📙

# Ciclo 03 - O sistema supervisionado de aprendizado II

🤖 Fundamentos Machine Learning



# Como funciona o aprendizado supervisionado em regressão

## Aprendizado supervisionado em regressão

No aprendizado supervisionado, temos um conjunto de dados de treinamento com característica observadas do fenômeno (colunas do dataset) e uma variável que representa o fenômeno de interesse.

Nesse tipo de aprendizado, estamos interessados em aprender duas coisas:

1. **Estudar o fenômeno**
    - Estudar e entender quais características conseguem explicar mais ou menos o comportamento do fenômeno de interesse, ou seja, quais impactam mais ou menos nas mudanças do fenômeno.
        - Quanto o custo de vida de uma região explica o salário de um profissional?
        - Quanto a taxa de analfabetismo explica o índice de violência?
        - Quanto o preço, o investimento em marketing, a região, o canal de distribuição, a idade, o valor vigente do salário mínimo no país e a quantidade do saldo na conta de poupança explicam a variação do número de vendas?

2. **Criar um modelo matemático**
    - Criar um modelo matemático que relaciona as variações das características observadas com as variações da variável de interesse.
        - Qual a previsão de vendas para os próximos 30 dias?
        - Qual a previsão do custo de aquisição de clientes?
        - Quantos alunos desistirão da formação nos próximo 14 dias?
        - Qual a previsão das respostas corretas de uma prova do Enem?

O processo de aprendizagem das relações entre uma variável alvo e suas características é analise de regressão ou aprendizado supervisionado do tipo regressão.

## Tarefas de Regressão

Usamos regressão quando o conjunto de treinamento possui o rótulo do fenômeno observado como uma variável real ou contínua, como o salário e o peso, por exemplo.

Nesse tipo de tarefa, o algoritmo tenta ajustar um novo ponto, ao conjunto de pontos do conjunto de dados, de modo que apresente a menor distância possível do conjunto de dados.

**Exemplos de algoritmos de regressão**

1. Linear Regression
2. Regularized Linear Regression
    a. Ridge Regression - L2 Norm
    b. Lasso Regression - L1 Norm
3. Polynomial Regression
4. Neural Network Regression
5. Decision Tree Regression
6. Random Forest Regression
7. KNN Regression
8. Gaussian Regression

# Linear Regression : Teoria

## O que é Regressão Linear?

A regressão linear é um método estatístico que permite resumir e estudar a relação entre duas variáveis contínuas (quantitativas).

No fenômeno sendo estudado a relação entre as duas variáveis podem possuir duas naturezas:

1. **Determinístico**
    - Em um relação determinística, há uma equação que descreve exatamente a relação entre as duas variáveis. Um exemplo seria a conversão de temperaturas entre as escalas Celsius e Fahrenheit, uma vez que eu sei a temperatura em uma das escalas, o valor na outra está automaticamente definido.

2. **Estatístico**
    - A relação estatística entre duas variáveis não é perfeita. É utilizado quando não sabemos ou não possível definir com certeza a relação exata entre as variáveis, portanto devemos levar em conta uma incerteza. Nesse caso para um mesmo valor da variável $x$ podemos obter um valor ligeiramente diferente para variável $y$.

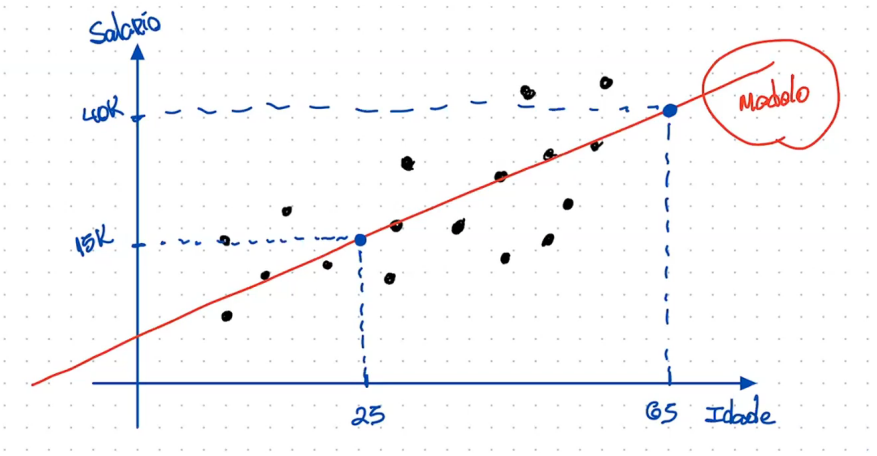
Em uma regressão linear estamos interessado em encontrar uma reta que se ajuste melhor aos dados. A linha que melhor se ajusta ao dados é aquela para a qual os $n$ erros de previsão (um para cada ponto observados) são os menores possíveis.

Imagine que queiramos estudar a relação entre o salário ganho por uma pessoa com a sua idade. Podemos inferir o comportamento geral entre as variáveis através de um modelo, nesse caso uma reta, que nos dirá qual será uma aproximação do salário para uma dada idade.

Nesse modelo hipotético, previmos (pontos em azul) que uma pessoa com $25$ anos deve ganhar em média $\$15000$, enquanto que alguém com $65$ anos ganha em média $\$45000$ segundo modelo.

No exemplo ao lado, observamos uma representação gráfica da reta, porém toda reta também possui uma representação algébrica dada por:

$$y = a + bx$$

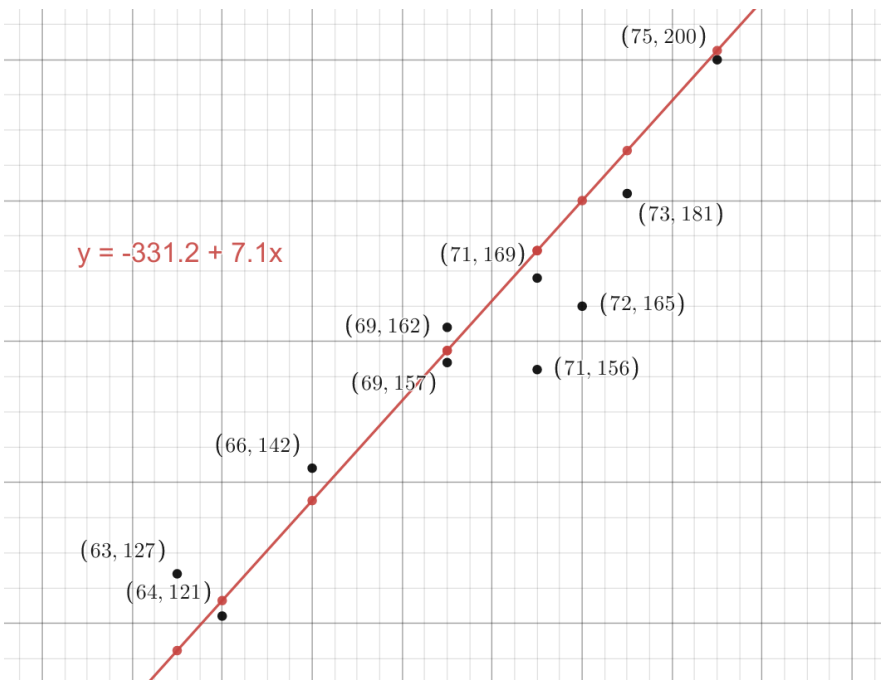



Onde:

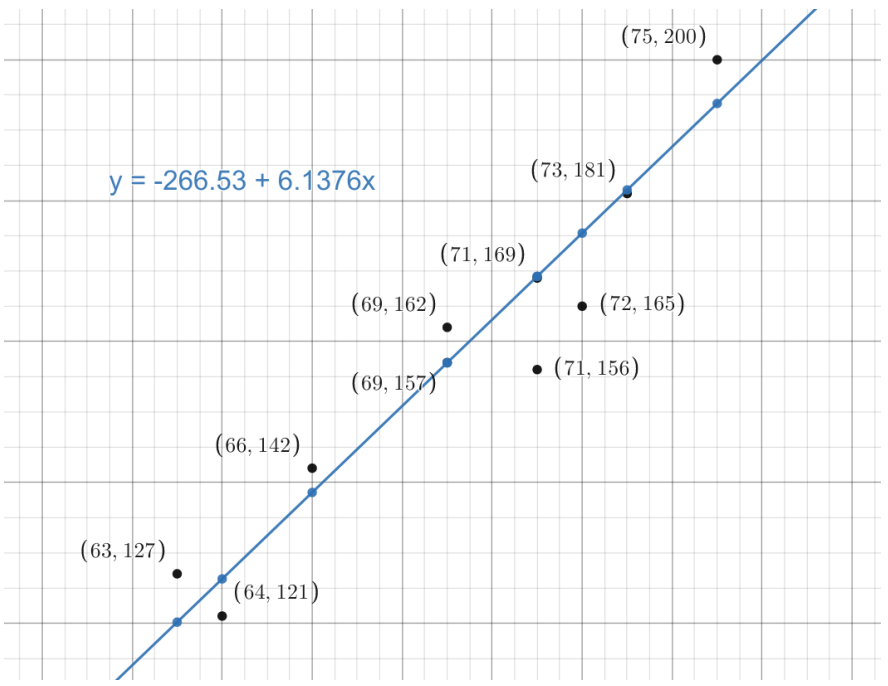
- $x$ → salário
- $y$ → idade
- $a$ → **intercepto** (coeficiente linear)
- $b$ → **slope** (coeficiente angular)

Uma das formas de encontrar os melhores valores dos parâmetros ($a$ e $b$) para produzir a reta com o menor erro geral é usando o critério dos mínimos quadrados. Esse critério minimiza a soma dos erros de previsão ao quadrado. Para um mesmo conjunto de dados podemos ajustar diferentes retas:

$$y = -331.2 + 7.1x$$




$$y = -266.53 + 6.1376x$$




Observando as retas acima, percebemos que a reta azul ajusta os dados melhor que a reta vermelha. O melhor ajuste é aquele que minimiza a soma do erro quadrático de todos os pontos, ou seja, devemos encontrar os valores de $a$ (intercepto) e $b$ (slope) que minimizam o valor de $SE$.

$$SE = \sum_{i=1}^{n} (y_i - (a - bx_i))^2$$

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

Os valores do intercepto e coeficiente angular que minimizam o erro quadrático (square error $SE$) estão listados exibidos abaixo:

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

## Os 5 passos para treinar um algoritmo de regressão linear

O treinamento de um modelo de regressão linear consiste principalmente em determinar os valores de $a$ e $b$.

Passo 1: Carregue os dados

Passo 2: Aplique a fórmula para encontrar o $b$ (coeficiente angular)

Passo 3: Use o valor de $b$ na fórmula para encontrar o valor de $a$ (intercepto)

Passo 4: Use o modelo da reta para prever os valores da variável resposta, a partir do modelo da reta.

Passo 5: Definir uma métrica de performance para calcular as previsões para dados nunca vistos

**Premissas assumidas**

1. A relação entre as características e a variável resposta é linear.
2. Os erros são independentes.
    - Um erro não interfere o outro.
3. Os erros são normalmente distribuídos.
4. Os erros para cada valor previsto tem variâncias iguais.
    - Esperamos que uma quantidade parecida de pontos sejam superestimados e subestimados pelo modelo.

> 💡 A falta de homogeneidade e representatividade nos dados pode quebrar uma ou mais premissas.

# Linear Regression : Prática

O modelo de regressão linear é usado para estudar a relação linear de influência entre as variáveis e a variável alvo, mas também é usado para criar um modelo e prever o valor futuro da variável alvo.

## Exemplos

1. **Previsão de vendas**
    - Quantos produtos serão vendidos nos próximos 4 meses?
2. **Estudos de performance esportiva**
    - Análise das relações entre fatores como temperatura, umidade, distância total do trajeto, hora do dia, inclinação do relevo, estação do ano, ingestão de água no dia e quantidade de sono com o tempo total de uma corrida.
    - Qual a contribuição de cada desses fatores na performance da corrida do atleta?
3. **Modelagem climática**
    - Qual a relação entre a emissão de gases de efeito estufa e o aumento da temperatura na Terra?
4. **Previsão do preço de venda da casa**

- Definição do valor de venda da casa em relação as suas características como o número de quartos, o tamanho da sala, a quantidade de garagens, localização, andar do apartamento e etc.

5. **Elasticidade do preço**
    - Se diminuir 10% o preço do produto, quanto % aumenta o número de vendas?

## Vantagens e desvantagens

**Vantagens**

1. Os resultados de um modelo de regressão linear são simples para explicar ao time de negócios.
2. Os modelos de regressão linear são eficientes computacionalmente para um grande volume de dados.
3. Modelos de regressão linear são úteis para estudar correlações entre variáveis numéricas.

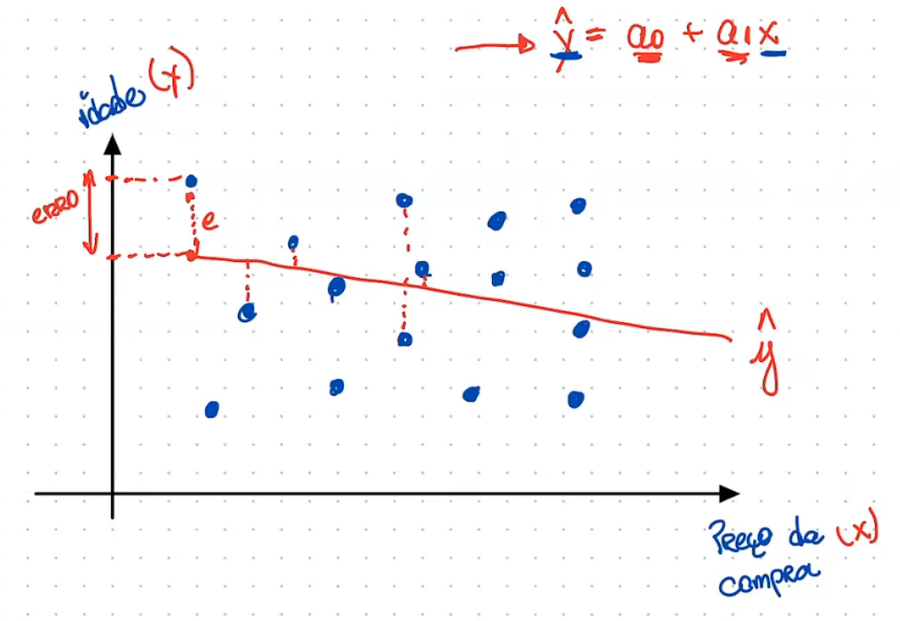
**Desvantagens**

1. Modelos de regressão linear assume linearidade entre as características e a variável resposta.
2. Não apresenta robustez na presença de outliers.
3. Seus coeficientes podem sofrer altas variações, a partir de pequenas mudanças nos dados, na presença de características altamente correlacionadas (multicolinearidade).

# Métricas de avaliação I: R2 - Teoria

Imagine que queremos entender qual é a relação entre a idade e o preço da compra.

Nesse gráfico observamos que não há um comportamento definido, temos um número parecido de pessoas de idade baixa que gasta pouco e pessoas de idade baixa que gasta muito, o mesmo vale para pessoas de idade alta. Percebemos que o comportamento de uma variável não nos diz muito sobre o valor da outra.

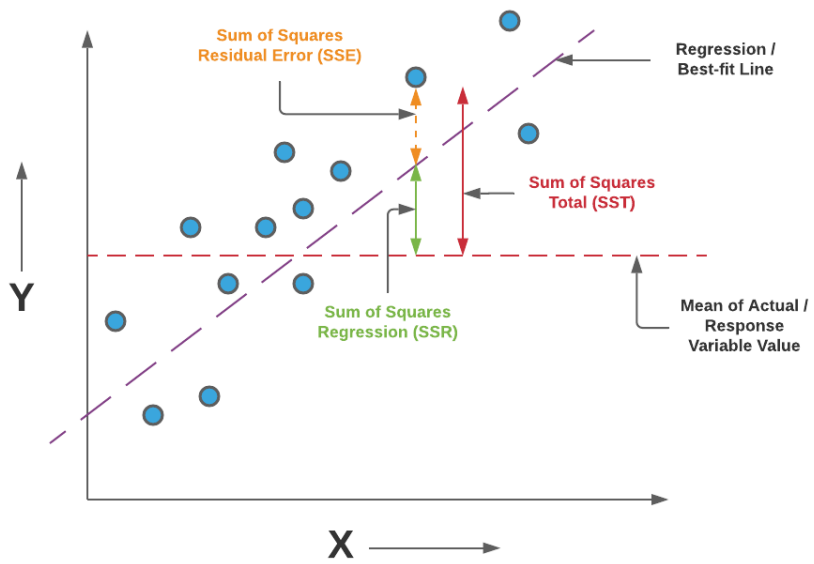



Usando regressão podemos ajustar uma reta para tentarmos entender a tendência nesse conjunto. Na regressão estamos interessado em encontrar os parâmetros $a_0$ e $a_1$que nos provê a reta $\hat{y}$ que melhor se encaixa no conjunto de dados, ou seja, estamos interessados no ajuste que minimiza a soma dos erros ($e_i = y_i - \hat{y}_i$) ao quadrado.

> 💡 Em estatística usamos o símbolo “^” para representar um estimador

$$SSE = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2$$

A somatória representada acima é chamada de **Sum of Square Residual Error (SSE)**.




Para avaliarmos se o ajuste é bom vamos introduzir outras duas medidas de erros. A primeira é chamada de **Sum of Squares Total (SST)** e ela mede a distância entre os pontos do gráfico e o ajuste dado pelo valor médio do variável alvo.

Esse ajuste é considerado a pior medida de tendência aceitável, nela estamos estimando o valor de variável alvo $y_i$ como sendo o valor médio $\bar{y}$ para qualquer valor de $x$.

$$SST = \sum_i (y_i - \bar{y_i})^2$$

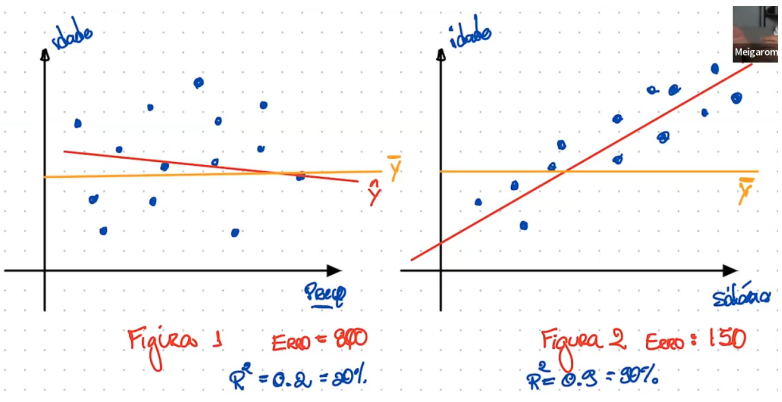
A outra medida é chamada de **Sum of Squares Regression (SSR)** e representa a soma das distâncias entre o melhor ajuste $\hat{y}$ e o ajuste da reta média $\bar{y}$:

$$SSR = \sum_i (\hat{y_i} - \bar{y_i})^2$$

Essas 3 medidas satisfazem a seguinte relação:

$$SST = SSE + SSR$$

Usando essas medidas de erro introduzimos um métrica chamada de $R^2$ que vai nos dizer, em termos vagos, o quanto conseguimos explicar o valor da variável alvo a partir das outras variáveis presentes no nosso conjunto de dados.




Em outras palavras, a métrica $R^2$ nos diz o quanto o nosso fenômeno está sendo explicado pela nossa reta ajustada.

- $R^2 \rightarrow 0$ : As features analisadas **não** são capazes de explicar o fenômeno. O comportamento (tendência) obtida pelo ajuste não é muito melhor do que simplesmente ajustar um média.

- $R^2 \rightarrow 1$: As features analisadas conseguem explicar o fenômeno bem. Quanto mais próximo de $1$ o valor de $R^2$ for, menor será a necessidade de analisarmos outras features.

$$R^2 = 1 - \frac{SSE}{SST}$$

# Métricas de avaliação II: MSE

O MSE (Mean Square Error) calcula a média do quadrado das diferenças entre os valores reais e preditos. O MSE é usado como uma função de perda (métrica de performance) e representa quão bem o modelo ajustou ao dados do treinamento. Quanto menor o MSE, melhor o ajuste do modelo ao dados.

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

> O MSE é um número único que nos diz o quão próximo as previsões estão dos valores reais na média.

> 💡 Todo o algoritmo de machine learning possui uma função de perda ( **Loss Function** ), que é usada para determinar o quanto o algoritmo errou após uma iteração do treinamento. Esse erro determinado pela **loss function** é levado em consideração na próxima iteração do algoritmo de modo a diminuir o valor obtido da função de perda.

**Os problemas do MSE:**

1. **Sensível na presença de outliers:** Um único grande erro pode impactar significativamente o valor final do MSE.
2. **Fora de escala:** O MSE é o erro quadrático, que não está na mesma escala da variável resposta. Por exemplo, o MSE da previsão de metro é metro ao quadrado. Não é possível comparar se o erro está bom ou ruim em relação a escala original.

# Métricas de avaliação III: RMSE

O RMSE (Root Square Mean Error) calcula a raiz quadrada do erro médio quadrático (MSE) entre as previsões e os valores reais. O valor do RMSE está na mesma unidade de medida da variável resposta, o que faz a interpretação do erro ser direta. Essa métrica de erro atribui um erro maior para previsões com altos valores de erro.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

O RMSE é uma medida de performance do modelo na mesma escala da variável alvo. O RMSE pode ser interpretado como o erro médio que as previsões do modelo tem com os dados reais, sendo que o erro está na mesma escala.

Um valor de RMSE de R$1.000 para a previsão do preço de venda de uma casa parece bom, uma vez que os preços das casas tendem a ser maior que R$100.000. No entanto, o mesmo RMSE de R$1.000 para a previsão de vendas de um computador entre R$800 e R$5.000 seria terrível.

**Vantagens**

1. Atribui pesos maiores para grandes erros.
2. Apresenta a mesma unidade de medida da variável resposta.

**Desvantagens**

1. Não é robusto na presença de outliers.