Aula 15 - Algoritmos básicos: Regressão Linear e Regressão Logística.

Profa. Gabrielly Queiroz

Estrutura de um Algoritmo de ML

1. Coleta de Dados
2. Pré-processamento (limpeza, normalização, seleção de atributos)
3. Divisão em treino e teste
4. Treinamento do modelo
5. Avaliação (métricas)
6. Aplicação prática

Regressão Linear

A regressão linear é um método de aprendizado supervisionado usado para prever valores contínuos. Ela modela a relação entre uma variável dependente (y) e uma ou mais variáveis independentes (x), assumindo uma relação linear entre elas.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{\sum((x - \bar{x})(y - \bar{y}))}{\sum((x - \bar{x})^2)}$$

- y: variável dependente (o valor que queremos prever).
- x: variável independente (o valor usado para prever).
- β0, β1: coeficientes a serem ajustados.
- β0: **intercepto** (bias), o valor de y quando x=0.
- β1: **coeficiente angular**, que representa a inclinação da linha e a variação de y para cada unidade de mudança em x.

Objetivo: Minimizar o erro (soma dos resíduos ao quadrado).
Aplicações: Previsão de vendas, crescimento populacional, etc.

Vantagens e Desvantagens:

- Simplicidade e interpretabilidade.
- Limitação em problemas não lineares.

Em Machine Learning

O modelo recebe um conjunto de dados com pares de entrada (x) e saída (y).

Ele tenta encontrar os coeficientes β0 (intercepto) e β1 (inclinação) que minimizam o **erro** entre os valores previstos e os valores reais.

$$\text{Erro Total} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

y_i é o valor real,

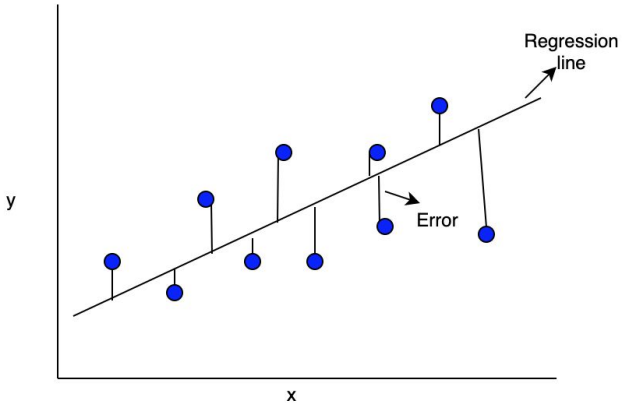
$\hat{y}_i = \beta_0 + \beta_1 x_i$ é o valor previsto pelo modelo.

Após ajustar a equação (ou seja, encontrar β0 e β1), o modelo é testado com novos dados (xteste).

Ele usa a equação ajustada para prever yteste e compara com os valores reais para verificar a precisão.

Erro Quadrático Médio (MSE):

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$



Exemplo

Prever o preço de uma casa com base na área (em m²).

Área (m²)	Preço (R$)
50	150.000
60	180.000
70	200.000
80	230.000
90	250.000

Exemplo

Prever o preço de uma casa com base na área (em m²).

Área (m²)	Preço (R$)
50	150.000
60	180.000
70	200.000
80	230.000
90	250.000

$$\bar{x} = \frac{\text{soma dos valores de área}}{\text{número de entradas}} = \frac{50 + 60 + 70 + 80 + 90}{5} = 70$$

$$\bar{y} = \frac{\text{soma dos valores de preço}}{\text{número de entradas}} = \frac{150000 + 180000 + 200000 + 230000 + 250000}{5} = 202000$$

$$\beta_1 = \frac{(50 - 70)(150000 - 202000) + \dots}{(50 - 70)^2 + \dots} = 4000$$

$$\beta_0 = 202000 - (4000)(70) = -80000$$

$$y = -80000 + 4000x$$

Prever o preço para x=75

$$y = -80000 + 4000(75) = 220000$$

Regressão Logística

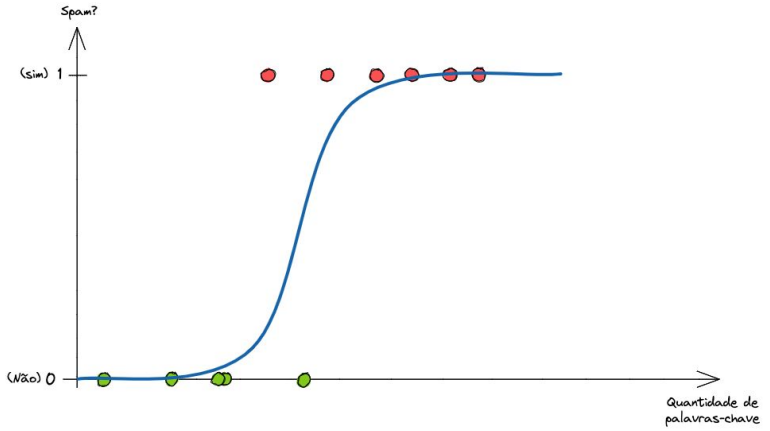
A regressão logística é usada para problemas de classificação, onde a saída é categórica (ex.: sim/não). Diferentemente da regressão linear, que gera valores contínuos, a regressão logística usa a função sigmóide para limitar os valores preditos entre 0 e 1, representando a probabilidade de pertencer a uma classe.

Fórmula:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

P(y=1 | x): probabilidade da classe positiva.

Se a probabilidade for maior que 0.5, y é classificado como 1; caso contrário, como 0.



Regressão Logística

Na regressão logística, **os parâmetros** (β0 e β1) **são encontrados usando um método numérico de otimização**, chamado **Máxima Verossimilhança**.

Em regressão logística: você **ajusta** os parâmetros buscando maximizar a chance dos dados observados acontecerem (**não dá pra calcular direto por fórmula básica**).

Esse processo é feito com **algoritmos iterativos**, como o **Gradiente Descendente**. Esses algoritmos funcionam assim:

1. Começam com valores aleatórios ou nulos para β0 e β1;
2. Calculam o quanto esses valores erram nas previsões (usando uma função chamada *log-verossimilhança*);
3. Ajustam os valores de β0 e β1 **um pouquinho por vez**, sempre tentando **reduzir o erro**;
4. Repetem isso centenas ou milhares de vezes, **até encontrar os melhores valores possíveis**.

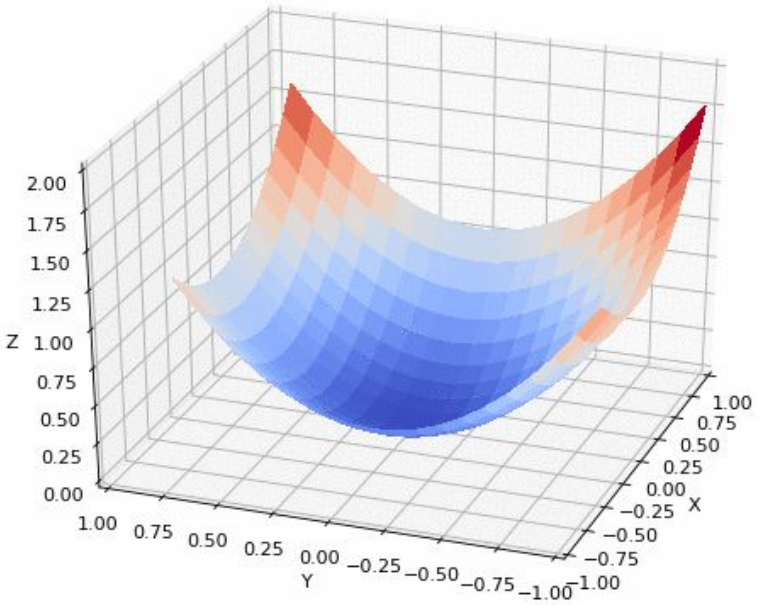
Gradiente Descendente

A função de erro (ou função de custo) usada na regressão logística é a **log-verossimilhança:**

$$J(\beta) = -\sum_{i=1}^{n} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)]$$

Onde $p_i = \frac{1}{1+e^{-(\beta_0+\beta_1 x_i)}}$ é a **probabilidade prevista.**

O gradiente descendente testa diferentes combinações de β0 e β1, calcula o erro para cada uma delas, e vai ajustando os valores até encontrar o ponto onde o erro é o menor possível.



Exemplo

Classificar pacientes como "doente" ou "saudável" com base na idade.

Idade	Diagnóstico (0 = saudavel, 1 = doente)
25	0
30	0
35	1
40	1
45	1

Exemplo

Classificar pacientes como "doente" ou "saudável" com base na idade.

Idade	Diagnóstico (0 = saudavel, 1 = doente)
25	0
30	0
35	1
40	1
45	1

$$\beta_0 = -10,\ \beta_1 = 0.25$$

Previsão para idade x=28

$$P(y = 1|x = 28) = \frac{1}{1 + e^{-(-10+0.25(28))}}$$

$$P(y = 1|x = 28) = \frac{1}{1 + e^{-3}} \approx 0.952$$

Se $P > 0.5$, $y = 1$ (doente). Caso contrário, $y = 0$ (saudável).

Para $x = 28$, o paciente é classificado como **doente**.

Atividade

Uma pesquisadora está estudando a relação entre o número de horas de estudo por semana e a nota final dos alunos em uma disciplina. A tabela a seguir mostra os dados coletados:

Horas de Estudo	Nota Final
5	55
10	60
15	65
20	75
25	85

Com base nesses dados, monte a equação da reta de regressão linear que melhor representa a relação entre as horas de estudo e a nota final. Depois, utilize essa equação para prever a nota de um aluno que estuda 18 horas por semana.