Aprendizado de Máquina II

Regressão Linear

Profª. Renata De Paris

Especialização em Ciência de Dados

UCS
UNIVERSIDADE
DE CAXIAS DO SUL

Roteiro da Aula

- Definição
- Regressão Linear
- Avaliação do desempenho
- Prática
- Atividade

Regressão

- Prever um valor para uma dada variável contínua, baseado nos valores de outras variáveis, assumindo um modelo de dependência linear ou não linear.
- Busca minimizar os **erros quadráticos**.
- Muito estudada em estatística, redes neurais, etc.
- Exemplos:
  - Prever **vendas de um novo produto**, baseado nos gastos com propaganda.
  - Prever **velocidade do vento** como uma função da temperatura, umidade, pressão do ar, etc.
  - **Previsão de séries temporais** para índices em mercados de ações.

Regressão – Exemplo Aplicação

- Qual é o valor de preço de venda da minha casa?

Tabela 1. Valores da casa para o modelo de regressão

Tamanho da casa (pés quadrados)	Tamanho do lote	Quartos	Granito	Banheiro reformado?	Preço de venda
3529	9191	6	0	0	$205,000
3247	10061	5	1	1	$224,900
4032	10150	5	0	1	$197,900
2397	14156	4	1	0	$189,900
2200	9600	4	0	1`	$195,000
3536	19994	6	1	1	$325,000
2983	9365	5	0	1	$230,000
3198	9669	5	1	1	????

Regressão Linear:

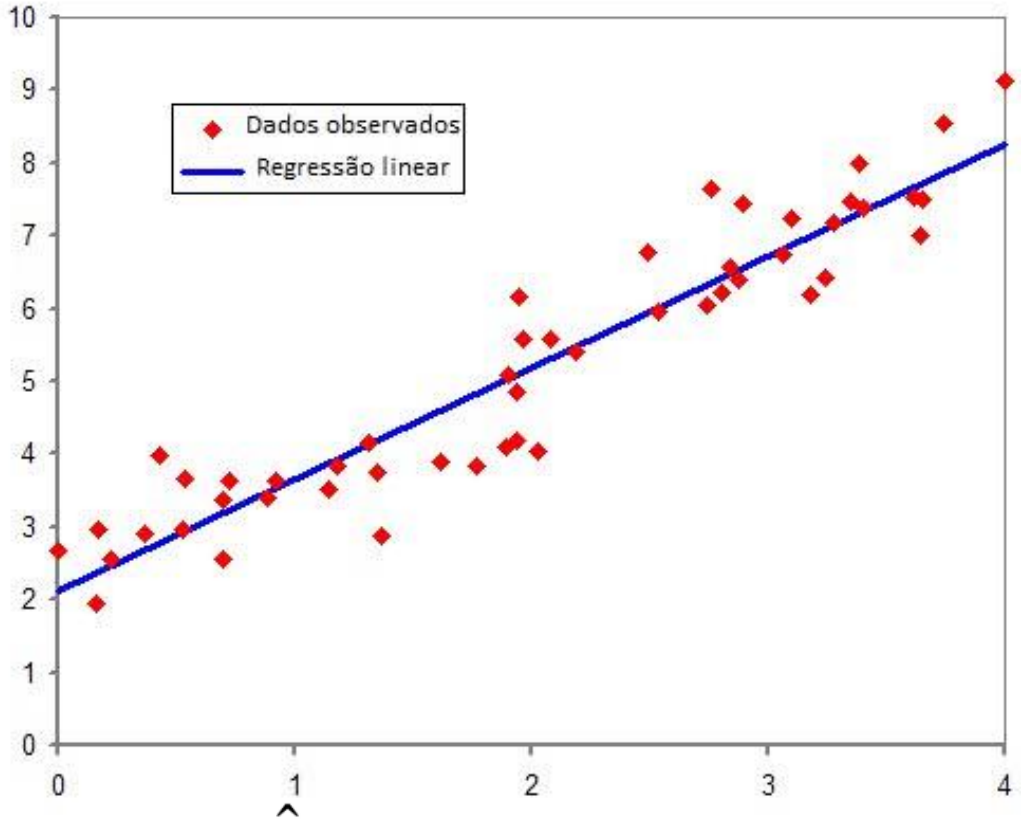
$x_1 + x_2 + x_3 + x_4 + x_5 = y$

Regressão Linear

- Utiliza **pesos/coeficientes** para aprender uma representação que se aproxime ao máximo dos dados do treino.
- Os pesos são atualizados conforme a função que minimiza os erros.
- Predição supervisionada/preditiva.
- Por exemplo:
  - Features (entradas):
    - $x_1 + x_2 + x_3 + x_4 + x_5$
  - Pesos:
    - $p_0 + p_1 + p_2 + p_3 + p_4 + p_5$
  - Aplicação dos pesos:
    - $p_0 + p_1*x_1 + p_2*x_2 + p_3*x_3 + p_4*x_4 + p_5*x_5 = y_i$
- O resultado de saída (yi - predição) é então comparado ao valor real de Y.

Regressão Linear

- O resultado da regressão linear seria uma reta



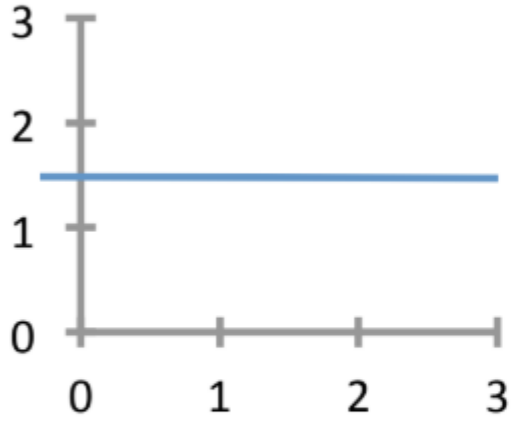
$$\hat{f}(x) = \theta_0 + \theta_1 x$$

Intercept
Representa o ponto y quando x=0

Slope
Representa a inclinação da reta para mais ou para menos

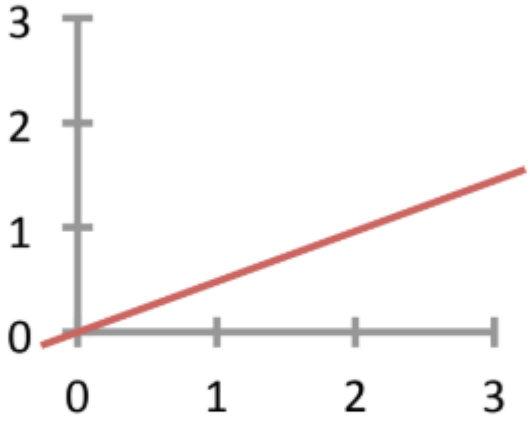
Regressão Linear

$$\hat{f}(x) = \theta_0 + \theta_1 x$$



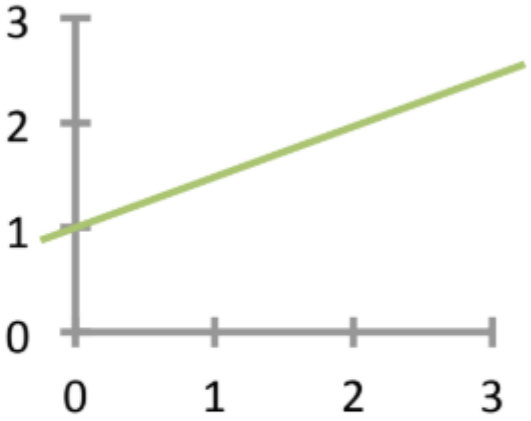
$\theta_0 = 1.5$
$\theta_1 = 0$

$$\hat{f}(x) = 1.5 + 0x$$



$\theta_0 = 0$
$\theta_1 = 0.5$

$$\hat{f}(x) = 0 + 0.5x$$



$\theta_0 = 1$
$\theta_1 = 0.5$

$$\hat{f}(x) = 1 + 0.5x$$

Regressão Linear

- Regressão Univariada

$$\hat{f}(x) = \theta_0 + \theta_1 x$$

- Regressão Multi-variada

$$\hat{f}(\mathbf{x}) = \theta_0 + \theta_1 x^1 + \theta_2 x^2 + \ldots + \theta_m x^m$$

$$\hat{f}(\mathbf{x}) = \theta_0 + \sum_{i=1}^{m} \theta_i x^i$$

Regressão Linear

- Exemplo: Cotações Diárias da Petrobrás
  - Usar regressão linear para prever o valor de fechamento da ação da Petrobras em um dia específico
  - Qual será a cotação da ação para os seguintes valores:
    - Abertura: 12.30
    - Máxima: 12.35
    - Mínima: 12.20
    - Fechamento: ?
  - A regressão linear aplica os pesos nos dados.
  - Digamos que o regressor definiu os seguintes pesos:
    - P0 = 1
    - P1 = 0.7
    - P2 = 0.06
    - P3 = 0.08

Regressão Linear

- Exemplo: Cotações Diárias da Petrobrás

  - Abertura: 12.30
  - Máxima: 12.35
  - Mínima: 12.20
  - Fechamento: ?

  - P0 = 1
  - P1 = 0.7
  - P2 = 0.06
  - P3 = 0.08

- Aplicando os valores teríamos a seguinte equação:
  - Y = p0+p1*x1+p2*x2+p3*x3
  - Y = 1+0.7*12.30+0.06*12.35+0.08*12.20
  - Y =1+8.61+0.74+0.97 = **11.32**

- Valor predito: **11.32**

Regressão Linear

- Exemplo: Cotações Diárias da Petrobrás
  - Usar regressão linear para prever o valor de fechamento da ação da Petrobras em um dia específico
  - O valor predito pelo regressor foi de 11.32
  - Qual é a acurácia desse valor?
  - Levando em consideração que o valor real é de 12.33, temos que calcular o erro e ajustar novamente os pesos
  - Erro absoluto: 11.32-12.33 = -1,01

Regressão Linear

- Como minimizar o erro?
  - Gradiente Descendente
    - Algoritmo usado para minimizar o erro dos pesos do modelo
    - Utiliza o erro médio quadrático entre o valor predito e o valor real
    - Utiliza todos os dados de treinamento de forma interativa até o menor erro possível
    - É preciso parametrizar o valor da **taxa de aprendizado** (*learning rate*).
      - Controlar o nível de aprendizado a cada iteração do algoritmo

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$$

Regressão Linear

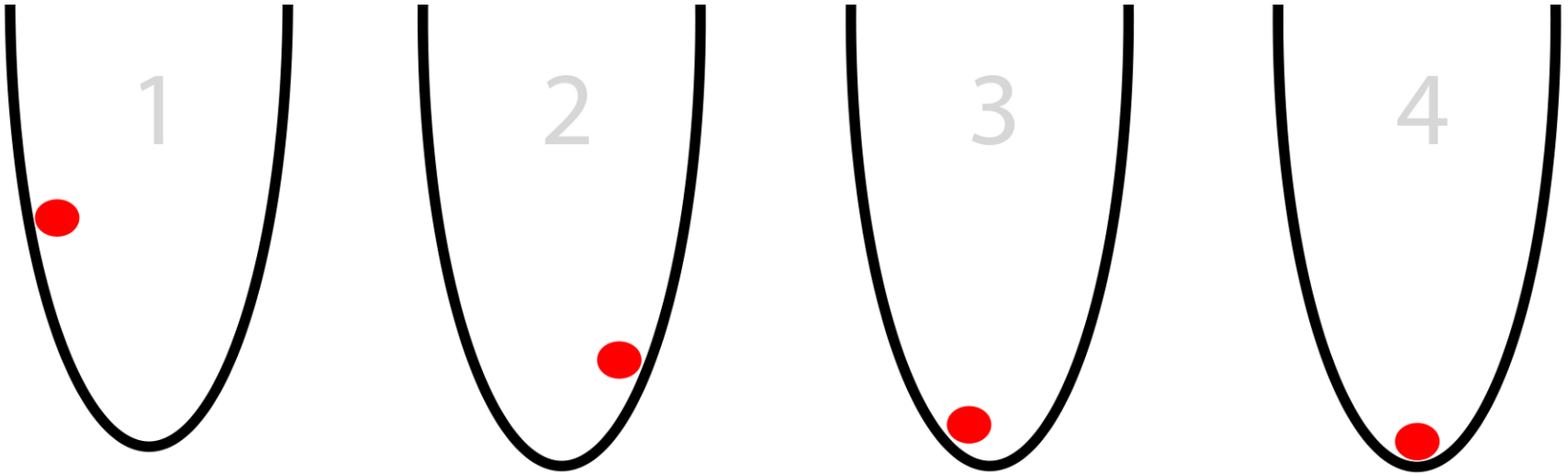
- Gradiente Descendente
  - Exemplo
    - p0=1, p1=0.9, x1=12.30
    - Y = 12.33 (valor real)
  - Aplicando os valores na equação:
    - Yi = 1+0.9*12.30
    - Yi = 12.07
  - Calculando o **erro** a partir do valor predito:
    - ε = 12.07 – 12.33
    - ε = -0.26

Regressão Linear

- Gradiente Descendente – Atualização de pesos
  - O objetivo é calcular o novo valor dos pesos p0 e p1
    - Valores originais: p0=1, p1=0.9, x1=12.30
  - A **taxa de aprendizado** é chamada de alpha (**α**)
  - Para α = 0.01 temos
    - p0 = p0 – α * **ε**
    - p0 = 1 - 0.01*-0.26
    - p0 = 1,0026
  - O valor de p1 deve ter influência no valor da feature associada a ele.
    - p1 = p1-α * **ε** * x1
    - p1 = 0.9 – 0.01*-0.26*12.30
    - p1 = 0.96198

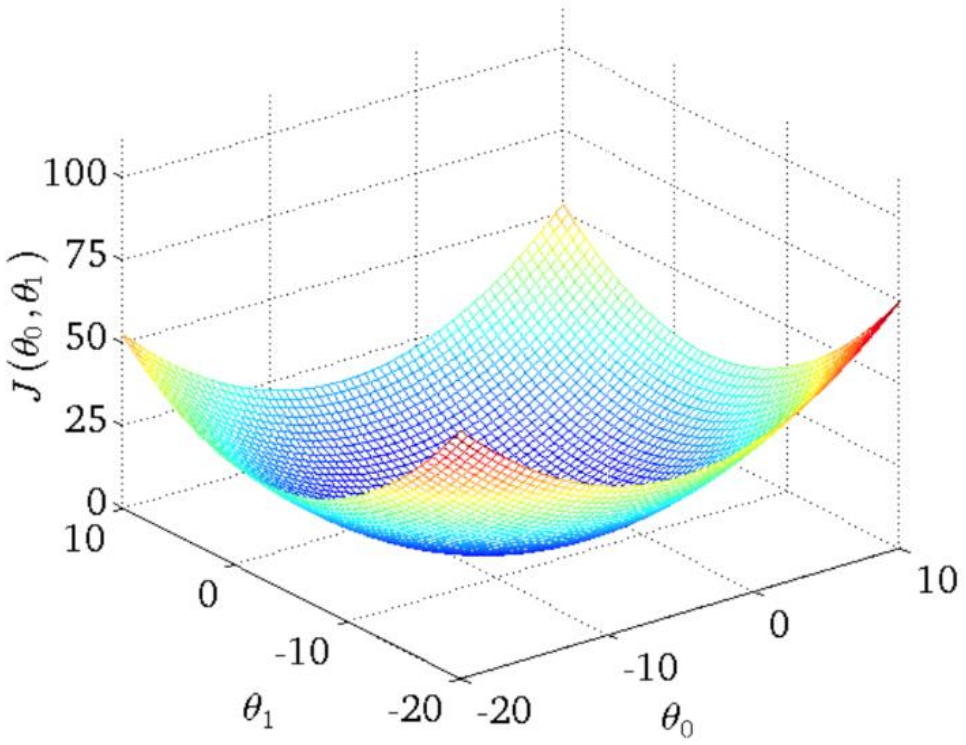
Regressão Linear

- Gradiente Descendente – Atualização de pesos
  - Gradiente Descente repete esse processo a cada instância do treino até que os pesos se ajustem com o **mínimo de erro** (bolinhas vermelhas).
  - **Época** = cada ciclo completo
  - Após várias épocas é possível chegar ao ponto mínimo de erro.



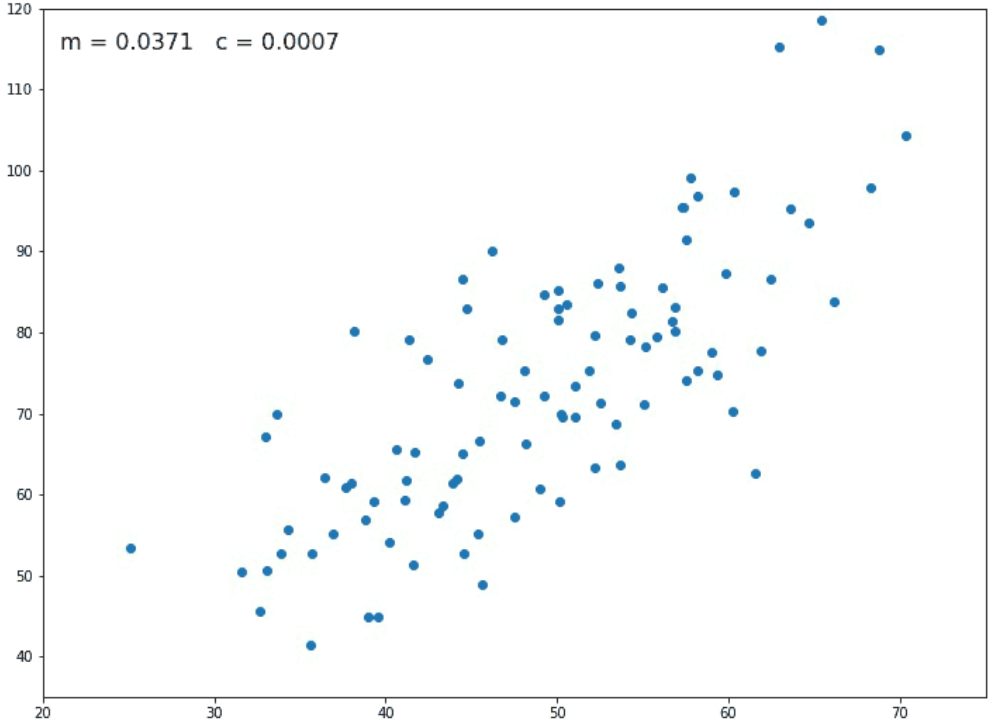
Regressão Linear

- Gradiente Descendente – Atualização de pesos
  - Função de custo do erro quadrático é convexa
  - Função tem único mínimo
  - Formato de bacia (*bowl shaped)*



Regressão Linear

- Gradiente Descendente
  - Deve reduzir a função de custo a cada iteração.
  - Quando parar de executar?
    - Quando convergir
    - A execução do algoritmo deve parar quando duas iterações consecutivas for menor que o limiar ε.



Regressão Linear – Prática

Dataset: Bolsa de Ações da Petrobras

Atividade

- A partir do dataset escolhido para trabalhar com os métodos supervisionados e utilizando a biblioteca scikitlearn da linguagem de programação Python, realize as seguintes tarefas:

1. Escolha 2 ou mais atributos contínuos do dataset, sendo 1 deles o atributo classe.
2. Execute o algoritmo de regressão linear com cross-validation e depois hold-out, escolhendo as divisões mais adequadas para ambos os métodos conforme o tamanho do dataset.
3. Avalie o desempenho do algoritmo por meio do erro médio absoluto (MAE).
4. Aumente ou diminua a quantidade de atributos/features para tentar melhorar a acurácia do resultado.

Referências

- Breiman, L., Freidman, J., Olshen, R. e Stone, C. (1984). *Classification and Regression Trees.* Wadsworth International Group., USA.
- Faceli, K.; Lorena, A.C.; Gama, J.; de Carvalho, A.C.P.L.F. *Inteligência Artificial: Uma abordagem de aprendizado de máquina.* LTC, Rio de Janeiro, 2011.
- Quilan, R. (1979). *Discovering rules by induction from large collections of examples.* In: Michie, D. (Ed.) Expert Systems in the Microelectronic Age, p. 168-201. Edinburgh University Press.
- Quilan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Mateo, CA, USA.
- TAN, P-N; STEINBACH, M.; KUMAR,V. *Introduction to Data Mining.* Pearson, 2006.