

Universidade do Minho
Escola de Engenharia
Departamento de Informática

ADI

Avaliação de Modelos

Licenciatura em Engenharia Informática, 3º ano
Mestrado integrado em Engenharia Informática, 4º ano

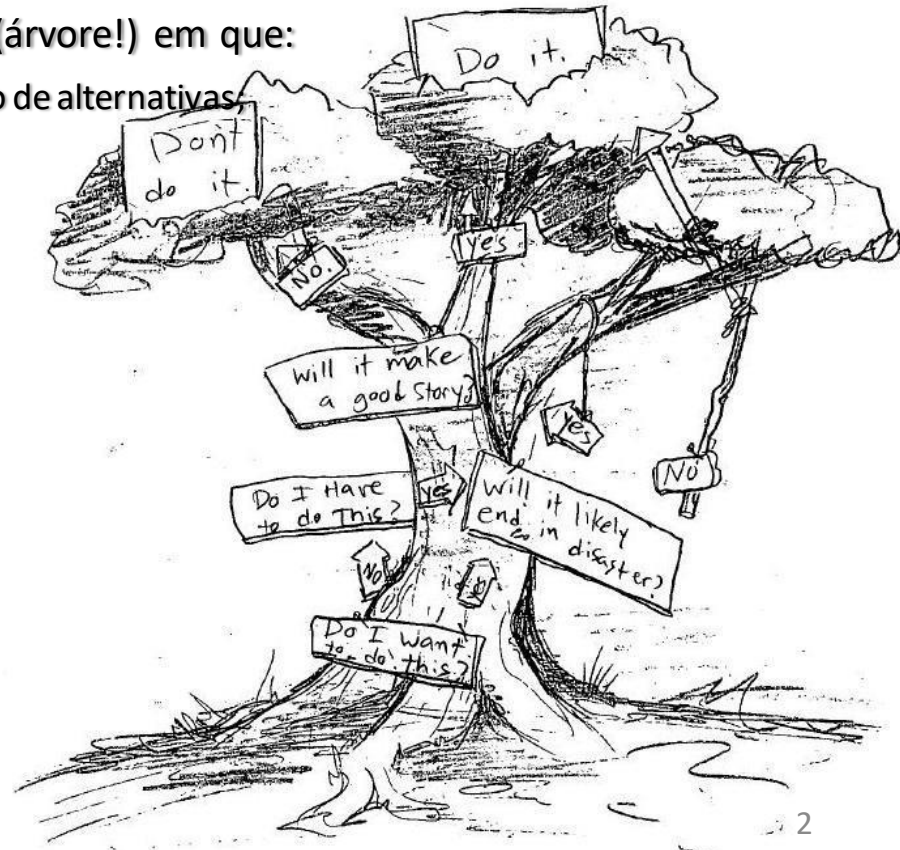


ISLab

Synthetic Intelligence Lab

Um Modelo: Árvores de Decisão (Decision Trees)

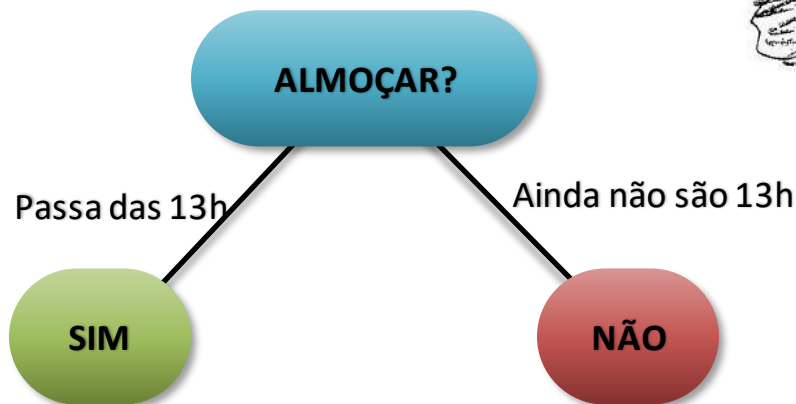
- Uma Árvore de Decisão é um grafo hierarquizado (árvore!) em que:
 - Cada ramo representa a seleção entre um conjunto de alternativas;
 - Cada folha representa uma decisão;





Um Modelo: Árvores de Decisão (Decision Trees)

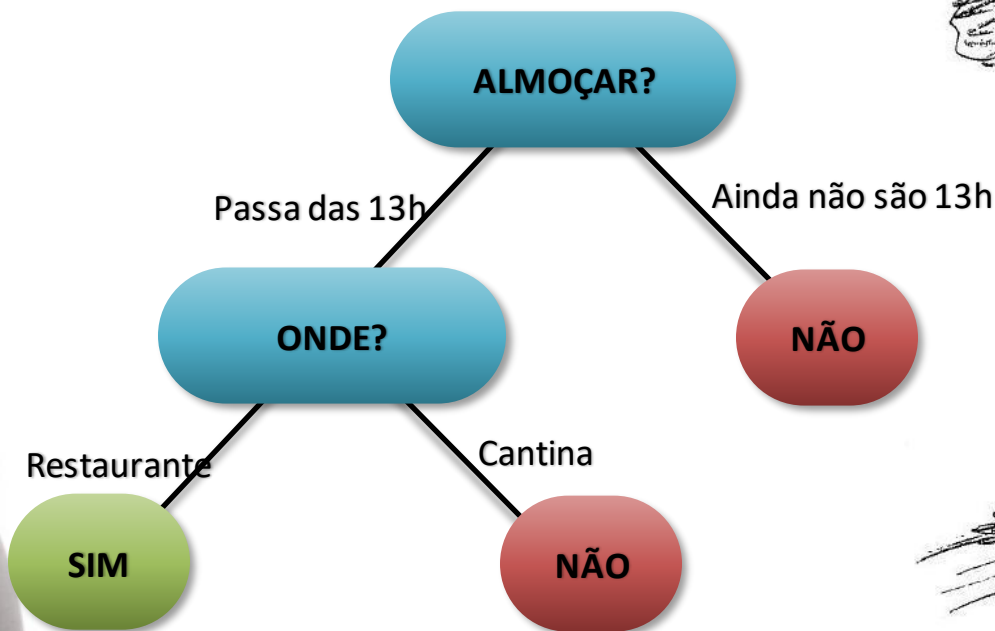
- Uma Árvore de Decisão é um grafo hierarquizado (árvore!) em que:
 - Cada ramo representa a seleção entre um conjunto de alternativas;
 - Cada folha representa uma decisão;





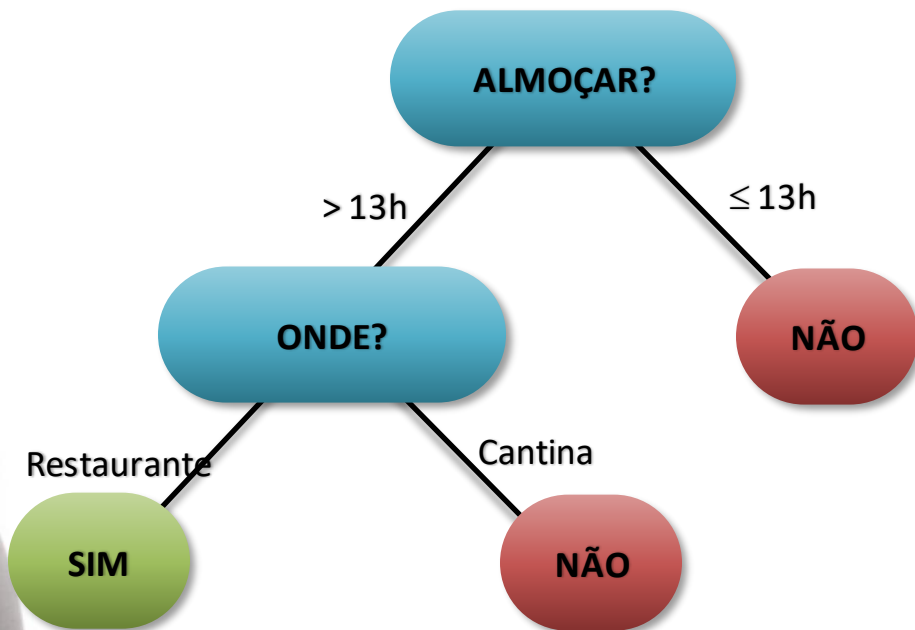
Um Modelo: Árvores de Decisão (Decision Trees)

- Uma Árvore de Decisão é um grafo hierarquizado (árvore!) em que:
 - Cada ramo representa a seleção entre um conjunto de alternativas;
 - Cada folha representa uma decisão;



Um Modelo: Árvores de Decisão (Decision Trees)

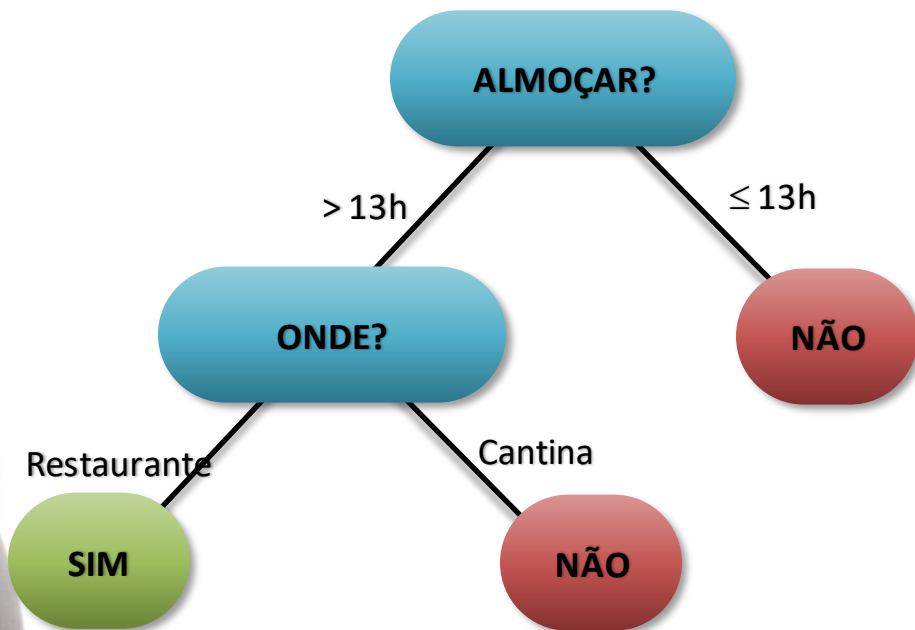
- Uma Árvore de Decisão é um grafo hierarquizado (árvore!) em que:
 - Cada nodo interno testa um atributo do *dataset*;
 - Cada ramo identifica um valor (ou conjunto de valores) do nodo testado;
 - Cada folha representa uma decisão;



| ALMOÇAR | ONDE | DECISÃO |
|---------|-------------|---------|
| 12h30 | Cantina | |
| 13h15 | Cantina | |
| 13h10 | Restaurante | |
| 11h00 | Restaurante | |
| 13:30 | Cantina | |

Um Modelo: Árvores de Decisão (Decision Trees)

- Uma Árvore de Decisão é um grafo hierarquizado (árvore!) em que:
 - Cada nodo interno testa um atributo do *dataset*;
 - Cada ramo identifica um valor (ou conjunto de valores) do nodo testado;
 - Cada folha representa uma decisão;

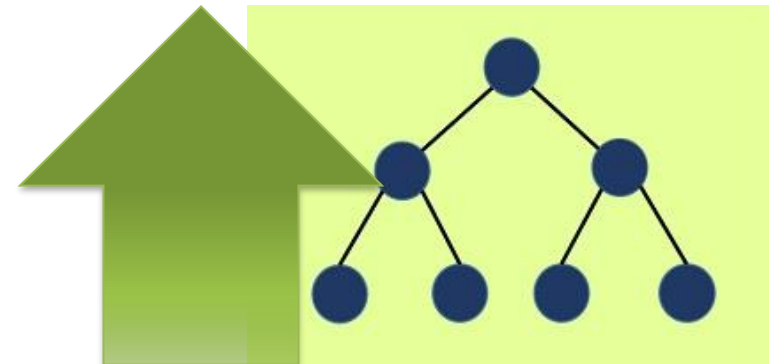
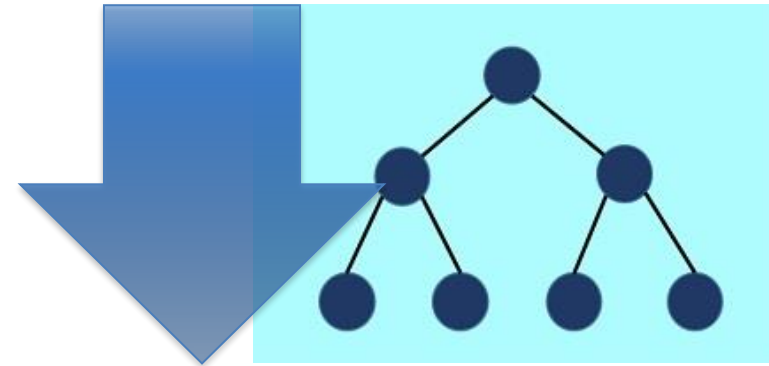


| ALMOÇAR | ONDE | DECISÃO |
|---------|-------------|---------|
| 12h30 | Cantina | NÃO |
| 13h15 | Cantina | NÃO |
| 13h10 | Restaurante | SIM |
| 11h00 | Restaurante | NÃO |
| 13:30 | Cantina | NÃO |



Modelos de Decisão

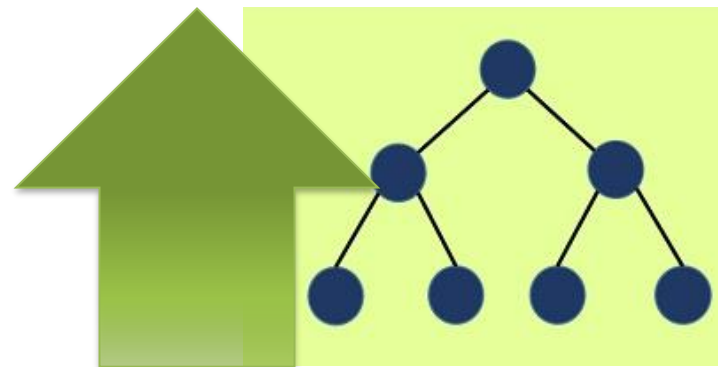
- Paradigmas de criação de modelos de decisão:
 - Top-down:
 - O modelo é construído a partir do conhecimento de especialistas;
 - O “todo” é dividido em “partes”;
 - Bottom-up:
 - O modelo é construído pela identificação de relações entre os atributos do *dataset*;
 - O modelo é induzido por “generalização” dos dados;



Modelos de Decisão

Árvores de Decisão

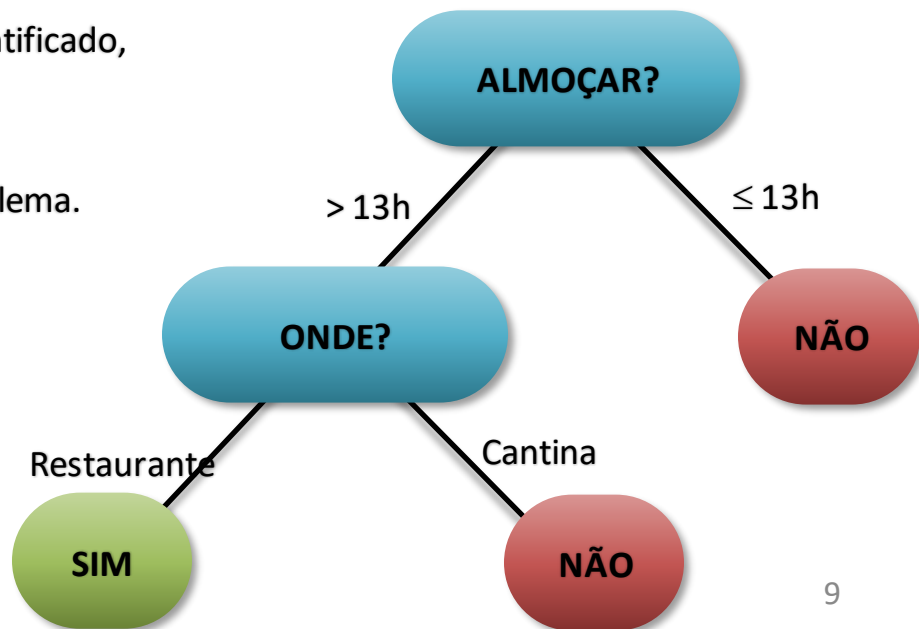
- Árvores de Decisão seguem o Paradigma Bottom-up:
 - Toda a informação sobre cada item de dados (ou objeto) deve estar definido numa coleção fixa e finita de atributos;
 - Deste modo, objetos distintos não podem requerer coleções distintas de atributos;
 - Quando o conjunto dos níveis de decisão é conhecido *a priori*, a construção do modelo segue um paradigma de aprendizagem supervisionado;
 - Quando o conjunto dos níveis de decisão é calculado pelo modelo, a sua construção segue um paradigma de aprendizagem não supervisionado;
 - Os níveis de decisão podem ser de 2 tipos:
 - Discretos: problemas de classificação;
 - Contínuos: problemas de regressão;
 - Quantidade de objetos \gg níveis de decisão;



Ciclo de Execução Resolução de Problemas

- Dada uma árvore de decisão (treinada), o processo de decisão desenvolve-se do seguinte modo:

1. Começar no nodo correspondente ao atributo “raiz”;
2. Identificar o valor do atributo;
3. Seguir pelo ramo correspondente ao valor identificado,
4. Alcançar o nodo relativo ao ramo percorrido;
5. Voltar a 2. até que o nodo seja uma “folha”;
6. O nodo alcançado indica a decisão para o problema.





ISLab

Synthetic Intelligence Lab

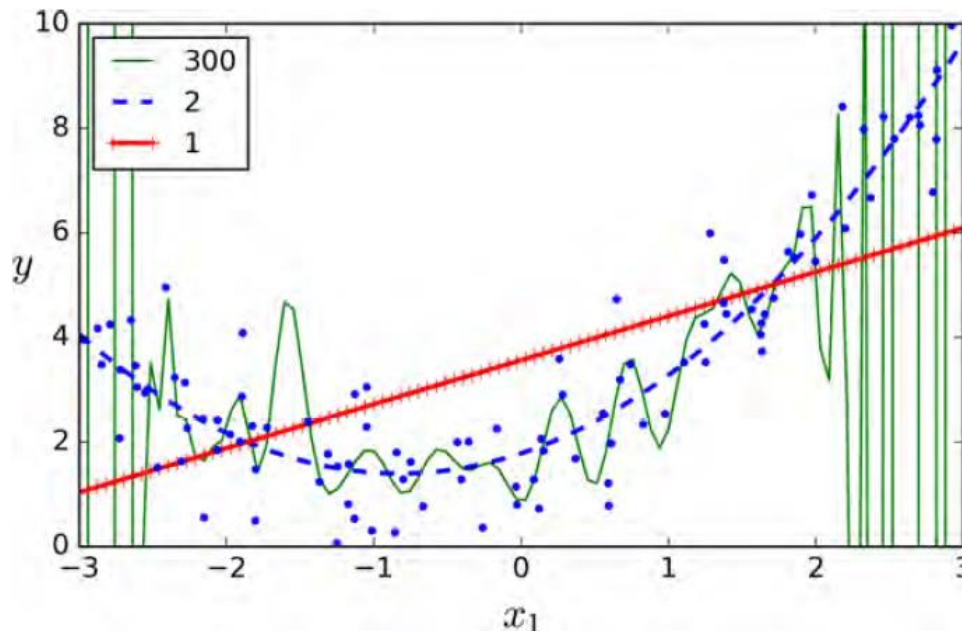
Árvores de Decisão Classificação

- Uma Árvore de Decisão pode ser utilizada para fazer **classificação**:
 - Decidir sobre se ou onde almoçar: classificação binária (SIM/NÃO)
 - Prever quem sobreviveu ao acidente do Titanic: classificação binária (SIM/NÃO)
 - Classificar um conjunto de imagens: classificação múltipla (laranja, kiwi, romã, ...)



Árvores de Decisão Regressão

- Uma Árvore de Decisão pode ser utilizada para fazer **regressão**:
 - Regressão linear, polinomial, múltipla, entre outras;
 - Prever o preço do petróleo/gás/combustíveis: escala contínua ou real, em € ou \$
 - Estimar a temperatura para o dia de amanhã: escala continua, em °C ou °F



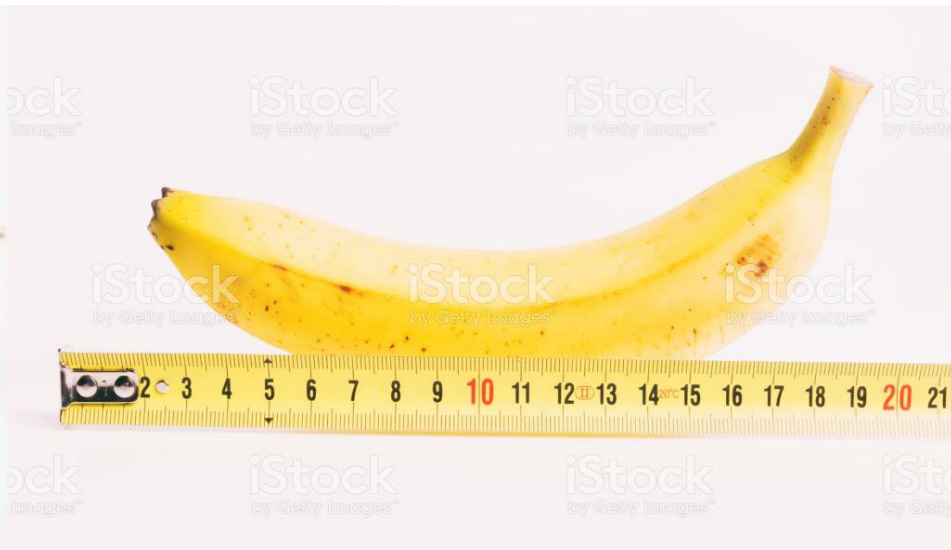


ISLab

Synthetic Intelligence Lab

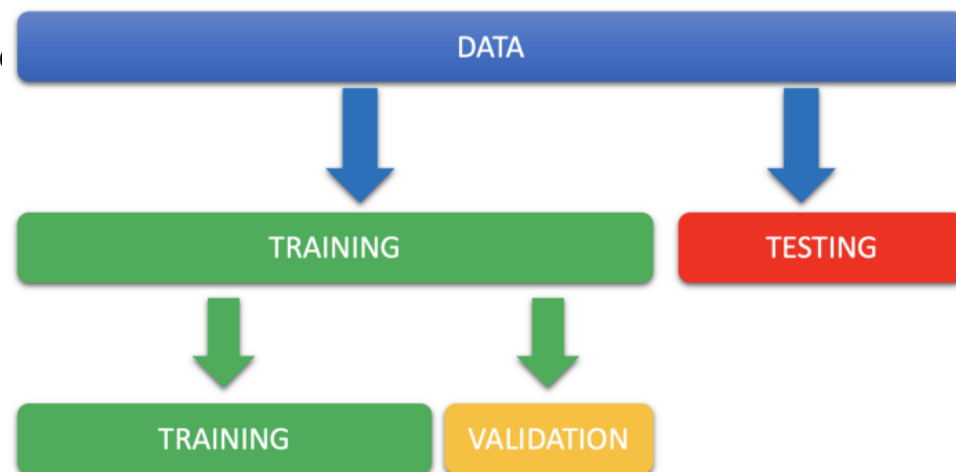
Avaliação de Modelos

- Após a criação (treino) de um modelo usando uma técnica de aprendizagem (*machine learning*), é necessário avaliar o seu desempenho;
- A medição do desempenho de um modelo é feita com dados não apresentados durante o treino;



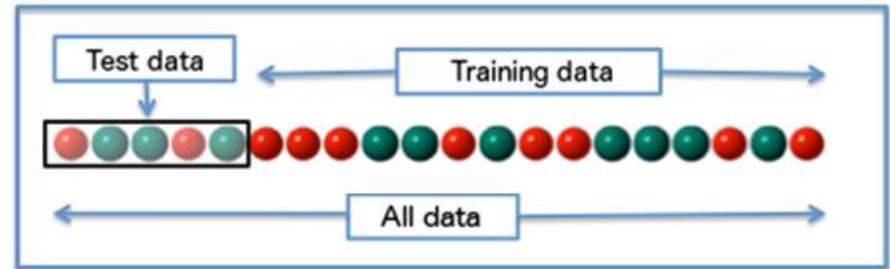
Avaliação de Modelos

- **Dados de treino:**
 - Conjunto de dados usado para ajustar o modelo;
- **Dados de validação:**
 - Conjunto de dados usado para fornecer uma avaliação imparcial de um ajuste do modelo, no conjunto de dados de treino;
- **Dados de teste:**
 - Conjunto de dados usado para fornecer uma avaliação imparcial de um modelo final ajustado ao conjunto de dados de treino.



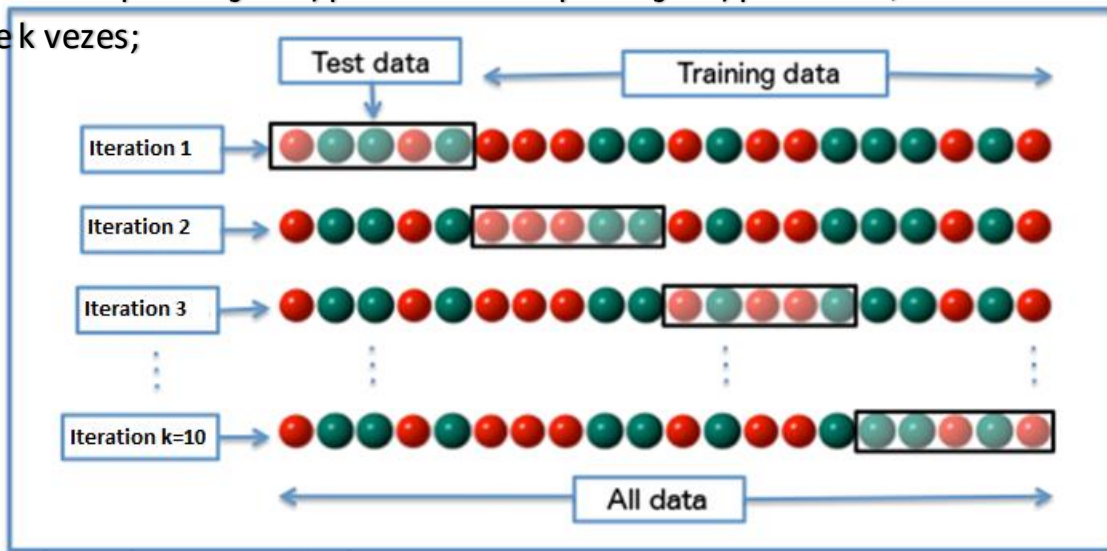
Hold-out Validation

- Método de particionamento de dados;
- Divide o conjunto de dados em dados de treino e dados de teste;



- Separa-se uma parte (*hold-out*) do conjunto de dados para treino/teste (80/20; 75/25; ...)

- Método de validação por cruzamento de dados;
- Consiste em dividir o conjunto de dados em k partes (k folds);
 - A cada iteração, o método utiliza $k-1$ partes ($folds$) para treino e 1 parte ($fold$) para teste;
 - O processo repete-se durante k vezes;



- O erro final é dado pela média dos valores parciais dos erros.

Leave-one-out Cross Validation ***($k=N$)***

- Método de validação por cruzamento de dados;
- Caso particular em que o número de casos N é igual ao número de *folds* k ;



Cross Validation

- Qual o número ideal para k (*folds*)?
- Se o *dataset* for grande, um valor pequeno para k pode ser suficiente, uma vez que teremos uma quantidade grande de dados para treino;
- Se o *dataset* for pequeno, um valor grande de $k \approx N$ pode revelar-se mais adequado para maximizar a quantidade de dados para treino;
- Quanto maior a quantidade de *folds*, melhor a estimativa do erro, mais baixo será o viés(*) (*bias*) e menor será o sobreajuste (*overfitting*);
- De facto, o valor de k depende do valor de N !

(*) viés = distorção
enviesar = entortar



Universidade do Minho
Escola de Engenharia
Departamento de Informática

ADI

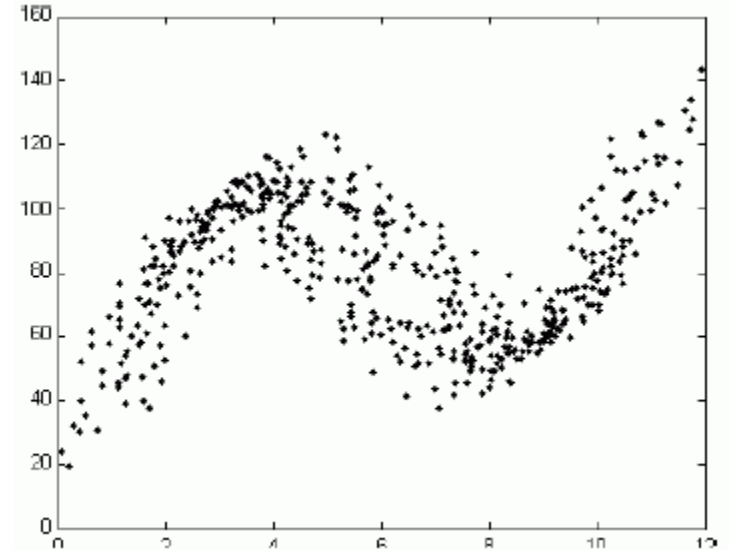
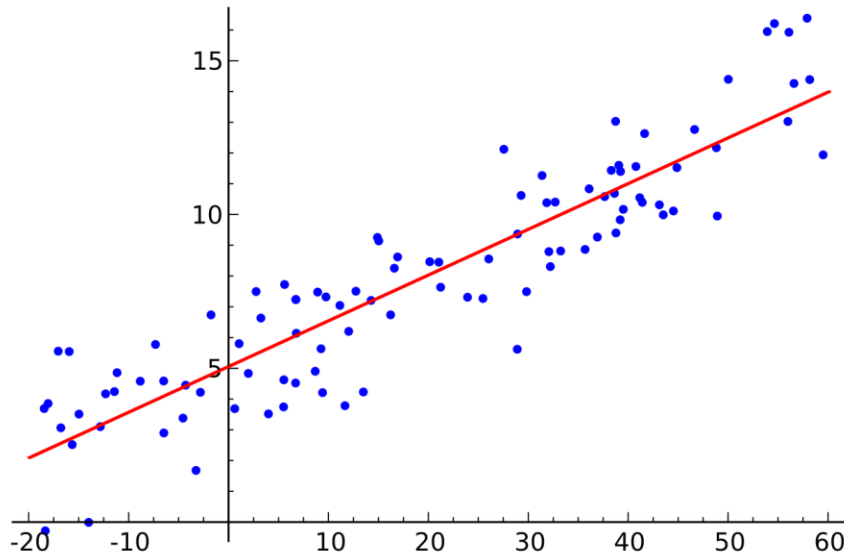
Técnicas de Regressão

Licenciatura em Engenharia Informática, 3º ano
Mestrado integrado em Engenharia Informática, 4º ano



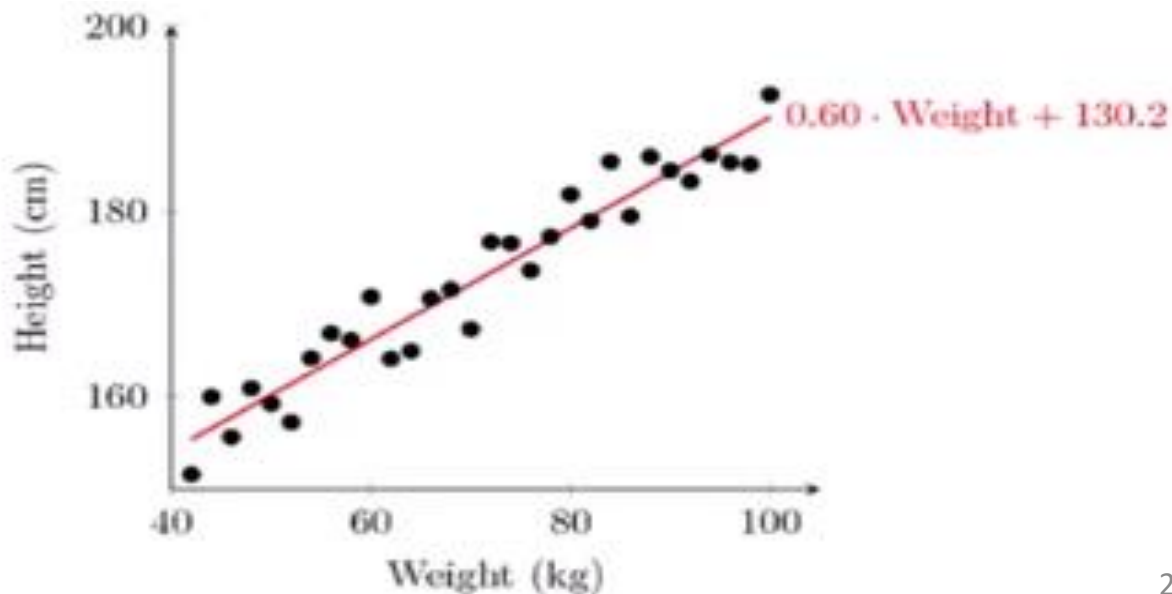
Regressão

- Quão bem uma determinada variável independente prevê outra variável dependente?
- A regressão é um procedimento estatístico que determina a equação para a linha reta que melhor se ajusta a um conjunto específico de dados.



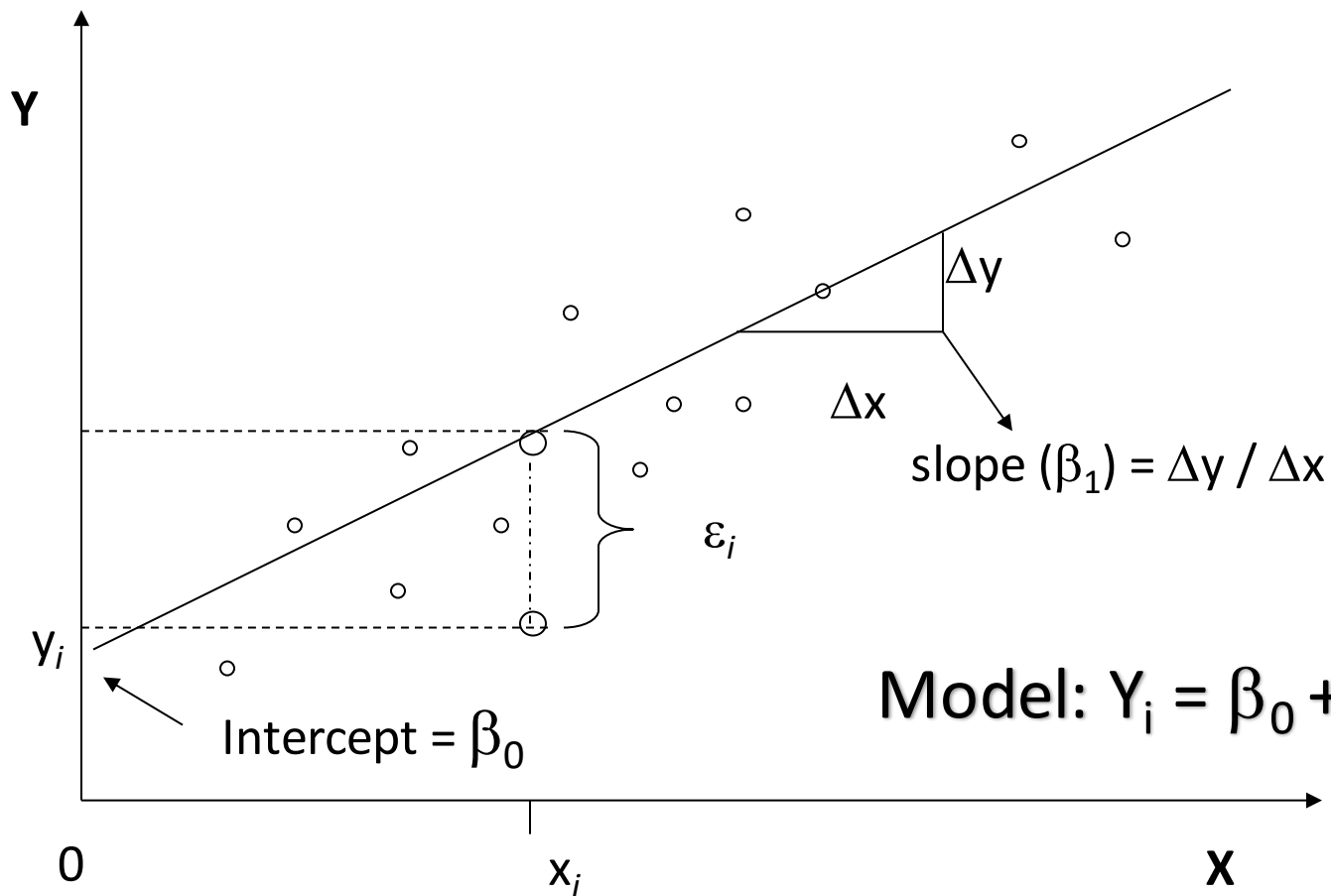
Regressão Linear

- Tem como objetivo prever o valor de um resultado, Y, com base no valor de uma variável de previsão, X;
 - Como “encaixar” uma linha reta num conjunto de dados;
 - Usar esta linha para estimar a resolução de problemas.





Regressão Linear



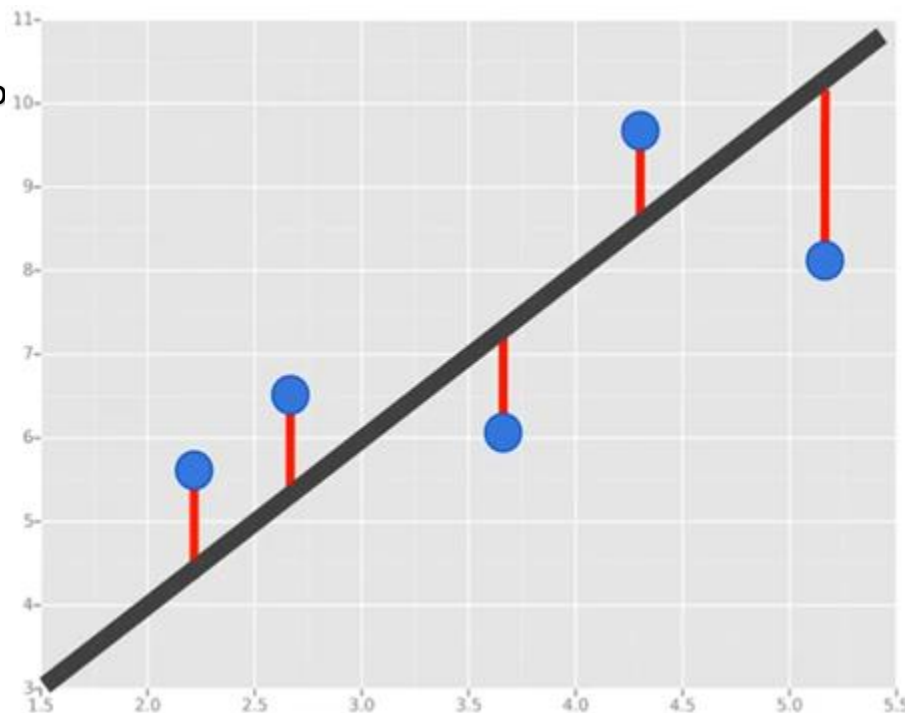
$$\text{Model: } Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Regressão Linear

■ Como funciona?

- O método dos mínimos quadrados minimiza a soma dos erros ao quadrado:
 - y_i : valor verdadeiro
 - $f(x_i, \beta)$: valor previsto / linha ajustada
- O resíduo para uma observação é a diferença entre a observação (valor y) e a linha ajustada:
 - $r_i = y_i - f(x_i, \beta)$
- O método dos mínimos quadrados procura os parâmetros ótimos, minimizando a soma S :
 - $S = \sum_{i=1}^n r_i^2$





ISLab

Synthetic Intelligence Lab

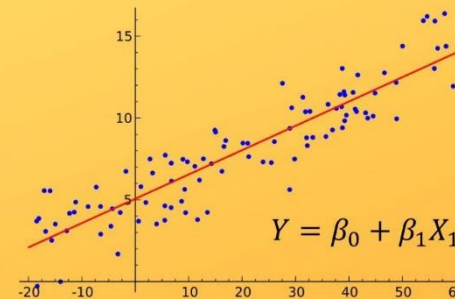
Regressão Linear Múltipla

- A regressão múltipla é usada para determinar o efeito de diversas variáveis independentes, x_1, x_2, x_3, \dots numa variável dependente, y ;
- As diferentes variáveis x_i são combinadas de forma linear e cada uma tem seu próprio coeficiente de regressão:

$$y = a_1 \cdot x_1 + a_2 \cdot x_2 + \dots + a_n \cdot x_n + b + \varepsilon$$

- Os parâmetros a_i refletem a contribuição independente de cada variável independente x_i , para o valor da variável dependente, y .

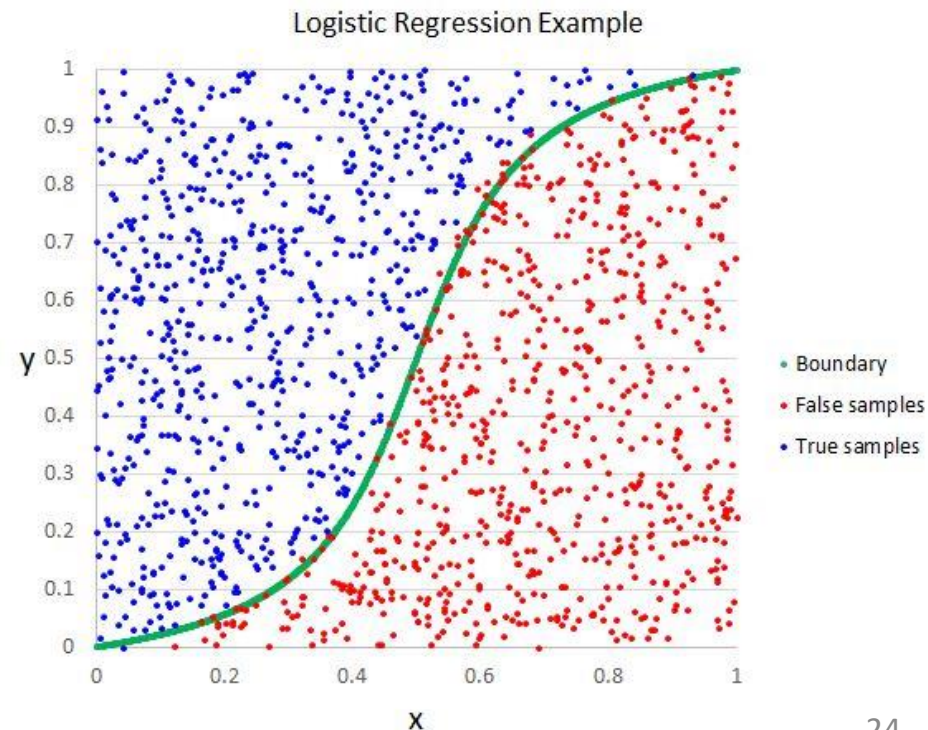
Multiple Linear Regression



number of predictors

Regressão Logística

- A diferença essencial entre regressão linear e **regressão logística** é que esta é usada quando a variável dependente é de natureza binária.
- Em contraste, a **regressão linear** é usada quando a variável dependente é contínua e a natureza da linha de regressão é linear.
- A Regressão Logística é uma técnica de **classificação**:
 - Empréstimo (SIM/NÃO)
 - Diagnóstico (São/Doente)
 - Vinho (Branco/Rosé/Tinto)



Universidade do Minho
Escola de Engenharia
Departamento de Informática

ADI

Métricas de Qualidade

Licenciatura em Engenharia Informática, 3º ano
Mestrado integrado em Engenharia Informática, 4º ano



ISLab

Synthetic Intelligence Lab

Métricas de Qualidade

- Porquê métricas de qualidade?
 - Para avaliar o desempenho do modelo.
- As métricas são usadas para monitorizar e medir o desempenho de um modelo:
 - Erro Médio Absoluto (*Mean Absolute Error* - MAE)
 - Erro Médio Quadrado (*Mean Squared Error* - MSE)
 - Precisão (*Precision*)
 - F1-Score,
 - entre outras...
- No entanto, depende do problema em mãos:
 - É um problema de classificação?
 - De regressão?
 - Séries temporais?



Métricas de Qualidade Modelos de Classificação

- Matrizes de Confusão
 - Tabela utilizada para descrever o desempenho de um modelo de classificação.

| | | True Class | |
|-----------------|----------|------------|----------|
| | | Positive | Negative |
| Predicted Class | Positive | TP | FP |
| | Negative | FN | TN |

Métricas de Qualidade Modelos de Classificação

- Matrizes de Confusão

- Tabela utilizada para descrever o desempenho de um modelo de classificação.

- *Accuracy*

- Quantidade de previsões corretas dividido pela quantidade total de observações:

- $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

| | | True Class | |
|-----------------|----------|------------|----------|
| | | Positive | Negative |
| Predicted Class | Positive | TP | FP |
| | Negative | FN | TN |

■ Matrizes de Confusão

- Tabela utilizada para descrever o desempenho de um modelo de classificação

■ Precisão (*Precision aka Sensitivity*)

- É uma medida da exatidão;
- Determina a proporção de itens relevantes entre todos os itens:

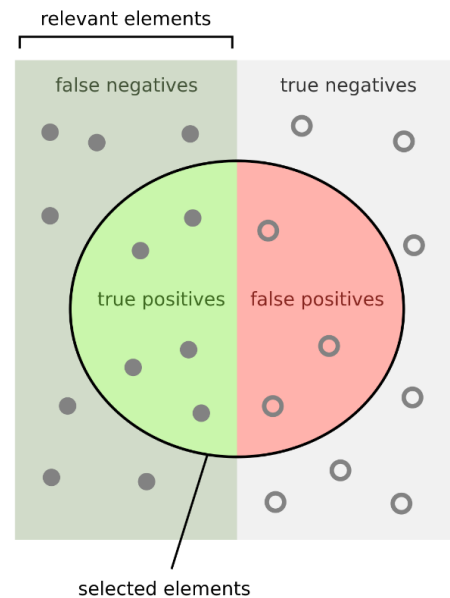
$$\bullet \text{ Precision} = \frac{TP}{TP+FP}$$

■ Recall (*aka Specificity*)

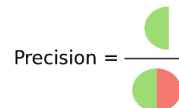
- É uma medida de completude;
- Determina a proporção de itens relevantes obtidos:

$$\bullet \text{ Precision} = \frac{TP}{TP+FN}$$

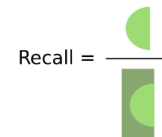
Métricas de Qualidade Modelos de Classificação



How many selected
items are relevant?



How many relevant
items are selected?



Métricas de Qualidade Modelos de Classificação

■ Matrizes de Confusão

- Tabela utilizada para descrever o desempenho de um modelo de classificação.

■ Precisão (*Precision aka Sensitivity*)

- É uma medida da exatidão;
- Determina a proporção de itens relevantes entre todos os itens:

$$\bullet \text{ Precision} = \frac{TP}{TP+FP}$$

■ Recall (*aka Specificity*)

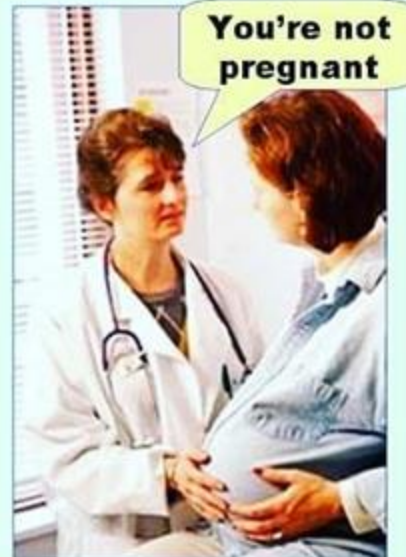
- É uma medida de completude;
- Determina a proporção de itens relevantes obtidos:

$$\bullet \text{ Precision} = \frac{TP}{TP+FN}$$

Type I error
(false positive)



Type II error
(false negative)



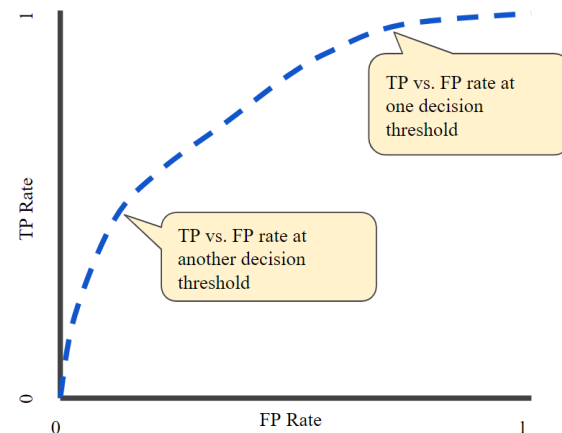
Métricas de Qualidade Modelos de Classificação

- Matrizes de Confusão

- Tabela utilizada para descrever o desempenho de um modelo de classificação.

- ROC curve:

- A curva *Receiver Operating Characteristics* (ROC) encontra o desempenho de um modelo de classificação em diferentes limites de classificação;
 - Reduzindo o patamar (*threshold*) de classificação, são classificados mais itens como positivos, aumentando os falsos positivos e os verdadeiros positivos.



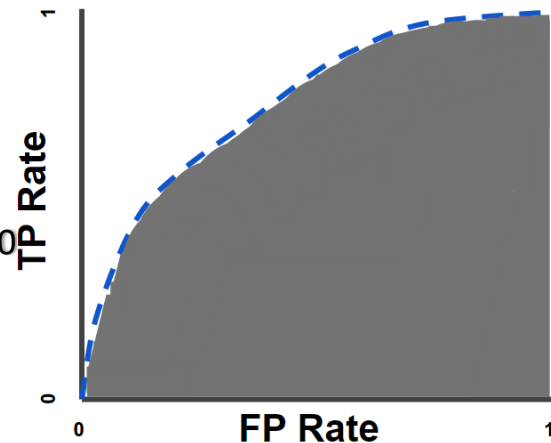
Métricas de Qualidade Modelos de Classificação

- Matrizes de Confusão

- Tabela utilizada para descrever o desempenho de um modelo de classificação.

- AUC curve:

- A Area Under the Curve (AUC) mede a área abaixo da curva ROC;
 - Mede quão bem as previsões são classificadas, em vez de avaliar os seus valores absolutos (varia de 0 a 1);
 - Um modelo cujas previsões estão 100% erradas tem uma AUC de 0, enquanto que aquele cujas previsões estão 100% corretas tem uma AUC de 1.



Métricas de Qualidade Modelos de Regressão

- Erro Médio Absoluto (*Mean Absolute Error* - MAE)
 - Mede a magnitude média dos erros num conjunto de previsões (não considera a direção):
 - $MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$

em que n é a quantidade de observações, y_j e \hat{y}_j são, respetivamente, a observação atual e o valor previsto.

$$MAE = \frac{1}{n} \sum |y - \hat{y}|$$

Diagram illustrating the components of the MAE formula:

- $\frac{1}{n}$: Divide by the total number of data points
- \sum : Sum of
- $|y - \hat{y}|$: The absolute value of the residual
 - y : Actual output value
 - \hat{y} : Predicted output value



Métricas de Qualidade Modelos de Regressão

- Erro Médio Quadrado (*Mean Squared Error* - MSE)
 - Consiste no cálculo da média das diferenças, ao quadrado, entre os erros num conjunto de previsões (não considera a direção):

$$MSE = \frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2$$

em que n é a quantidade de observações, y_j e \hat{y}_j são, respetivamente, a observação atual e o valor previsto.

$$MSE = \frac{1}{n} \sum (y - \hat{y})^2$$



Métricas de Qualidade Modelos de Regressão

- Raiz Quadrada do Erro Médio Quadrado (*Root Mean Squared Error* - RMSE)
 - Consiste no cálculo da média das diferenças, ao quadrado, entre os erros num conjunto de previsões (não considera a direção):

- $RMSE = \frac{1}{n} \sqrt{\sum_{j=1}^n (y_j - \hat{y}_j)^2}$

em que n é a quantidade de observações, y_j e \hat{y}_j são, respetivamente, a observação atual e o valor previsto.

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

Métricas de Qualidade Modelos de Regressão

$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

$$MSE = \frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2$$

$$RMSE = \frac{1}{n} \sqrt{\sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

- Três das métricas mais comuns usadas para medir a precisão de variáveis contínuas;
- Todas expressam o erro médio de previsão do modelo (valores mais baixos são melhores);
- Todos variam de 0 a ∞ e são indiferentes à direção dos erros;
- MAE e RMSE expressam o erro de previsão na mesma unidade da variável de interesse;
- MSE e RMSE, ao elevar o erro ao quadrado, dão um peso relativamente alto para erros grandes;
- MSE e RMSE são mais úteis quando grandes erros são especialmente indesejáveis.

Métricas de Qualidade Modelos de Regressão

$$MAE = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

$$MSE = \frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2$$

$$RMSE = \frac{1}{n} \sqrt{\sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

| # | Error | Error | Error ² |
|---|-------|-------|--------------------|
| 1 | 1 | 1 | 1 |
| 2 | -1 | 1 | 1 |
| 3 | 3 | 3 | 9 |
| 4 | 3 | 3 | 9 |

| MAE | MSE | RMS E |
|-----|-----|----------|
| 2 | 5 | 2.24 |

| # | Error | Error | Error ² |
|---|-------|-------|--------------------|
| 1 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 |
| 4 | 10 | 10 | 100 |

| MAE | MSE | RMS E |
|-----|-----|----------|
| 2.5 | 25 | 5 |

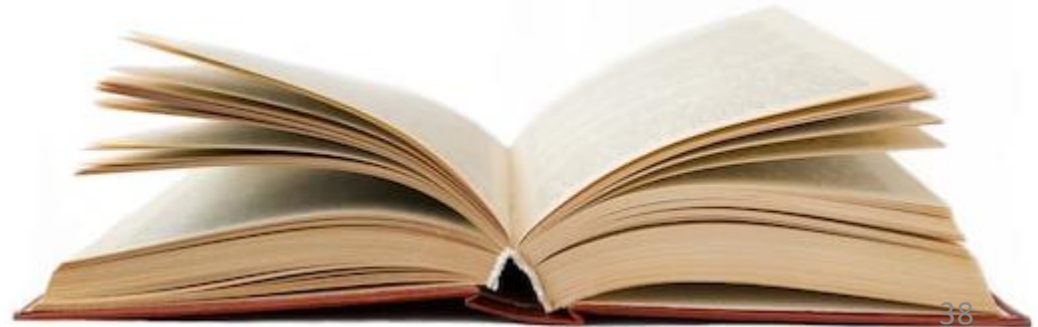


ISLab

Synthetic Intelligence Lab

Referências bibliográficas

- Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. John Wiley & Sons, 2021
- Ranganathan, Priya, C. S. Pramesh, and Rakesh Aggarwal. "Common pitfalls in statistical analysis: logistic regression." Perspectives in clinical research 8.3, 2017
- Breiman, Leo; Friedman, J. H.; Olshen, R. A.; Stone, C. J. (1984), "Classification and regression trees", Monterey, CA
- Ross Quinlan (1993), "C4.5 Programs for Machine Learning", Morgan Kaufmann



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Aprendizagem e Decisão Inteligentes

Licenciatura em Engenharia Informática, 3º ano
Mestrado integrado em Engenharia Informática, 4º ano