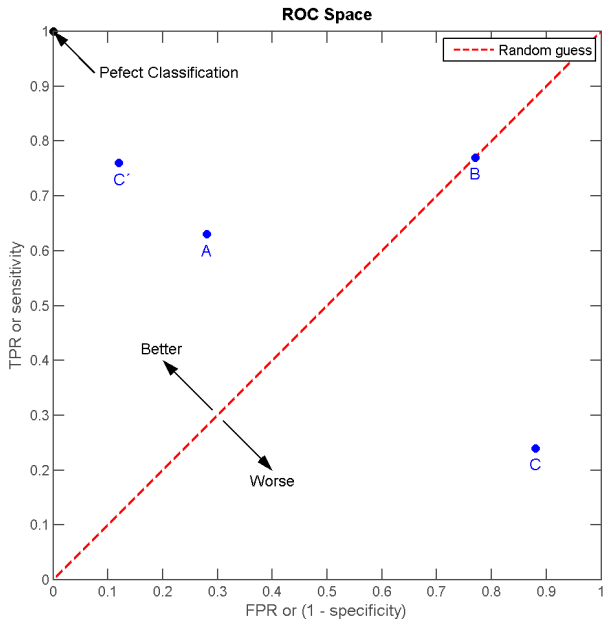
- Em classificadores binários, temos $\hat{y}_* = \begin{cases} 1, & \text{se } p(\hat{y}_*|\mathbf{x}_*) \geq \tau, \\ 0, & \text{caso contrário.} \end{cases}$
- Apesar do valor $\tau = 0.5$ ser usual, podemos usar $\tau \in [0, 1]$.
- A **curva ROC** é obtida para diversos valores de $\tau \in [0, 1]$.

$$\text{taxa de falsos positivos (FPR)} = \frac{\text{falsos positivos}}{\text{falsos positivos} + \text{verdadeiros negativos}} \text{ e}$$

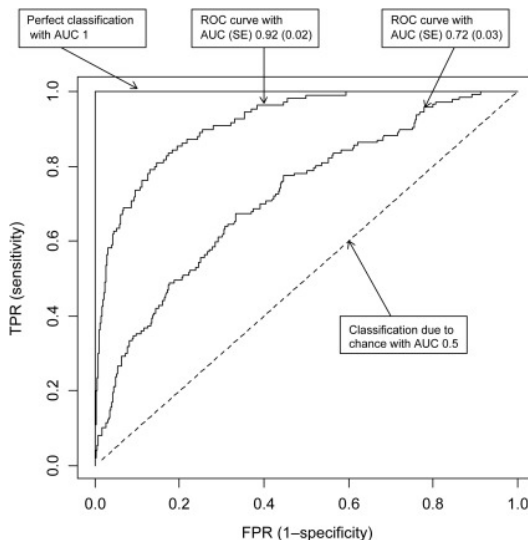
$$\text{taxa de verdadeiros positivos (TPR)} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos negativos}}$$

- **AUROC (area under the ROC curve)**: Área abaixo da curva ROC de um classificador. Seu pior valor é 0 e o melhor é 1.
 - Probabilidade do classificador atribuir um valor maior a um padrão positivo qualquer comparado a um negativo qualquer.
- **EER (equal error rate) ou cross-over rate**: Valor em que $\text{FPR} = 1 - \text{TPR}$. Seu melhor valor é 0.

Curva ROC - Ilustração



Curva ROC - Ilustração



- **Observação:** Um classificador aleatório possui uma curva ROC em que $TPR = FPR$.

Curva Precision-Recall (PRC)

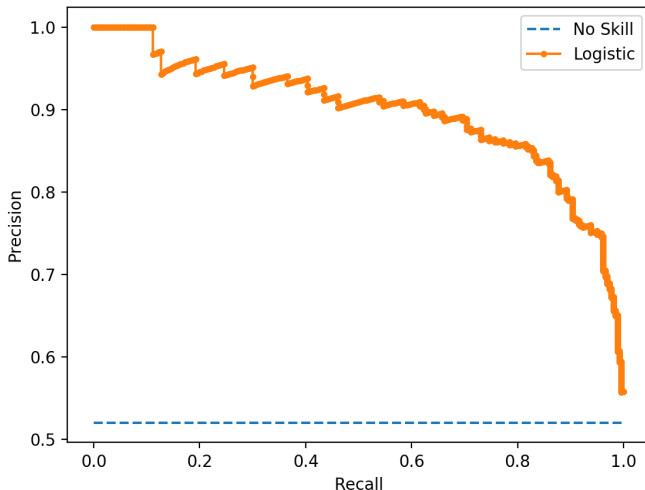
- Em classificadores binários, temos $\hat{y}_* = \begin{cases} 1, & \text{se } p(\hat{y}_*|\mathbf{x}_*) \geq \tau, \\ 0, & \text{caso contrário.} \end{cases}$
- A **curva PR** é obtida quando computamos

$$\text{revocação} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos negativos}} \text{ e}$$
$$\text{precisão} = \frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos positivos}}$$

para diversos valores de $\tau \in [0, 1]$.

- **AUPRC (area under the PR curve)**: Área abaixo da curva PR de um classificador. Corresponde à **precisão média**.

Curva PR - Ilustração

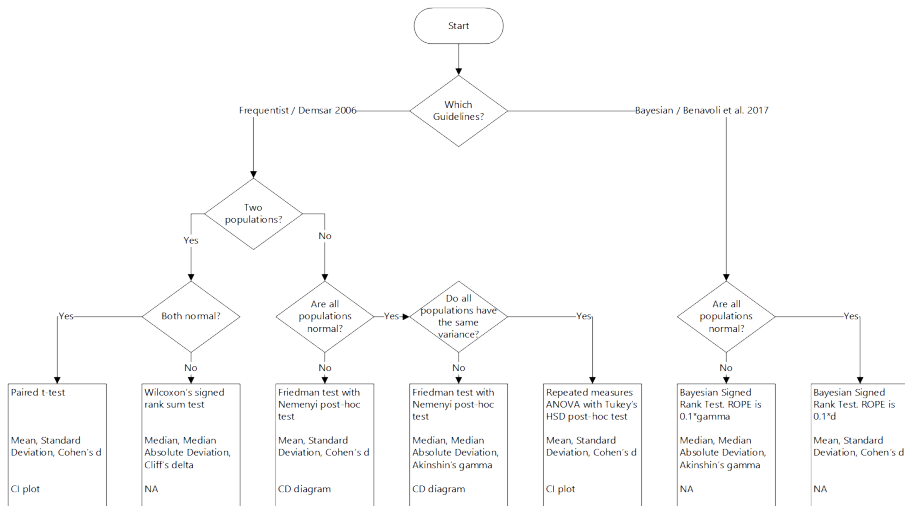


- **Observação:** Um classificador aleatório possui uma curva PR constante igual à proporção de exemplos positivos.

Tópicos adicionais sobre avaliação de modelos

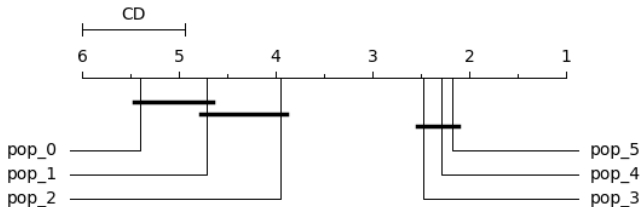
- DEMŠAR, Janez. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research, v. 7, p. 1-30, 2006.
- BENAOLLI, Alessio et al. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. Journal of Machine Learning Research, v. 18, n. 77, p. 1-36, 2017.
- Pacote Python para comparação estatística de múltiplos modelos: Autorank.

Tópicos adicionais sobre avaliação de modelos



Tópicos adicionais sobre avaliação de modelos

- Exemplo de teste não paramétrico de Friedman seguido de teste *post-hoc* Nemenyi (diagrama de distância crítica):



Referências bibliográficas

- **Caps. 5 e 16** - MURPHY, Kevin P. **Machine learning: a probabilistic perspective**, 2012.
- **Cap. 18** - MURPHY, Kevin P. **Probabilistic Machine Learning: An Introduction**, 2021.
- **Cap. 14** - BISHOP, C. **Pattern recognition and machine learning**, 2006.