

## Aula 5

# Regressão

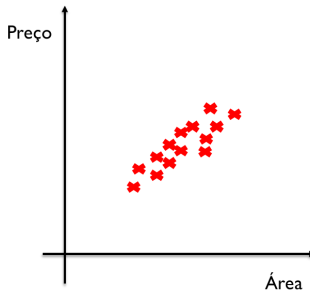
# Introdução

- A um modelo de regressão consiste um modelo, o qual foi gerado com base nos valores dos atributos dos exemplos, de uma de forma que esse modelo possa **predizer um valor numérico contínuo**
- Para isso pode-se utilizar
  - Métodos estatísticos
  - Métodos de aprendizado de máquina

# Estrutura dos Modelos de Regressão

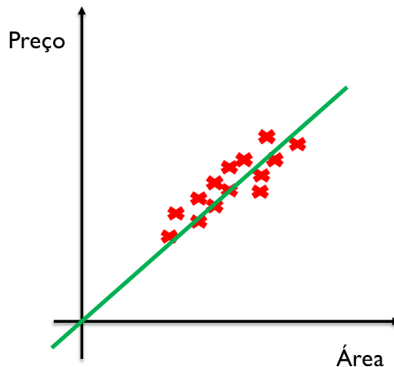
- O propósito dos modelos de regressão é identificar uma relacional funcional entre o **atributo alvo** (atributo cujo valor deseja-se prever no futuro) e os demais atributos que caracterizam os exemplos

Tamanho da casa (em m <sup>2</sup> )	Preço (em R\$ x 1000)
60	230
85	320
44	120
...	...
...	...
380	850



# Estrutura dos Modelos de Regressão

Tamanho da casa (em m <sup>2</sup> )	Preço (em R\$ x 1000)
60	230
85	320
44	120
...	...
...	...
380	850



# Estrutura dos Modelos de Regressão

- Sendo assim, os modelos de regressão podem ser utilizados para dois objetivos:
  - 1 **Ressaltar e interpretar a dependência entre o atributo alvo e os demais atributos**
  - 2 **Prever valores futuros relacionados ao atributo alvo**  
(considerando o modelo prévio que identificou a relação funcional entre os demais atributos e o atributo alvo)

## Exemplos de Uso

- **Qual o custo de produção para produzir um determinado número de itens**
  - Base histórica de custos de produção quando produzidos um determinado número de itens
    - 1 Descobrir a relação entre número de produtos produzidos e o custo de produção
    - 2 Prever qual será o custo de produção caso deseje aumentar o número de produtos produzidos

## Exemplos de Uso

- **Interpretação do impacto das vendas de produtos frente a investimentos em propagandas em diferentes mídias**
  - Base história com valores gastos em cada tipo de media (jornal, revistas, TV, rádio, ...) e o impacto na venda de produtos
    - ① Descobrir qual atributo tem maior relação com a variável alvo
    - ② Prever qual será o impacto nas vendas dado uma especificação de investimento nas diferentes medias

## Definição Formal


- Seja um conjunto de dados  $D$  composto de  $m$  observações e  $n + 1$  atributos
  - $n$  atributos explanatórios (atributos/variáveis independentes, atributos preditores, ...)
  - 1 atributo alvo (variável dependente, resposta, saída desejada, ...)
- Os valores das variáveis independentes do  $i$ -ésimo exemplo será denotado por  $\mathbf{x}_i$  ( $\mathbf{x}_i \in \mathbb{R}^n$ )
- O valor do atributo alvo de  $\mathbf{x}_i$  é denotado por  $y_i$



## Definição Formal

Regressão univariada (número de atributos preditores = 1)

Tamanho da casa (em m <sup>2</sup> )	Preço (em R\$ x 1000)
60	230
85	320
44	120
...	...
...	...
380	850



## Definição Formal

Regressão multivariada (número de atributos preditores  $> 1$ )

Gasto TV (milhares R\$)	Gasto Rádio (milhares R\$)	Aumento Vendas (%)
100	10	10%
150	15	16%
50	20	7%
300	300	40%
...	...	...
200	50	23%

$x$   $y$

## Definição Formal

- Os vetores das  $m$  observações de um conjunto  $\mathcal{D}$  serão denotados por  $\mathbf{X}$  ( $m \times n$ )
- Os valores associados à essas  $m$  observações serão denotados pelo vetor  $\mathbf{y} = (y_1, y_2, \dots, y_m)$
- Por fim:
  - Seja  $Y$  uma variável aleatória que representa os valores dos atributos alvo
  - Seja  $X_j$  uma variável aleatória associada com o valor do  $j$ -ésimo atributo

## Definição Formal

- Formalmente:
  - Um modelo de regressão visa encontrar uma função

$$f : \mathbb{R}^n \rightarrow \mathbb{R}$$

que expressa o relacionamento entre a variável dependente  $Y$  e as  $n$  variáveis explanatórias  $X_j$

$$Y = f(X_1, X_2, \dots, X_n)$$

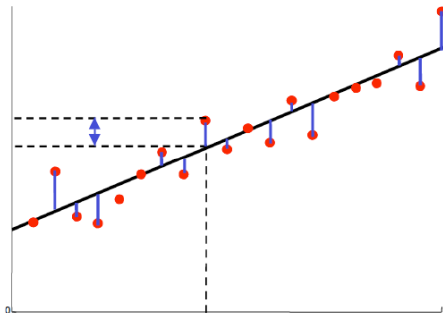
# Definição Formal

- Em geral, o processo de identificar a função  $f$  (também chamada de hipótese), pode ser dividida em duas fases

- Escolher qual classe de algoritmos utilizar**
- Escolher um modelo com maior poder de predição** (maior capacidade de generalização)

## Definição Formal

- Além disso, a grande maioria dos modelos de regressão irão tentar obter seu modelos de forma a minimizar o erro residual → diferença entre um valor predito por um modelo e um valor real



- Os modelos são obtidos obtendo os valores dos pesos ( $w$ ) que ponderam os atributos
- Os valores de  $w$  informarão o quão a sua variação afetará a variável  $Y$
- Portanto, o valor de  $w$  dará sua importância para prever a variável de saída → quanto maior, mais importante é o respectivo atributo

# Regressão Linear

- Modelos de regressão linear representam a família mais conhecida (e simples) de modelos de regressão
- Os possíveis tipos de hipóteses consistem em funções lineares

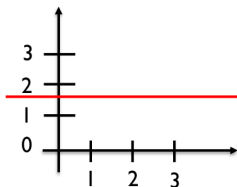
$$Y = w_1X_1 + w_2X_2 + \cdots + w_nX_n + b = \sum_{j=1}^n w_jX_j + b$$

- Se houver uma única variável ( $n = 1$ ), o modelo de regressão linear vira a equação da reta

$$Y = wX + b$$

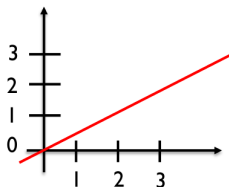


# Regressão Linear



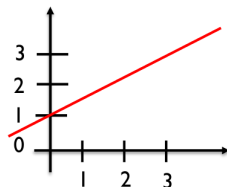
$$w = 0$$

$$b = 1,5$$



$$w = 0,5$$

$$b = 0$$

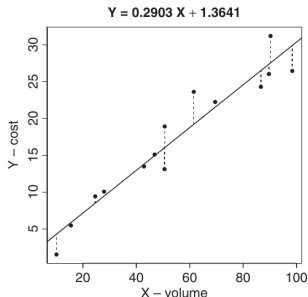


$$w = 0,5$$

$$b = 1$$

## Regressão Linear Univariada

- Para modelos de regressão linear, vamos considerar o modelo de regressão linear simples (um único atributos explanatório e um atributo alvo)
- Para isso, o conjunto de dados é reduzido a  $m$  pares de valores  $(x_i, y_i)$ ,  $1 \leq i \leq m$



# Correlação

- É possível obter os a solução (valores dos coeficientes) de forma analítica por meio das seguintes igualdades

$$w = \frac{\sigma_{xy}}{\sigma_{xx}}$$

$$b = \bar{\mu}_y - w\bar{\mu}_x$$

## Correlação

na qual

$$\bar{\mu}_x = \frac{\sum_{i=1}^m x_i}{m}, \quad \bar{\mu}_y = \frac{\sum_{i=1}^m y_i}{m}$$

representam a média do atributo explanatória e a média do atributo alvo respectivamente, e

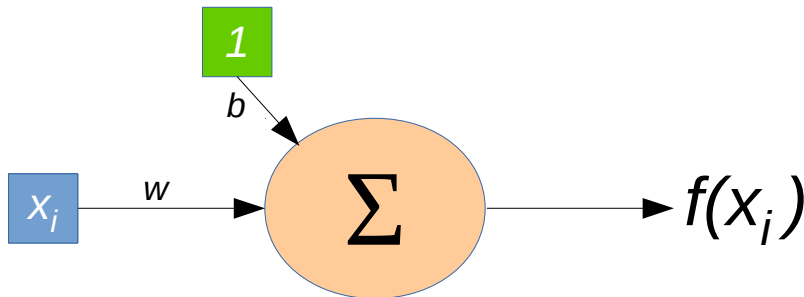
$$\sigma_{xy} = \sum_{i=1}^m (x_i - \bar{\mu}_x)(y_i - \bar{\mu}_y)$$

$$\sigma_{xx} = \sum_{i=1}^m (x_i - \bar{\mu}_x)^2, \text{ e } \sigma_{yy} = \sum_{i=1}^m (y_i - \bar{\mu}_y)^2$$

expressam a covariância de  $x$  e  $y$ , a variância de  $x$  e a variância de  $y$  respectivamente (**já considerando as simplificações matemáticas**)



# Perceptron



- Também irá obter os valores dos parâmetros  $w$  e  $b$  só que por meio de aprendizado supervisionado

# Perceptron

- **Treinamento**

$$w^{t+1} = w_t + \eta(y_i - f(x_i)) * x_i$$

$$b^{t+1} = b^t + \eta(y_i - f(x_i))$$

- **Realizar o treinamento até**

- Um número fixo de épocas
- Até que os valores dos parâmetros não se alterem
- Até que seja atingido um erro quadrático médio abaixo de um limiar

## Exercício

Calcule os modelos de regressão que utilizando o método da Correlação e Perceptron (com  $n^{\circ}$  máximo de épocas = 2,  $\eta = 0.1$ ,  $w^0 = 0$  e  $b = 0$  para o seguinte conjunto de dados

x	y
1	3
2	5
3	7
4	9
5	11

# Regressão Linear Multivariada

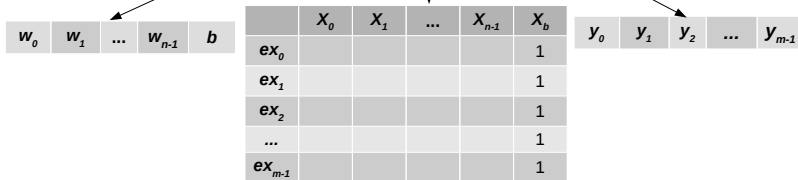
- No caso da regressão multivariada, há mais de um atributo descritivo para um determinado exemplo
- Nesse caso, tem-se que se aprender o peso associado a cada atributo
- O resto é igual à regressão linear simples



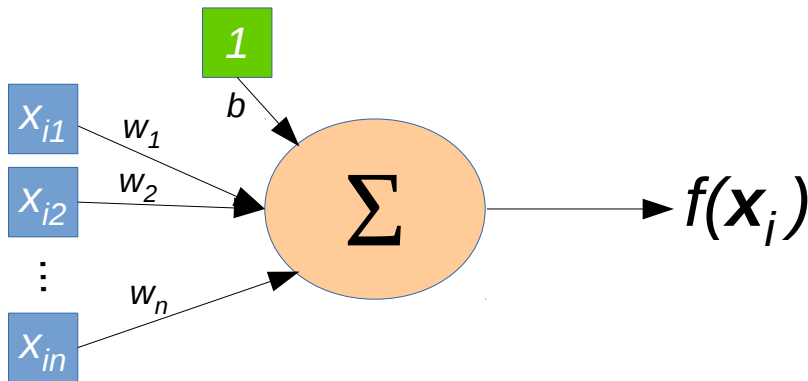
# Método Matricial

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$



# Perceptron



## Exercício

Considere os dados apresentados abaixo e todos os pesos iniciais de um perceptron iguais a 0. Considere também um  $\eta = 0,01$ . Faça 3 iterações completas e apresente os pesos obtidos.

Nota Disciplina 1	Nota Disciplina 2	Nota Disciplina 3	Nota Disciplina 4	Nota IA
8,5	9,0	10	10	8,0
9,0	2,0	5,5	6,0	7,5
3,5	5,0	6,0	7,0	5,0
4,0	2,0	2,5	7,5	6,6

# Validação de Modelos de Regressão

- Dentre as várias possibilidades para avaliar os resultados de um modelo de regressão (inclusive os vistos para validação de modelos de previsão em séries temporais), vamos o
  - Soma quadrática das diferenças (*sum of squared differences*)
  - Coeficiente de determinação (*coefficient of determination*)
  - Soma do erro quadrático (*sum of squared errors*)

## Soma quadrática das diferenças

- A primeira soma quadrática das diferenças

$$SSM = \sum_{i=1}^m (f(x) - \bar{\mu}_y)^2$$

- Representa o erro de predição quando a média do atributo  $y$  é utilizado como uma estimativa da resposta

## Soma quadrática das diferenças

- A segunda soma quadrática das diferenças

$$SST = \sum_{i=1}^m (y_i - \bar{\mu}_y)^2$$

- Representa a quantidade de erro no modelo de regressão

## Coeficiente de determinação

- O coeficiente de determinação  $R^2$  expressa a proporção das duas somas quadráticas das diferenças apresentadas anteriormente

$$R^2 = \frac{SSM}{SST} = \frac{\sum_{i=1}^m (f(x_i) - \bar{\mu}_y)^2}{\sum_{i=1}^m (y_i - \bar{\mu}_y)^2}$$

- Se o resultado é próximo de 1, pode-se concluir que a variabilidade explicada no atributo alvo pode ser explicada pelo modelo de regressão

## Soma dos erros quadráticos

- O bom e velho erro quadrático

$$SSE = \sum_{i=1}^m (f(x_i) - y_i)^2$$

- Nessa medida, quanto mais próximo de 0, melhor
- **OBSERVAÇÃO:** as demais medidas de avaliação de preditor de séries temporais também podem ser utilizados aqui



## Esquema de Avaliação

- Como podemos estimar quão bem um regressor irá se comportar se comportar frente a dados novos (ou não vistos)?
- Se utilizarmos o próprio conjunto de dados que foi utilizado para construir o regressor (também conhecido por **conjunto de treinamento**) para avaliá-lo, a estimativa tende a ser otimista
- Portanto, para avaliar/estimar a performance de um regressor, deve-se utilizar um **conjunto de teste**
  - Exemplos que não pertencem ao conjunto de treinamento
  - Os exemplos de teste também devem possuir os valores do atributo alvo

# Holdout

- O conjunto de exemplos é particionado aleatoriamente em dois conjuntos independentes: **treinamento** e **teste**
- Tipicamente  $2/3$  são utilizados para treinamento e o restante para teste
- Uma variação do método *Holdout* é o *Random Subsampling*, na qual o método *Holdout* é executado  $k$  vezes  
O resultado final é a média dos resultados de cada iteração

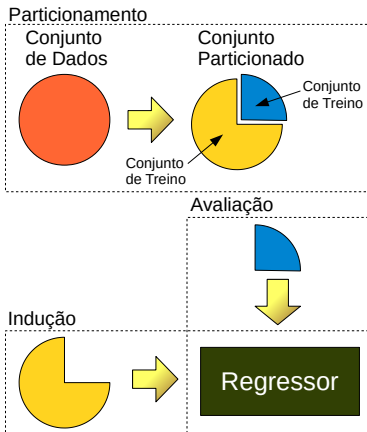


Figura: Holdout

## $k$ -Fold Cross-Validation

- Os dados são aleatoriamente particionados em  $k$  subconjuntos mutuamente exclusivos conhecidos como “*folds*”  
( $D_1, D_2, \dots, D_k$ )
- Os *folds* são de tamanhos iguais
- O procedimento de treino e teste é executado  $k$  vezes
- A cada iteração  $i$ ,  $D_i$  é utilizado como conjunto de teste, e os *folds* restantes são utilizados para treinamento
- Diferente dos métodos *Holdout* e *Random Subsampling*, cada exemplo é usado o mesmo número de vezes para treinamento e uma única vez para teste
- O resultado da regressão é a média dos resultados das iterações

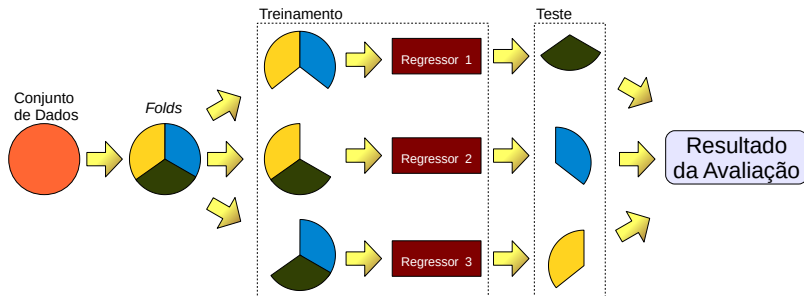


Figura: 3-fold cross-validation

- Normalmente é utilizado *10-fold cross-validation*

# Material Complementar

- Simple linear regression

[https://en.wikipedia.org/wiki/Simple\\_linear\\_regression](https://en.wikipedia.org/wiki/Simple_linear_regression)

- Isotonic Regression

[https://en.wikipedia.org/wiki/Isotonic\\_regression](https://en.wikipedia.org/wiki/Isotonic_regression)

- Commons Math: The Apache Commons Mathematics Library

<http://commons.apache.org/proper/commons-math/>

## Material Complementar

- Matriz Inversa

[https://pt.wikipedia.org/wiki/Matriz\\_inversa](https://pt.wikipedia.org/wiki/Matriz_inversa)

- Regressão não linear

[https://pt.wikipedia.org/wiki/Regress%C3%A3o\\_n%C3%A3o\\_linear](https://pt.wikipedia.org/wiki/Regress%C3%A3o_n%C3%A3o_linear)

- Como fazer uma regressão linear simples no Excel

<https://www.voitto.com.br/blog/artigo/regressao-linear-simples-no-excel>

- Análise de Regressão no LibreOffice

[https://help.libreoffice.org/Calc/Regression\\_Analysis/pt-BR](https://help.libreoffice.org/Calc/Regression_Analysis/pt-BR)

## Imagem do Dia





# Sistemas de Apoio à Decisão

<http://lives.ufms.br/moodle/>

Rafael Geraldeli Rossi  
rafael.g.rossi@ufms.br

Slides baseados no material do prof. Bruno Magalhães nogueira e em [Vercellis, 2011]

## Referências Bibliográficas I



Vercellis, C. (2011).

*Business intelligence: data mining and optimization for decision making.*

John Wiley & Sons.