Regressão
Mineração de Dados

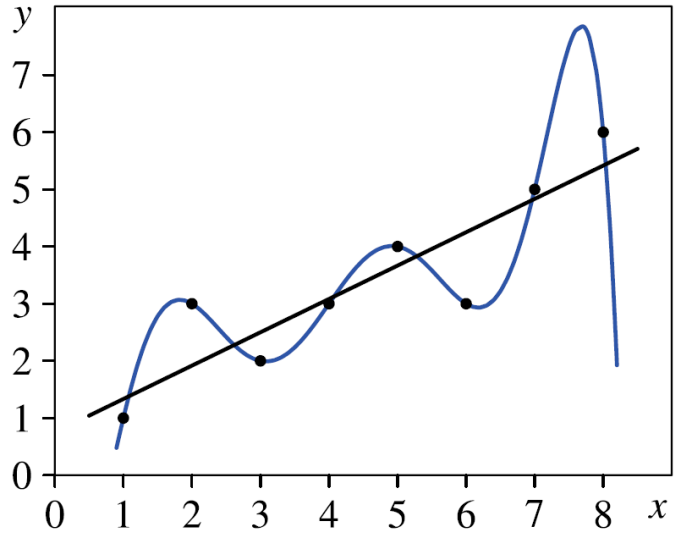
Ronaldo C. Prati

Universidade Federal do ABC (UFABC) - ronaldo.prati@ufabc.edu.br

A tarefa da regressão

Objetivo: Predizer como a variável alvo em função dos outros atributos

- Atributo alvo -> ***Variável resposta***
- Atributos descritivos -> **Variáveis regressoras**



Abordagem Paramétrica

- Estimar os parâmetros de uma classe de funções *f* que descrevem esse relacionamento
- Examplo: Linha (preto) v.s. polinômio (azul) de grau 7

Abordagem Não-Paramétrica

- Predizer o valor sem explicitar o relacionamento
- K-vizinhos (média), redes neurais

Tarefa de Regressão

Dado um conjunto de dados com n exemplos:

- encontrar uma função $f(x)$ que prediga o valor de y para novos exemplos

- minimizando a expectativa erro $E(f(x),y)$

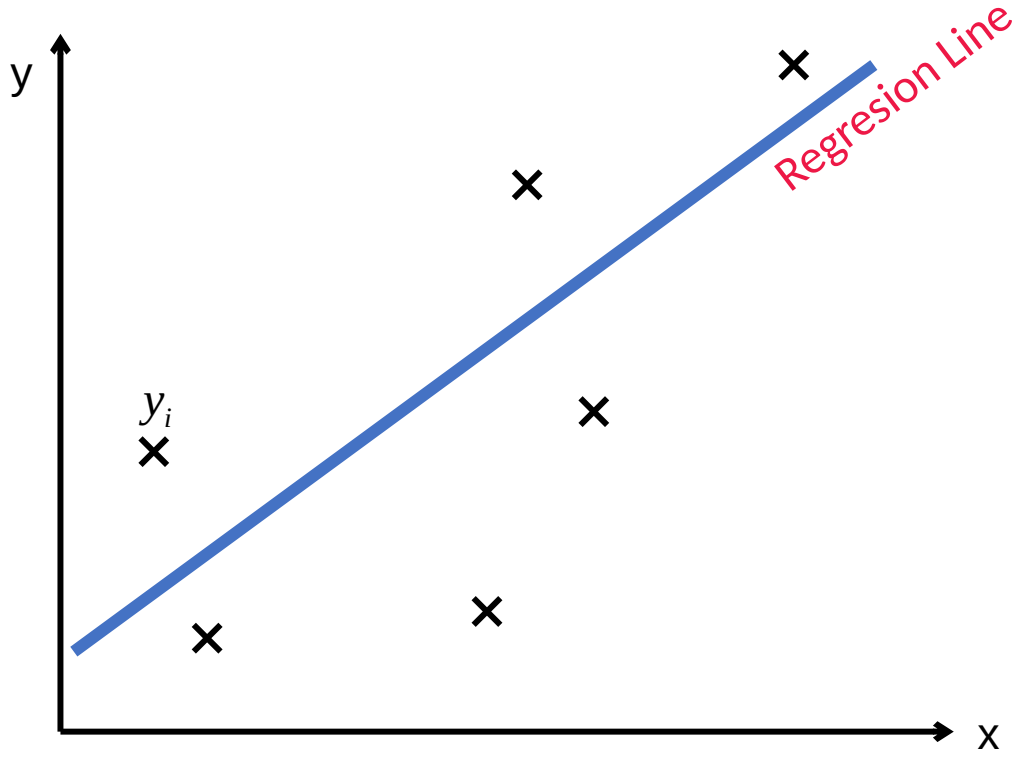
Regressão Linear

- Dado um conjunto de dados com dois atributos contínuos, *x* e *y*, encontrar uma reta $y \approx a + bx$ que aproxima a denpendência linear entre *x* e *y*.

- Para encontrar essa reta de regressão, queremos encontrar os coeficientes *a* e *b* que melhor se ajustam a essa relação

Regressão Linear

- O que é um bom ajuste?

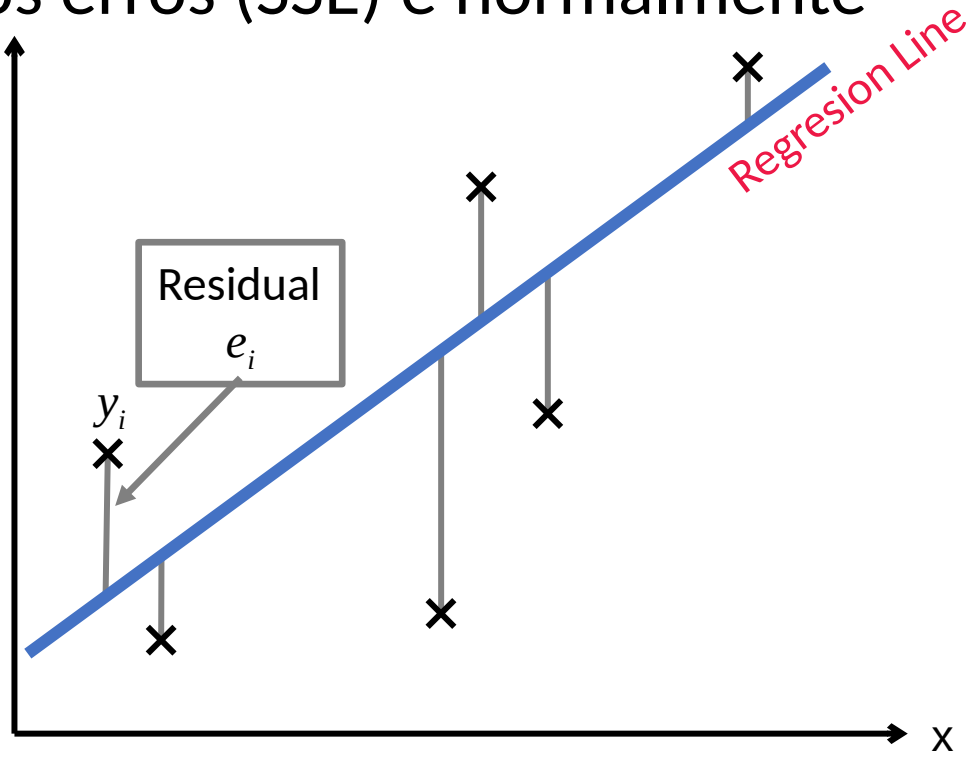


Função de custo

O erro, ou **resíduo**, é calculado a cada ponto

A soma dos quadrados dos erros (SSE) é normalmente escolha como função de custo (a ser minimizada)

É chamado de ***método dos mínimos quadrados***



Função de custo

Soma dos quadrados dos erros

Outras funções possíveis

- Erro absoluto médio
- Distância Euclidiana média
- Distância máxima absoula na direção y
- Distância Euclideana máxima
- . . .

Construção

- Considere a $\hat{y} = f(x) = a + bx$

- Encontre a e b para os exemplos (x_i, y_i) o mais próximo possível
- Minimizar

$$\text{SSE } F(a,b) = \sum_{i=1}^{n} (f(x) - y_i)^2 = \sum_{i=1}^{n} (a + bx_i - y_i)^2$$

- Objetivo: os valores de y computados pela equação linear (soma dos quadrados) deve desviar o menos possível dos valores medidos

Construção

SSE é minimizado se as derivadas parciais são zero

$$\frac{\partial F}{\partial a} = \sum_{i=1}^{n} 2(a + bx_i - y_i) = 0$$

$$\frac{\partial F}{\partial b} = \sum_{i=1}^{n} 2(a + bx_i - y_i)\, x_i = 0$$

Construção

Como resultado, temos a chamada equações normais da regressão:

$$na + \left(\sum_{i=1}^{n} x_i\right) b = \sum_{i=1}^{n} y_i$$

$$\left(\sum_{i=1}^{n} x_i\right) a + \left(\sum_{i=1}^{n} x_i^2\right) b = \sum_{i=1}^{n} x_i\, y_i$$

Esse é um sistema de duas equações e duas incógnitas que tem uma única solução, e que podem ser encotradas resolvendo o sistema

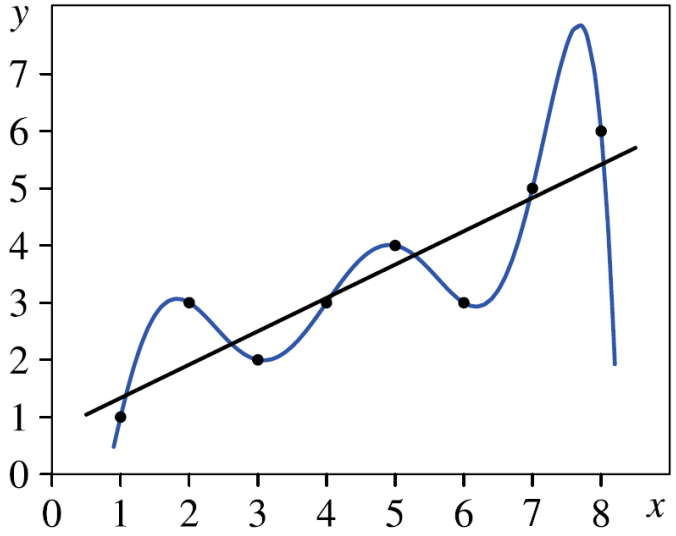
Exemplo: reta de regressão

Considere os dados

x	1	2	3	4	5	6	7	8
y	1	3	2	3	4	3	5	6

A reta regressão resultate é

$$y = \frac{3}{4} + \frac{7}{12}x$$



- A reta determina dessa maneira é chamada de **reta de regressão**.
- A linha de regressão pode ser interpretada como o **estimador de máxima verossimilhança** (MLE)
- Suposição: O processo de geração dos dados pode ser descrito pelo modelo $f(x) = a + bx + \xi$ em que ξ é o erro aleatório
- Os parâmetros *a* e *b* que minimizam a soma dos quadrados (na direção de y) a partir dos dados maximizam a probabilidade dos dados, assumindo essa classe de modelos.

Outros métodos

Além da equação normal, também é possível encontrar a reta da regressão linear usando:

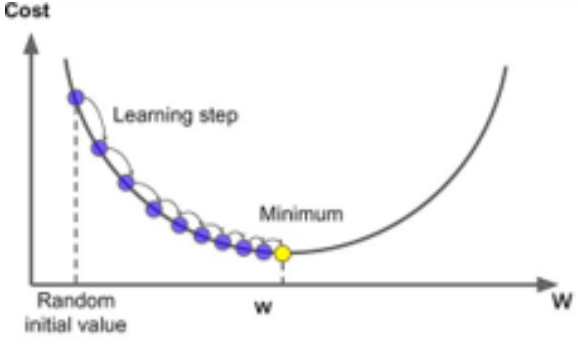
- Descida do gradiente

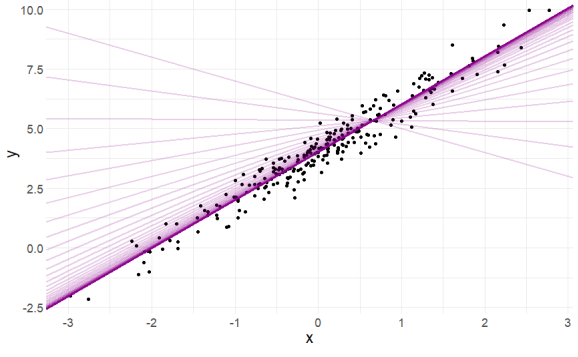
- Método de Adams

- Decomposição em valores singulares

Descida do Gradiente

- Método interativo
- Inicializa com valores aleatórios
- Dá um "pequeno passo" na direção do gradiente
- Repete o processo até convergir



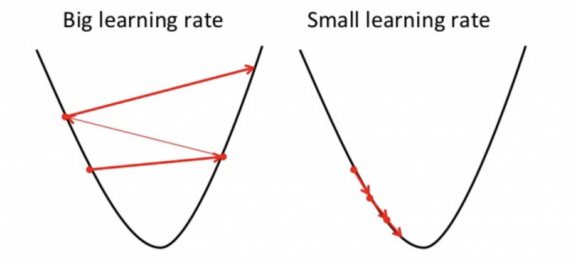


Taxa de aprendizado

O tamanho do passo é dada pela taxa de aprendizado, e controla a taxa de convergência

- Passo pequeno, demora em convergir

- Passo grange, pode não convergir



Outras regressões

Regressão Multivariada

- Podemos aplicar a regressão linear quando temos mais de uma variável regressores
- Nesse caso, ao invés de de uma reta, temos um hiperplano:

$$y = f\left(x_1, x_2, \ldots, x_m\right) = a_0 + \sum_{k=1}^{m} a_k x_k$$

com a função de custo:

$$F\left(a_0, a_1, \ldots, a_m\right) = \sum_{i=1}^{n} \left(f\left(\mathbf{x}_i\right) - y_i \right)^2$$

$$\sum_{i=1}^{n} \left(a_0 + a_1 x_{i1} + a_2 x_{i2} + \cdots + a_m x_{im} - y_i \right)^2$$

Regressão Polinomial

- O método dos mínimos quadrados pode ser ampliado para polinômios de grau m

$$y = p(x) = a_0 + a_1 + a_2 x^2 + \ldots + a_m x^m$$

- Encontrar a_i que minimiza a função de erro

$$F(a_0, a_1, \ldots, a_m) = \sum_{i=1}^{n} \left(p(x) - y_i \right)^2 = \sum_{i=1}^{n} \left(a_0 + a_1 + a_2 x^2 + \ldots + a_m x^m - y_i \right)^2$$

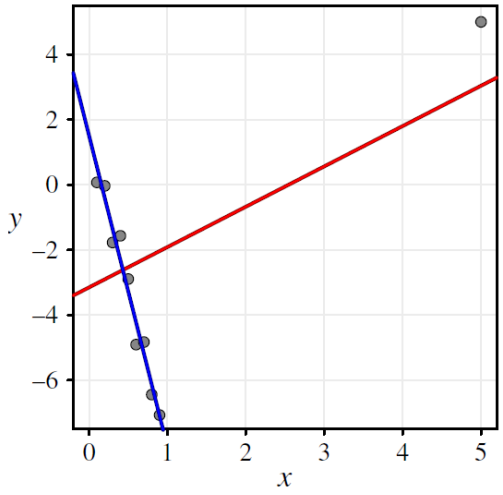
Regressão Robusta

Regressão Robusta

- A regressão por mínimos quadrados é sensível a outliers
- Solução: ***regressão robusta***

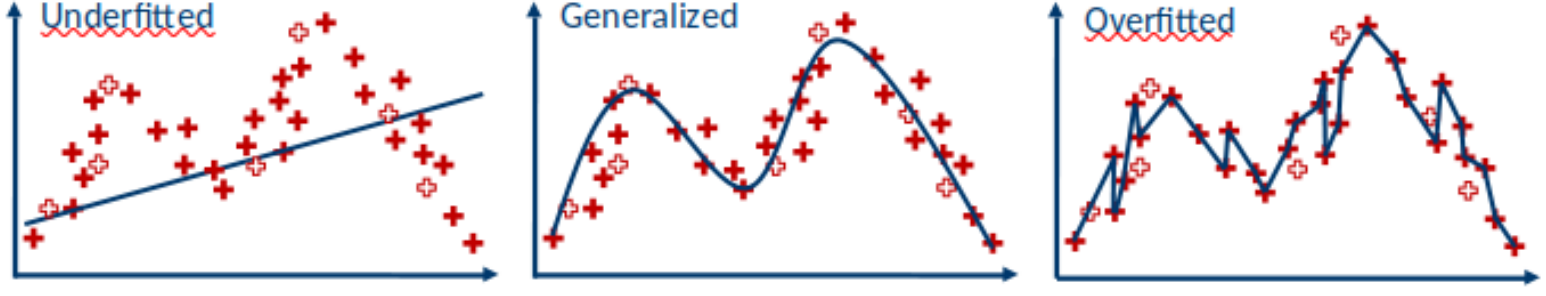
Regressão Robusta

- Um outlier extremo influencia a regressão por mínimos quadrados
- A influência do outlier é atenuada pelo uso de regressão robusta



Regressão Robusta

Overfitting	Underfitting
• Modelo que se ajusta aos dados de treinamento muito bem, incluindo outliers • Impacto negativo na capacidade do modelo em generalizar	• Um modelos que nem se ajusta bem aos dados de treino, nem generaliza para novos dados



Regularização

- Existe um trade-off entre a simplicidade do modelo e o seu fit
- **Regularização** é um termo matemático que permite introduzir informações adicionais para resolver diversos problemas pouco definidos.
- Dentro do contexto de aprendizado de máquina, ele funciona como uma penalização para modelos complexos, e pode ser usado para controlar o trade-off entre simplicidade e especificidade.
- Pode ser usado para evitar o overfitting.

Regressão LASSO

- Na regressão LASSO, é adicionado um termo à função de custo correspondente à soma dos coeficientes:

$$\mathrm{RSS}_{\mathrm{lasso}} = \sum_{i=1}^{n} \left[y_i - (\mathbf{w} \cdot \mathbf{x}_i + b) \right]^2, \boxed{+ \alpha \sum_{j=1}^{p} |w_j|}$$

$$\text{regularização } \ell_1$$

- Quanto há muitos atributos altamente correlacionadas, a regressão LASSO “força” a escolha de apenas uma delas, pois selecionar um novo atributio pode contribuir pouco para o modelo

Regressão Rígida

- Assim como na regressão LASSO, na regressão rígida é adicionada um termo de penalização. Mas esse termo é o quadrado dos coeficientes, ao invés do módulo

$$\mathrm{RSS}_{\mathrm{ridge}} = \sum_{i=1}^{n} \left[y_i - (\mathbf{w} \cdot \mathbf{x}_i + b) \right]^2 + \alpha \sum_{j=1}^{p} w_j^2$$

regularização ℓ_2

- A regressão rígida previne que poucos atributos dominem o modelo

Elastic Net

- A elastic net é uma combinação da regressão lasso com a regressão rígida

$$\mathrm{RSS}_{\mathrm{elasticnet}} = \sum_{i=1}^{n} \left[y_i - (\mathbf{w} \cdot \mathbf{x}_i + b) \right]^2 + \alpha_1 \sum_{j=1}^{p} \left| w_j \right| + \alpha_2 \sum_{j=1}^{p} w_j^2$$

Outros algoritmos de regressão

Regressão Não Linear

- Podemos aplicar uma função não linear a cada atributo, antes de calcular o coeficiente

  Nesse caso, podemos ver a regressão linear como um caso particular, em que aplicamos a função identidade

  Funções pode ser o quadrado, logarítmo, raíz, módulo, seno, exponencial, etc...

- Identificar as funções apropriadas é um problema

Regressao usando árvores

- Um problema com regressão linear é que a maioria dos conjuntos de dados é não-linear.

  A formulação não-linear também é problemática, pois nãoé trivial definir a função que iremos usar!

- Árvores de regressão

  Ideia: Usar uma estrutura de árvore (dividir e conquistar) para dividir os exemplos de tal maneira que eles possam ser melhor modelados linearmente pelos exemplos que estão no nó “folha”.

  Dessa maneira, os ramos são similiares a árvore de decisão, mas ao invés da classe no nó folha, temos que prever um valor numérico.

Regressao usando árvores

- Árvore de regressão: a folha contém um valor numérico,

  geralmente a média dos exemplos que caem nela.

- Model Trees: As folhas contém uma regressão (geralmente linear) para prever o valor dos exemplos que caem nela.

  A árvore de regressão é uma model tree “constante”.

Regressao usando árvores

- Para fazer o crescimento da árvore, uma alternativa para escolher o atributo é o uso do Standard Deviation Reduction

  trata o desvio padrão do atributo meta como uma medida de erro no nó e maximiza a redução desse valor a cada divisão.

K-Vizinhos mais próximos

- Calcula a média ou mediana dos k-vizinhos mais próximos, ao invés da classe mais frequente

Outras opções

- SVR (vetores de suporte para regressão)
- Redes Neurais
- Ensembles

Análise do Modelo

- Existem várias abordagens para avaliar os erros obtidos por um modelo de regressão
- Um gráfico que costuma ser útil consiste no gráfico de dispersão de valores preditos vs reais
- Os valores dos coeficientes de um modelo linear podem dar uma estimativa de importância para variáveis base sob algumas condições