



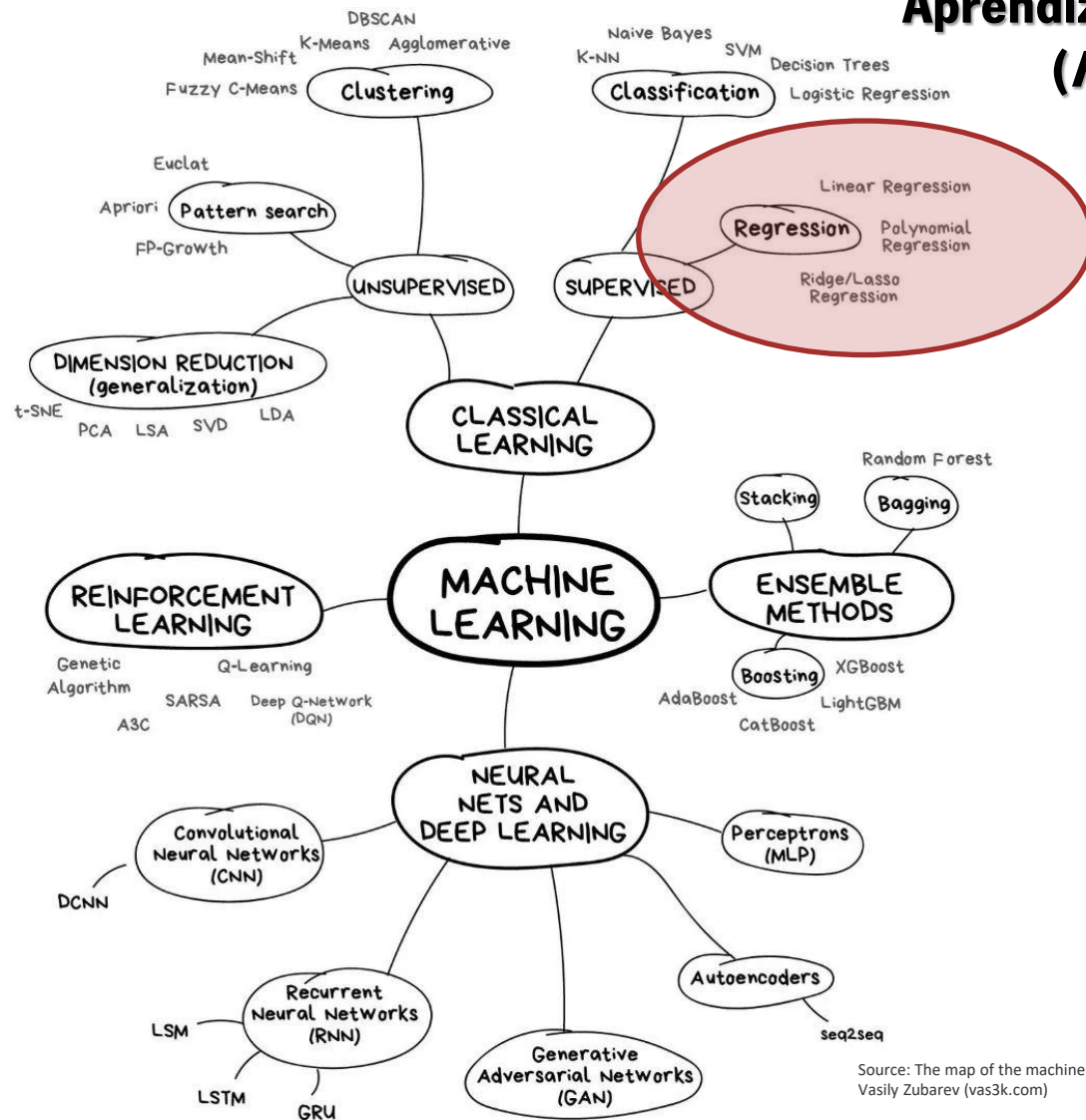
Universidade do Minho
Departamento de Informática

Técnicas de Regressão

ADI³ - LEI/MiEI @ 2024/2025, 2º sem



Aprendizagem Automática (Machine Learning)



Source: The map of the machine learning world
Vasily Zubarev (vas3k.com)



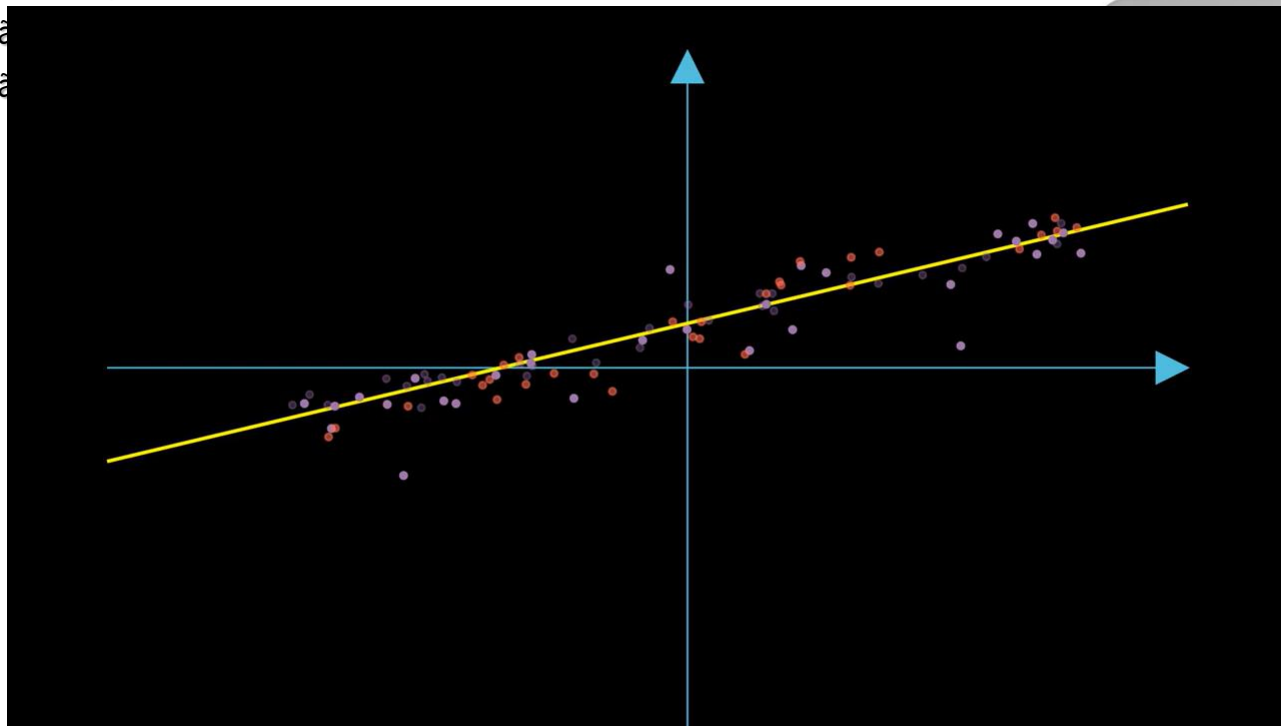
- Técnicas de Regressão
 - Regressão Linear
 - Regressão Múltipla
 - Regressão Polinomial
 - Regressão Logística





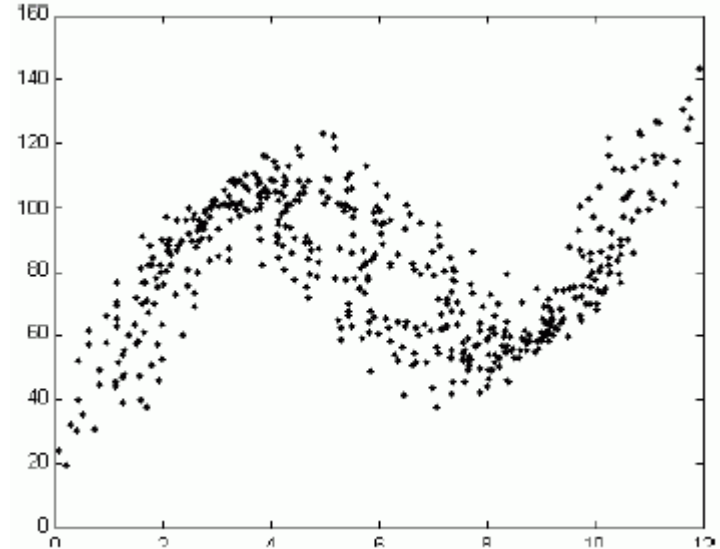
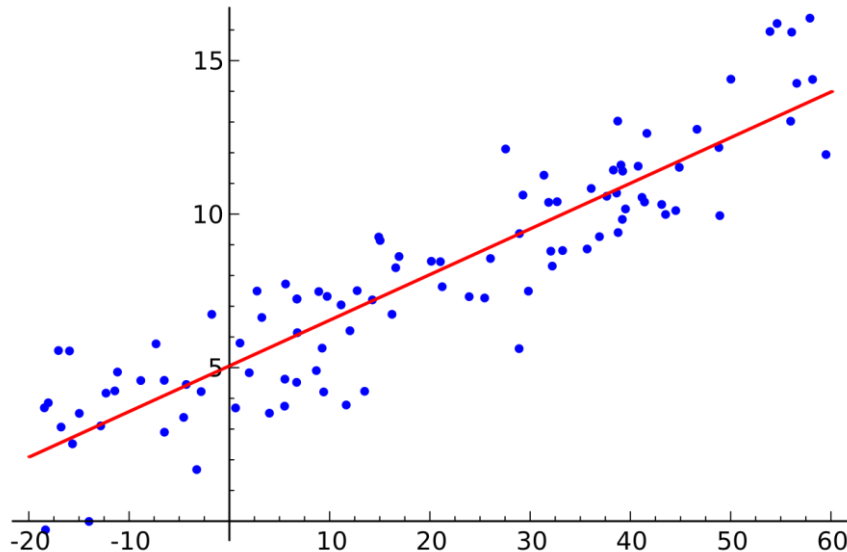
■ Técnicas de Regressão

- Regressão Linear
- Regressão Múltipla
- Regressão
- Regressão



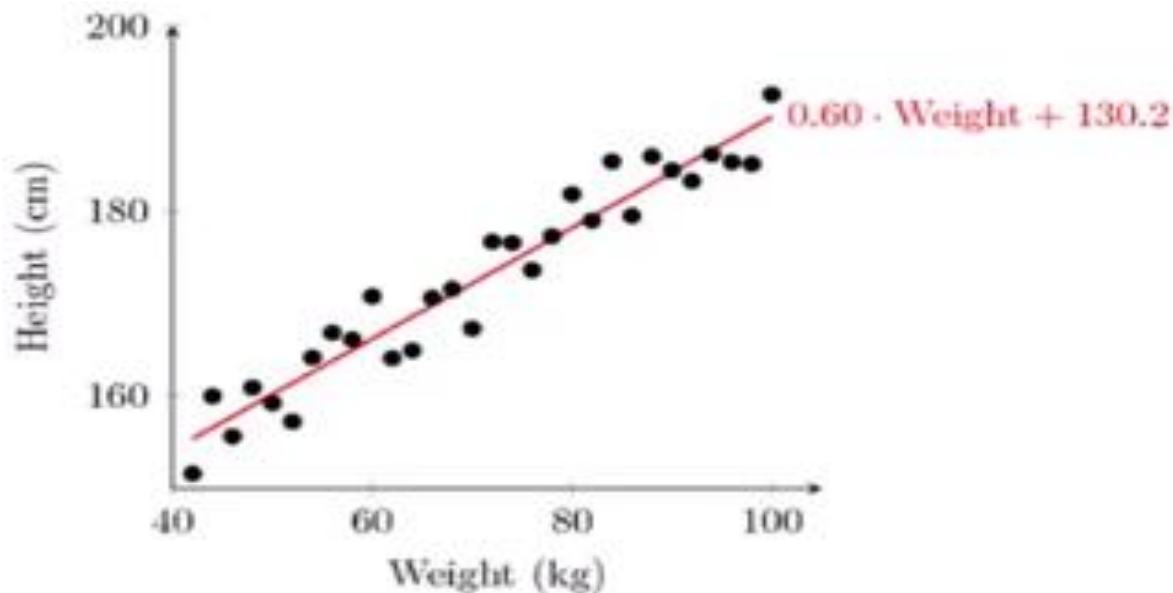


- Quão bem uma determinada variável independente prevê outra variável dependente?
- A regressão é um procedimento estatístico que determina a equação para a reta/curva que melhor se ajusta a um conjunto específico de dados.





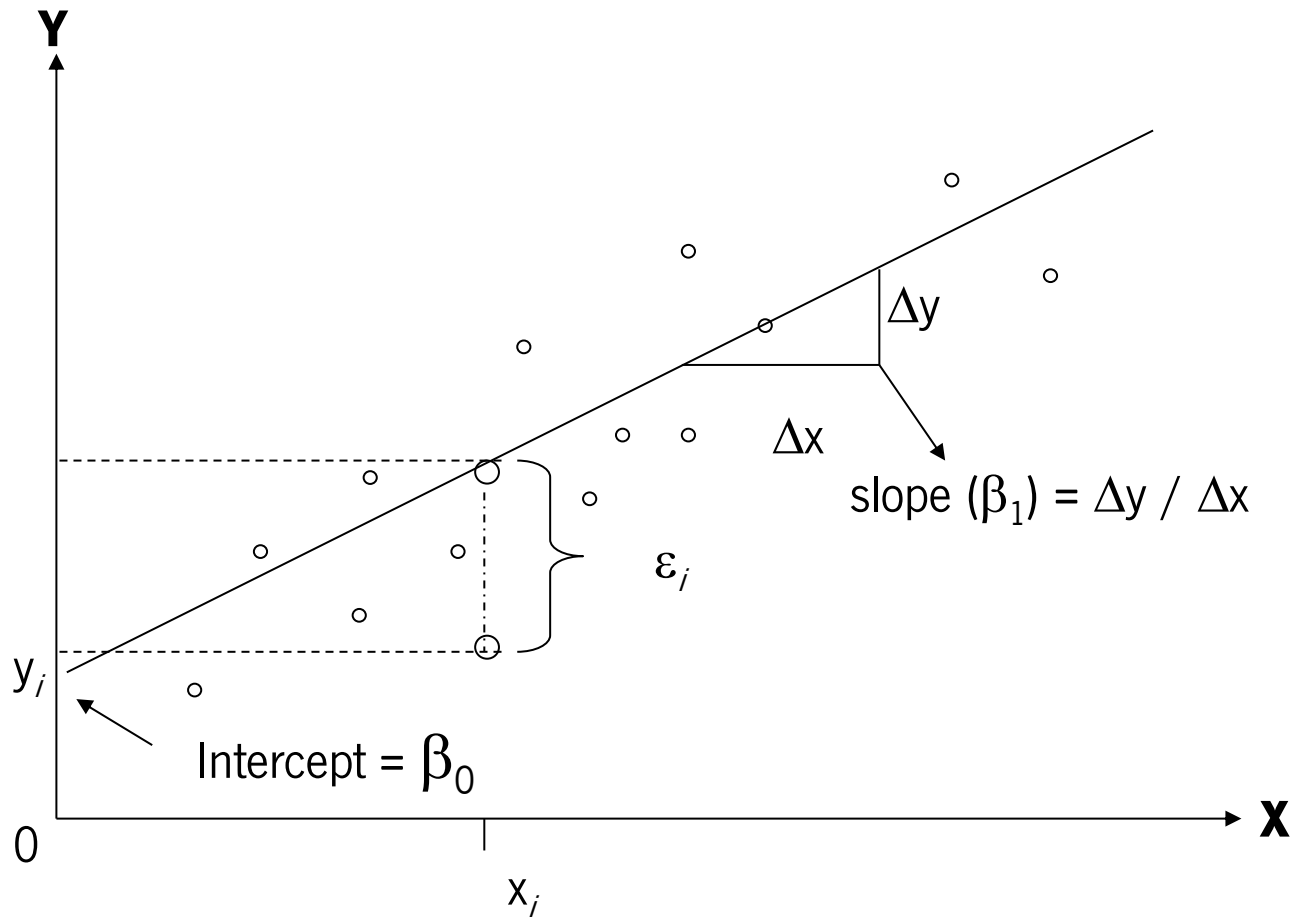
- Tem como objetivo prever o valor de um resultado, Y , com base no valor de uma variável de previsão, X ;
 - Como “encaixar” uma linha reta num conjunto de dados?
 - Como usar esta linha para estimar a resolução de problemas?





- Usando a equação de reta:

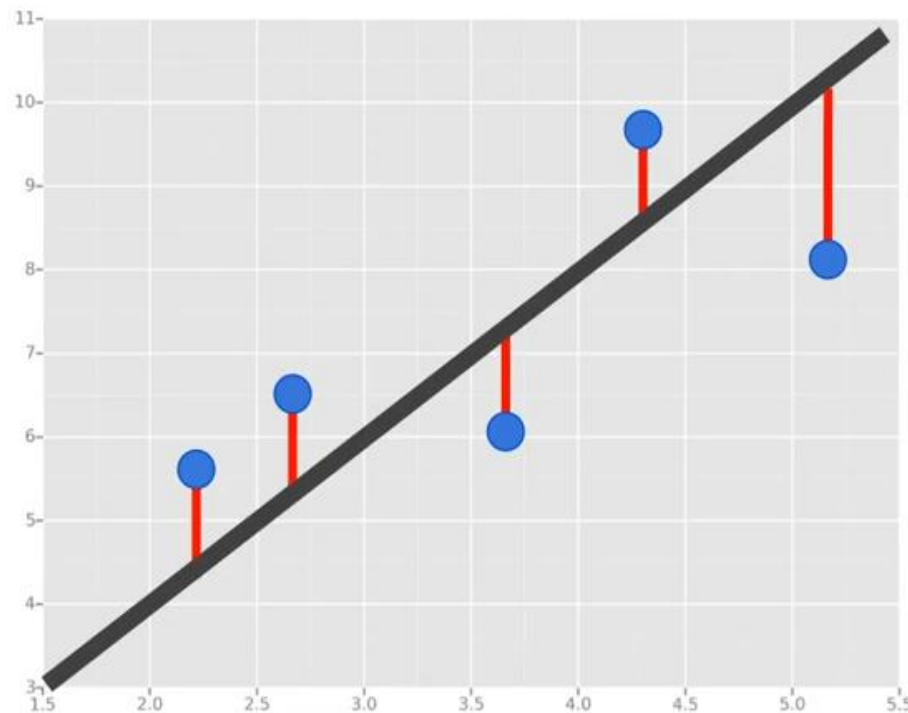
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$





■ 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 :
 - $S = \sum_{i=1}^n r_i^2$





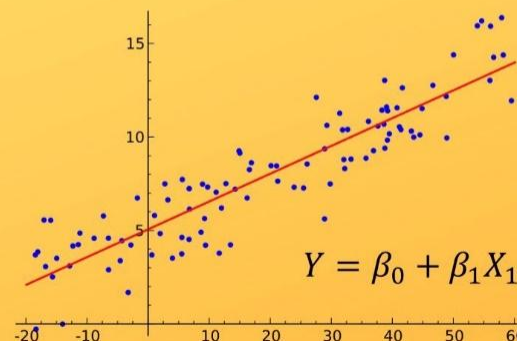
Regressão Linear Múltipla

- 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



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

number of predictors

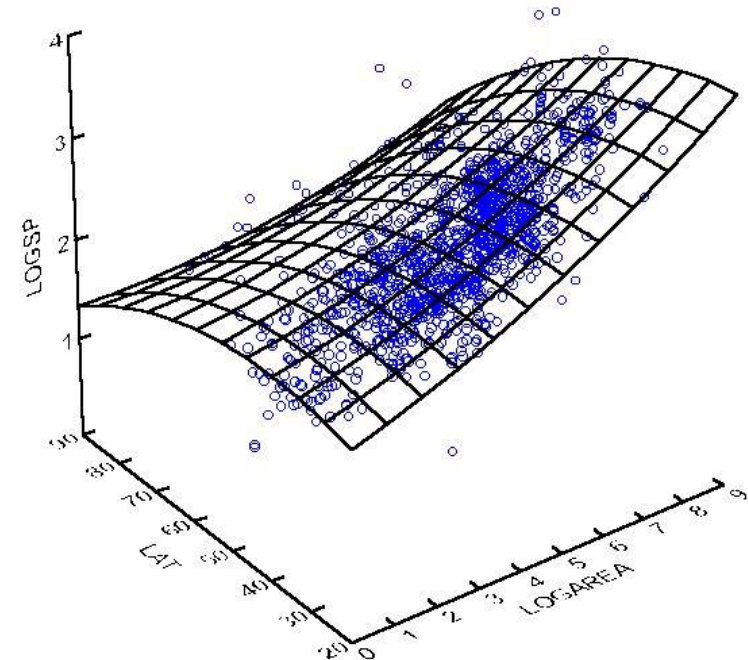


Regressão Polinomial Múltipla

- A regressão polinomial 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_1 e x_2 são combinadas de forma polinomial e cada uma tem seu coeficiente de regressão:

$$y = a_1 \cdot x_1 + a_2 \cdot x_1^2 + a_3 \cdot x_2 + a_4 \cdot x_2^2 + a_5 \cdot x_1 \cdot x_2 + b + \varepsilon$$

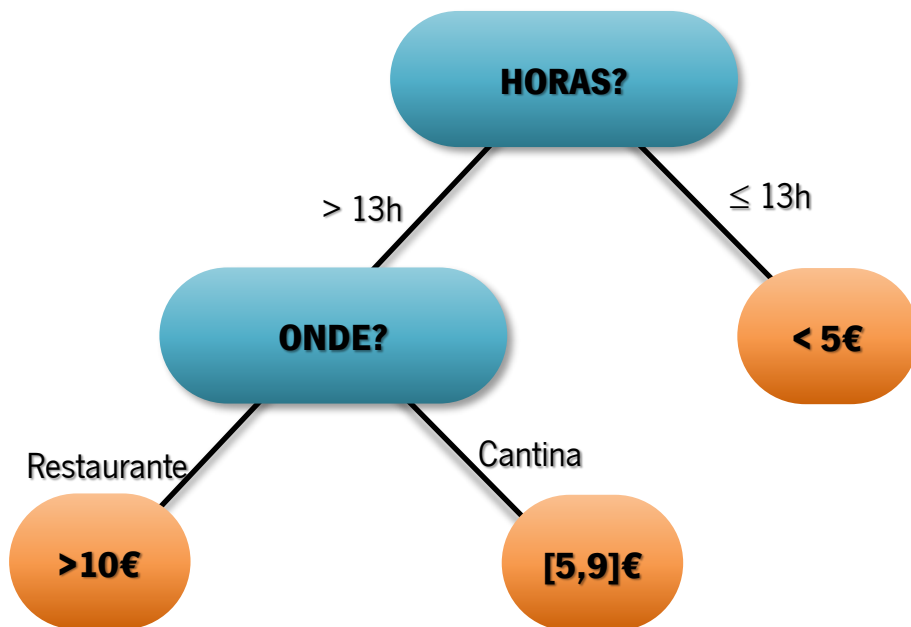
- Os parâmetros a_i refletem a contribuição independente de cada variável independente x_1 e x_2 , para o valor da variável dependente, y .
- A regressão polinomial é um caso particular da regressão linear.





Regressão Árvores de Decisão

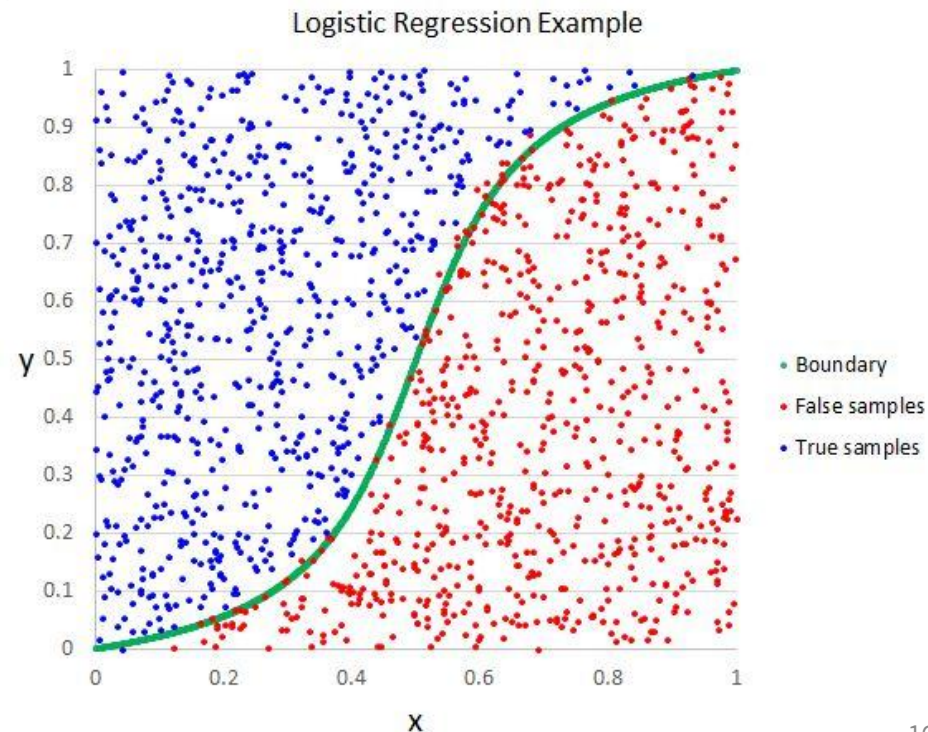
- Uma Árvore de Decisão é um grafo hierarquizado (árvore!) em que:
 - Cada nodo interno testa um atributo do *dataset*;
 - Cada ramo identifica um valor (ou conjunto de valores) do nodo testado;
 - Cada folha representa uma decisão;



HORAS	ONDE	CUSTO
12h30	Cantina	< 5€
13h15	Cantina	> 5€ , < 9€
13h10	Restaurante	> 10€
11h00	Restaurante	< 5€
13:30	Cantina	> 5€ , < 9€

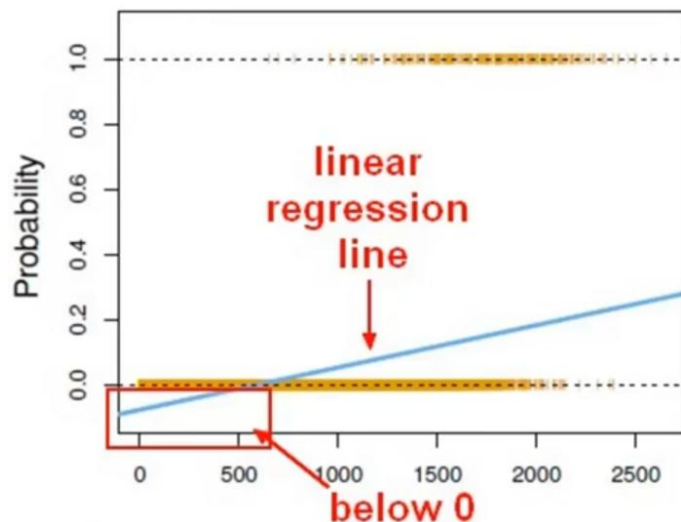


- A diferença essencial entre regressão (linear, múltipla, ...) e **regressão logística** é que esta é usada quando a variável dependente é de natureza discreta.
- Em contraste, a **regressão linear** (múltipla, polinomial, ...) é usada quando a variável dependente é **contínua** e a natureza da linha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



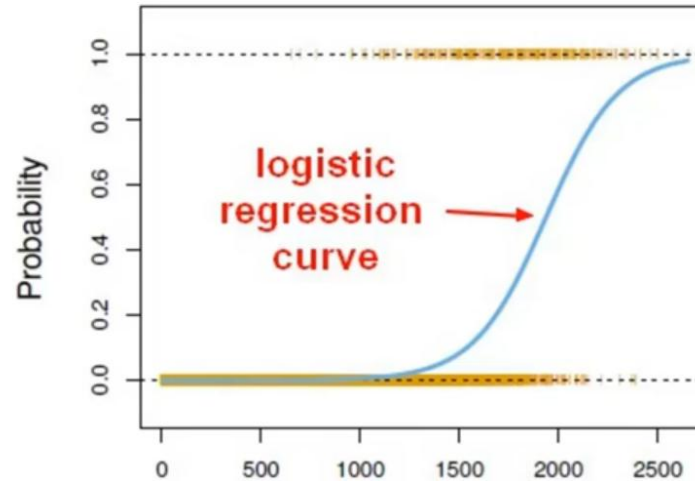
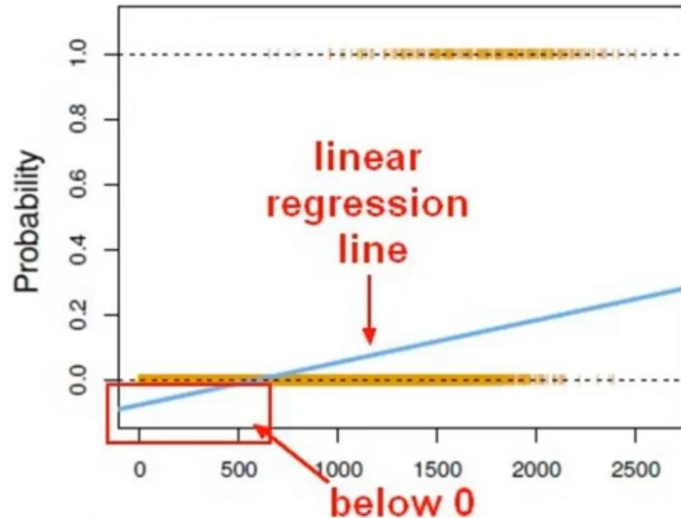


- **Técnicas de regressão** são usadas, normalmente, para prever uma variável dependente contínua;
- Apesar de a designação poder originar alguma confusão, a **regressão logística** permite resolver **problemas de classificação**, em que se estimam categorias (valores discretos);
- Usar uma função linear de regressão não produz bons resultados na previsão de uma variável binária:



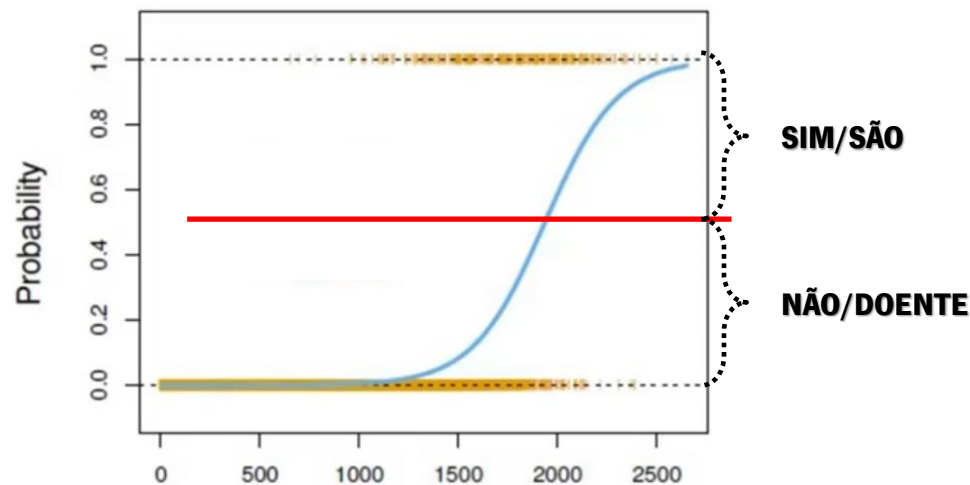
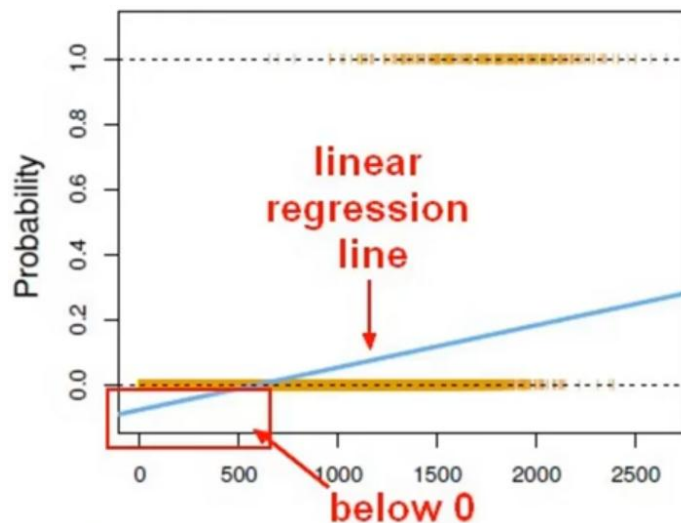


- **Técnicas de regressão** são usadas, normalmente, para prever uma variável dependente contínua;
- Apesar de a designação poder originar alguma confusão, a **regressão logística** permite resolver **problemas de classificação**, em que se estimam categorias (valores discretos);
- Substitui-se uma linha de regressão por uma curva de regressão logística:





- **Técnicas de regressão** são usadas, normalmente, para prever uma variável dependente contínua;
- Apesar de a designação poder originar alguma confusão, a **regressão logística** permite resolver **problemas de classificação**, em que se estimam categorias (valores discretos);
- Para conseguir uma previsão binária, usa-se um patamar de definição (*threshold*):





Regressão Logística - Classificação

Árvores de Decisão

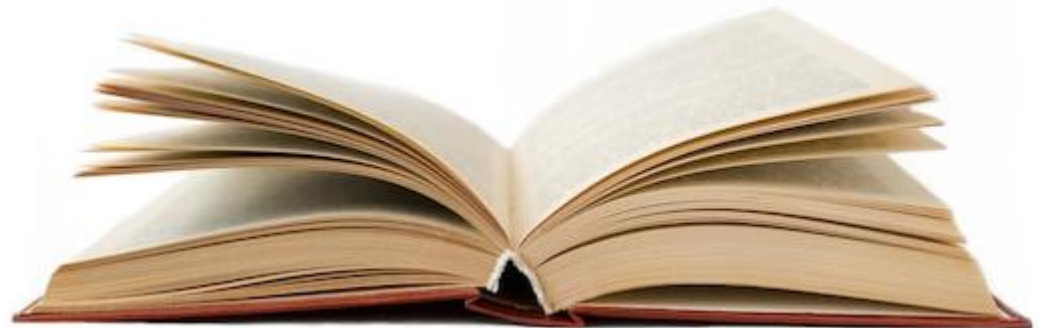
- Uma Árvore de Decisão pode ser utilizada para fazer **classificação**:
 - Decidir sobre se ou onde almoçar: classificação binária (SIM/NÃO)
 - Prever quem sobreviveu ao acidente do Titanic: classificação binária (SIM/NÃO)
 - Classificar um conjunto de imagens: classificação múltipla (laranja, kiwi,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





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





Universidade do Minho
Departamento de Informática

Técnicas de Regressão

ADI³ - LEI/MiEI @ 2024/2025, 2º sem