Módulo 5

Inteligência Artificial

Aula 5

Agenda

- **Introdução**
- **Processos Estocásticos**
- **Exemplos de Processos Estocásticos com Python**
- **Métricas de Desempenho: Matriz de Confusão e Exemplo em Python**
- **Outras Métricas de Desempenho**

Introdução
Determinístico x Estocástico

- Determinístico

  Um processo determinístico se caracteriza por não apresentar nenhum tipo de aleatoriedade. Ao se conhecer as entradas é possível saber exatamente qual a saída. Independe do número de vezes que o processo ocorre a saída sempre será a mesma.

  Exemplo: Converter Celsius para Kelvin.

- Estocástico

  Um processo estocástica é caracterizado a partir de eventos que envolvem aleatoriedade ou incerteza, **apresentando uma ou mais variáveis aleatórias**.

  Exemplo: Jogar uma Moeda.

Introdução
Determinístico x Estocástico

- Calcular a área de um apartamento?

  Estocástico ☐ Determinístico ☐

Introdução
Determinístico x Estocástico

- Calcular a área de um apartamento?

  Estocástico □ Determinístico ■

- Sortear um número entre 0 - 100?

  Estocástico □ Determinístico □

Introdução
Determinístico x Estocástico

- Calcular a área de um apartamento?
  Estocástico □ Determinístico ■

- Sortear um número entre 0 - 100?
  Estocástico ■ Determinístico □

- Jogar um dado para o alto e cair o número 4 com a face para cima?
  Estocástico □ Determinístico □

Introdução
Determinístico x Estocástico

- Calcular a área de um apartamento?
  Estocástico □ Determinístico ■

- Sortear um número entre 0 - 100?
  Estocástico ■ Determinístico □

- Jogar um dado para o alto e cair o número 4 com a face para cima?
  Estocástico ■ Determinístico □

Introdução
Processos Estocásticos

Todo o Processo que pode ser modelado como um problema de probabilidade é considerado como um **Processo Estocástico**.

Introdução
Processos Estocásticos na Indústria

- Grau de desgaste de uma ferramenta.
- Número de produtos defeituosos em uma linha de montagem.
- Quantidade de unidades de determinado produto vendidas por dia.
- Outros...?

Algoritmos Estocásticos

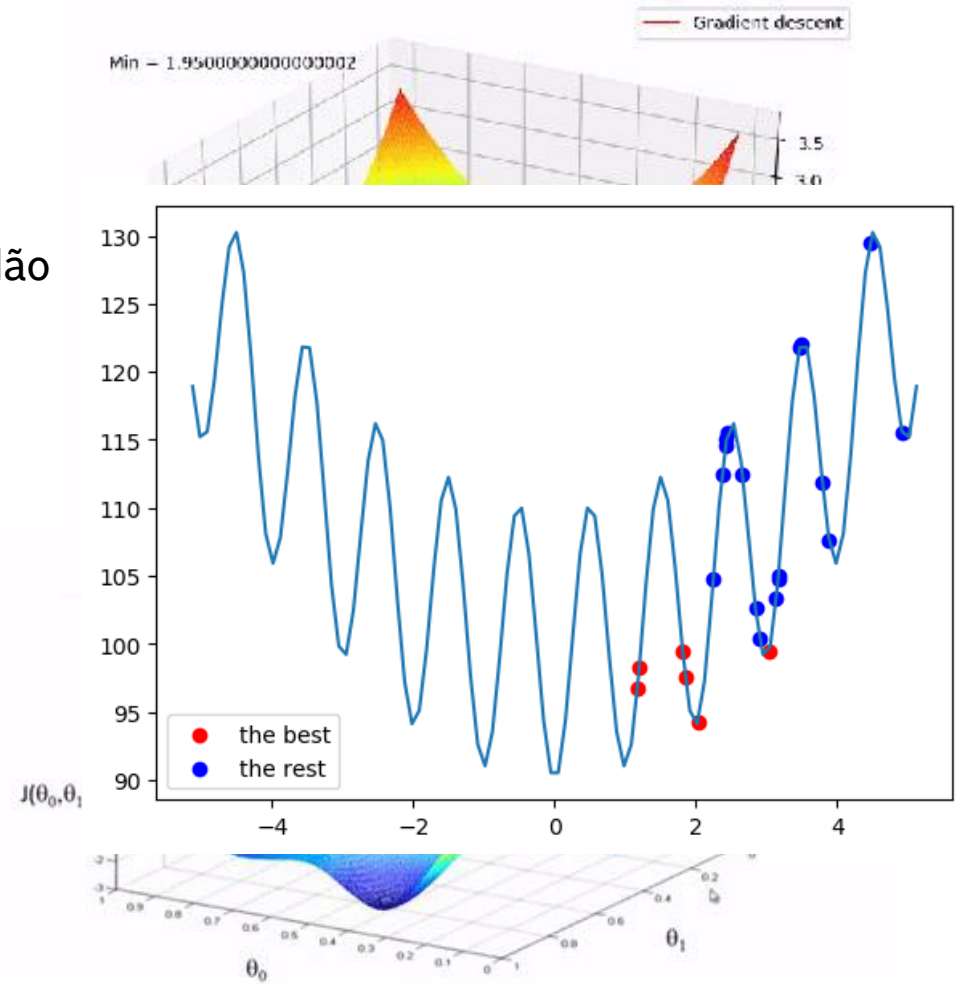
Algoritmos Estocásticos

Definição

- **Algoritmos Estocásticos**, diferente de algoritmos determinísticos, utilizam variáveis aleatórias para obter um resultado (aproximação do resultado). Portanto, mesmo conhecendo suas entradas não é possível dizer com exatidão qual serão suas saídas.
- Ao empregar variáveis aleatórias, temos que os algoritmos estocásticos apresentam maior flexibilidade em várias aplicações, atingindo resultados que os algoritmos determinísticos não conseguiriam ou não seriam viáveis.
- São muito utilizados em problemas complexos de muitas variáveis onde os métodos analíticos são ineficientes ou limitados.

Ex: simulação, otimização, etc...

- Os processos estocásticos geralmente são chamados por métodos de Monte Carlo (MC) ou Monte Carlo Cadeia de Markov (MCMC).



C/IDI1-LA | 2019-07-15

Algoritmos Estocásticos
Monte Carlo (MC) ou Monte Carlo Cadeia de Markov (MCMC)

- Foi proposto em 1946 pelo físico Stanislaw Ulam e desenvolvido em conjunto com o físico-matemático John Von Neumann e outros colegas cientistas durante o projeto Mahattan.
- Inspirado por Ulam numa tentativa de estimar as chances de ganhar no jogo de cartas “Paciência”.
- Foi muito importante para simulações da difusão de nêutrons em ogivas usadas na produção das bombas atômicas desenvolvidas no Projeto Manhattan.
- O **MC** ou **MCMC**, pode ser descrito como método de simulação estatística que utiliza geradores de números aleatórios em cada iteração para desenvolver as simulações.
- Trata-se de um método numérico universal para resolver problemas por meio de amostragem aleatória.

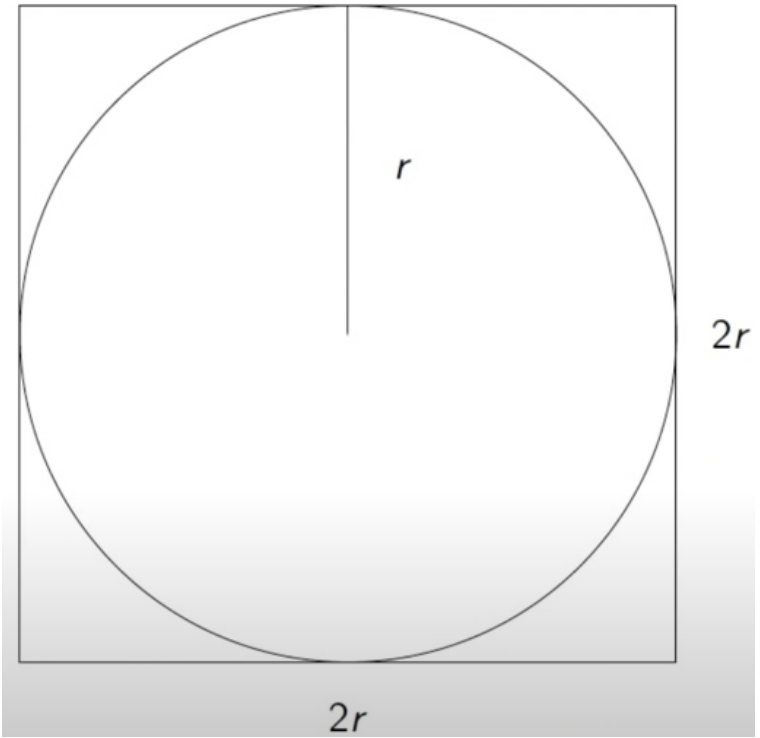
Refs:

http://www2.dbd.puc-rio.br/pergamum/tesesabertas/1012652_2012_cap_3.pdf

http://www.inf.ufpr.br/pfperroni/archive/MonteCarlo.pdf

Algoritmos Estocásticos
Monte Carlo – Exemplo 1: Obtendo o valor de Pi

- Podemos calcular o valor de π utilzando números aleatórios?



$$A_c = \pi r^2$$

$$A_q = 4r^2$$

$$P_{chuva\ c} = \frac{\pi r^2}{4r^2}$$

$$P_{chuva\ c} = \frac{\pi}{4}$$

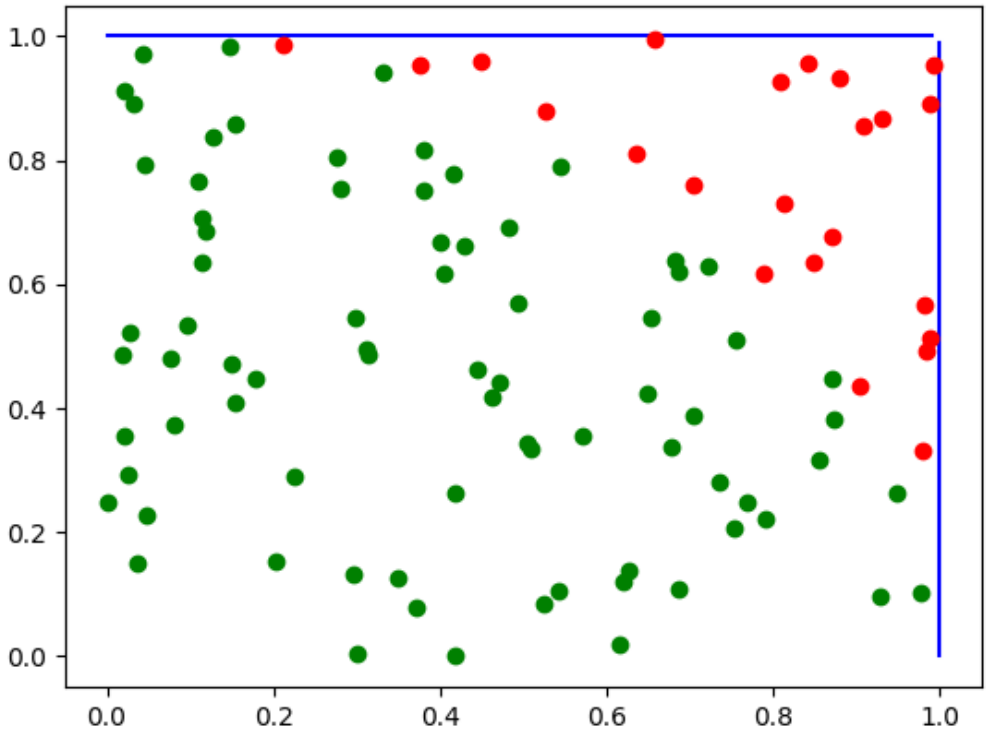
$$\pi = 4 \times P_{chuva\ c}$$

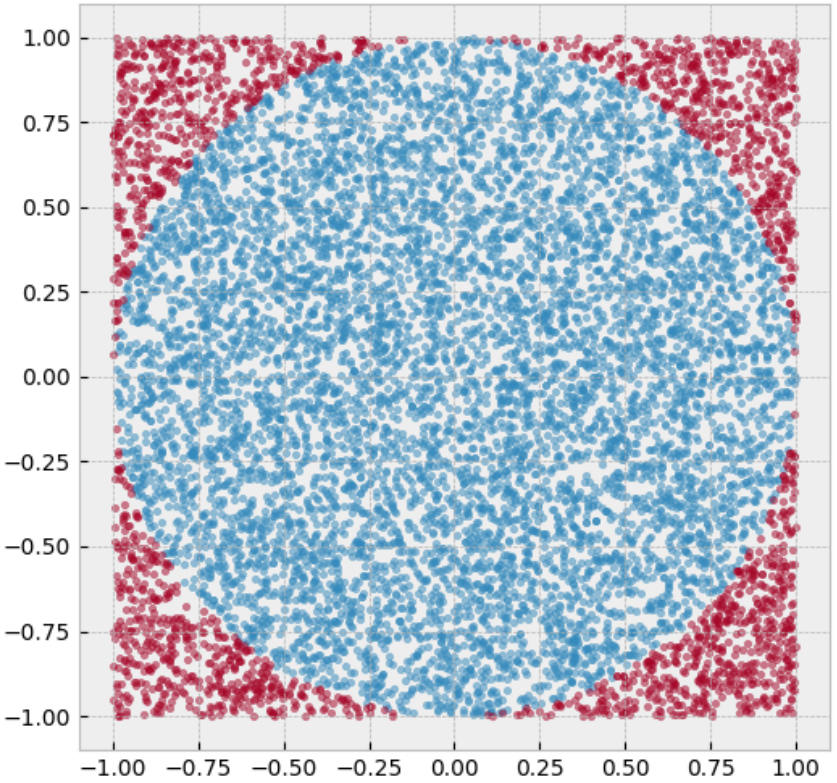
$$\pi = 4 \times \frac{N\ gotas\ círculo}{N\ gotas\ quadrado}$$

Algoritmos Estocásticos

Monte Carlo – Exemplo 1: Obtendo o valor de Pi

Como estimar o valor de π usando números aleatórios?





Algoritmos Estocásticos

Monte Carlo – Exemplo 2: Determinístico vs. Estocástico

Jogue uma moeda 10 vezes. Qual a probabilidade de obter mais que 3 caras?

Resolução Analítica: Distribuição Binomial.

Resolução via Método Estocástico: Monte Carlo.

Métricas de Desempenho

Métricas de Desempenho
Definição

- Consiste no conjunto de métricas utilizadas para avaliar o desempenho de um algoritmo.
- A avaliação de desempenho é uma etapa muito importante, dependendo do problema e do algoritmo utilizado pode ser uma tarefa difícil, e deve ser feita tomando o devido cuidado.
- É recomendado, em problemas mais complexos (a até mesmo nos mais simples), a utilização de mais de uma métrica de desempenho, a fim de ter um resultado mais fiel ao comportamento do algoritmo (ou algoritmos) em questão.
- Métricas diferentes podem representar noções diferentes para o mesmo resultado, a fim de, facilitar o processo de tomada de decisão.

Métricas de Desempenho
Definição

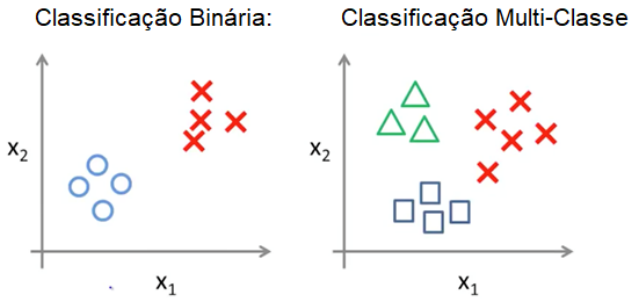
As métricas mais comuns utilizadas na avaliação e comparação de algoritmos:

- Métricas de Erro.
- Custo Computacional.
- Convergência.

Métricas de Desempenho
Métricas de Erro

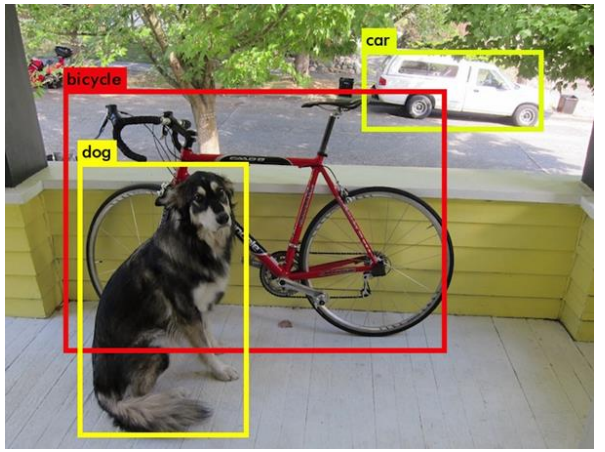
As métricas de erro consistem na avaliação do resultado final do algoritmo em relação a uma dada solução, e a mesma deve ser selecionada de acordo com a característica do problema.

Problemas de Classificação



Problemas de Regressão

Detecção de Objetos



Métricas de Desempenho
Métricas de Erro

As métricas de erro consistem na avaliação do resultado final do algoritmo em relação a uma dada solução, e a mesma deve ser selecionada de acordo com a característica do problema.

Problemas de Classificação

- Matriz de Confusão
- Acurácia
- Precisão
- Recall
- F1

Problemas de Regressão

- R^2
- MSE/RMSE
- MAE

Métricas de Desempenho
Métricas de Erro - Classificação

- **Matriz de Confusão**, é uma matriz que apresenta o número de observações reais e preditas pelo algoritmo.

		Valor Verdadeiro	
		Classe Positiva	Classe Negativa
Valor previsto	Classe Positiva	**VP** Verdadeiro Positivo	**FP** Falso Positivo
	Classe Negativa	**FN** Falso Negativo	**VN** Verdadeiro Negativo

TP-True Positive ou VP-Verdadeiro Positivo: Classe **Positiva** é classificada como **Positiva**.

FP-False Positive ou Falso Positivo: Classe **Negativa** é classificada como **Positiva**.

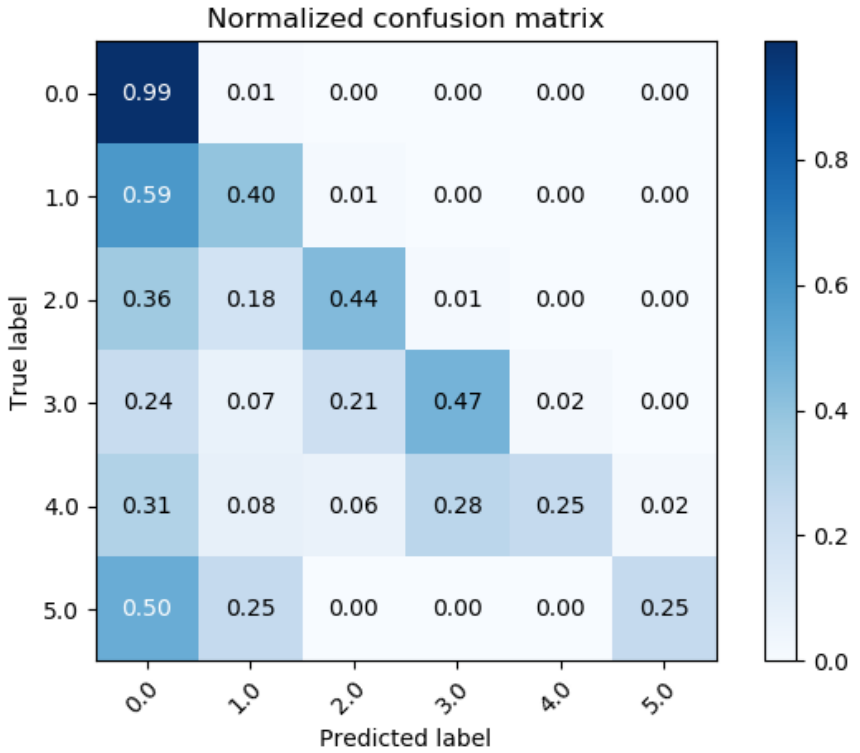
FN-False Negative ou Falso Negativo: Classe **Positiva** é classificada como **Negativa**.

TN-True Negative ou VN-Verdadeiro Negativo: Classe **Negativa** é classificada como **Negativa**.

Métricas de Desempenho
Métricas de Erro - Classificação

- **Matriz de Confusão**, é uma matriz que apresenta o número de observações reais e preditas pelo algoritmo.

N = 600		Previsão	
		Yes	No
Real	Yes	50 (TP)	40 (FN)
	No	10 (FP)	500 (TN)



Métricas de Desempenho
Métricas de Erro - Classificação

- **Acurácia** ou taxa de acerto é a métrica mais intuitiva, se resume à razão do número de observações corretamente preditas sob o número do total de observações. É uma ótima métrica para avaliar modelos nos quais os dados possuem classes balanceadas, ou seja o mesmo número de observações por classe.

$$\text{Acurácia} = \frac{TP + TN}{TP + FP + TN + FN}$$

- **Precisão** é a taxa de observações classificadas corretamente como positivas em relação a todas as observações classificadas como positivas. Uma alta precisão se relaciona com uma baixa taxa de falsos positivos (FP).

$$Precisão = \frac{TP}{TP + FP}$$

Métricas de Desempenho
Métricas de Erro - Classificação

- **Recall** (Sensibilidade) se remete à proporção de observações corretamente classificadas como positivas em relação a todas as observações positivas que poderiam ter sido feitas, ou seja, também levam em consideração observações positivas marcadas como negativas (FN).

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score** é a média harmônica entre **Precisão** e **Recall**. Essa métrica leva em consideração ambos falsos negativos e falsos positivo. É uma métrica mais utilizada que acurácia, principalmente se possui dados com classes desbalanceadas, ou seja, quantidades diferentes de observações por classe.

$$F1 = \frac{2 \ * (Recall * Precisão)}{Recall \ + Precisão}$$

Métricas de Desempenho
Métricas de Erro - Classificação

Em resumo das métricas de erro de classificação:

- **Acurácia** mede a taxa de acerto do modelo, com classes balanceadas.
- **Precisão** indica a proporção de itens da classe classificados corretamente (TP) entre as observações marcadas da mesma classe (TP + FP).
- **Recall** indica a proporção de itens da classe classificados corretamente (TP) em relação a eles mesmos (TP) mais os itens da classe classificados incorretamente (FN).
- **F1-Score** média harmônica de **Precisão** e **Recall**. Leva em consideração FN e FP. Mais utilizado que **Acurácia** para dados com classes desbalanceadas.

Métricas de Desempenho
Exemplo: Matriz de Confusão

Suponha que você está usando um algoritmo para um aprendizado de máquina que precisa prever se pacientes de uma clínica estão infectados ou não por um vírus. Após treinar seu algoritmo com dados de treino, você escolhe 10 valores de teste e monta a tabela a seguir:

a) A partir da tabela, construa a Matriz de Confusão para o problema;

b) Determine os parâmetros a seguir:

- Recall
- Especificidade
- Acurácia
- Precisão
- F1-Score

Predição	Real
Tem Vírus	Tem vírus
Não tem Vírus	Não tem Vírus
Tem Vírus	Tem Vírus
Tem Vírus	Tem Vírus
Não tem Vírus	Tem Vírus
Tem Vírus	Não tem Vírus
Tem Vírus	Tem Vírus
Tem Vírus	Não tem Vírus
Tem Vírus	Tem Vírus
Tem Vírus	Tem Vírus

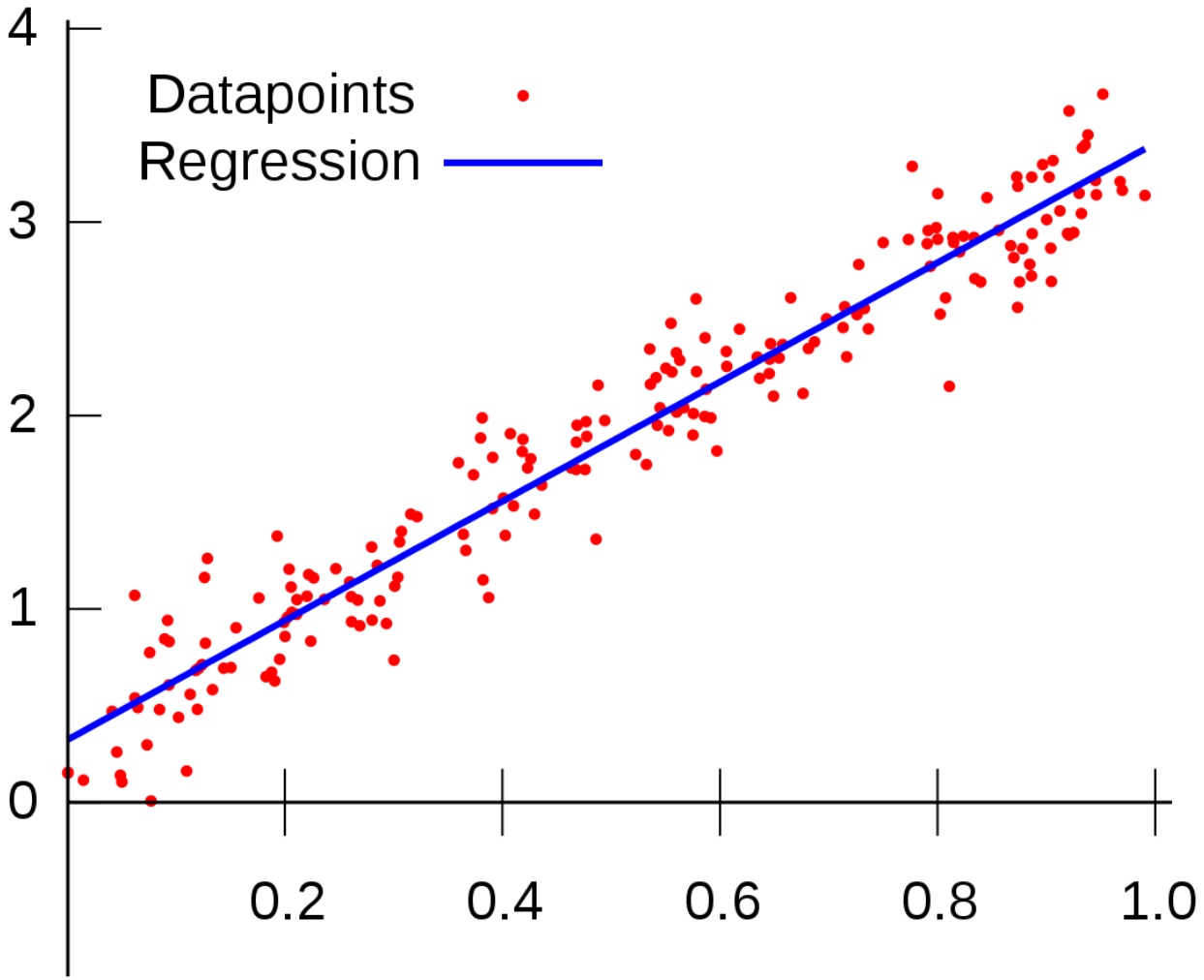
ALGORITMOS DE REGRESSÃO

- O algoritmo de regressão tem como objetivo predizer uma **saída contínua (numérica)** como estimar a temperatura do próximo dia ou predizer o preço de certo produto, baseado nas características envolvidas.
- Os algoritmos de regressão desenham uma **linha** que explica a relação entre a variável dependente (Y) com as variáveis independentes (x1, x2, ..., xn).

$$Y = f(x1, x1, ..., xn)$$

Uma **função f** é aplicada sobre uma ou mais **variáveis independentes Xn** e resulta em determinada **variável dependente Y**.

O objetivo do algoritmo de regressão é encontrar esta função de conversão.

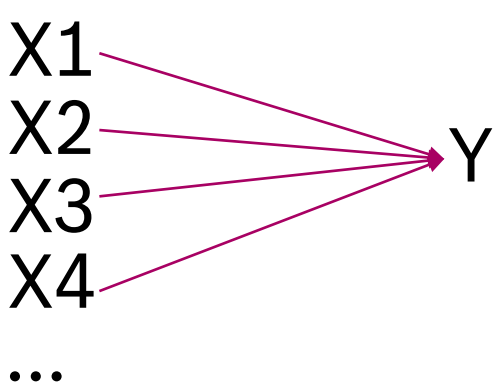


Fonte: https://upload.wikimedia.org/wikipedia/commons/thumb/4/41/LinearRegression.svg/1200px-LinearRegression.svg.png

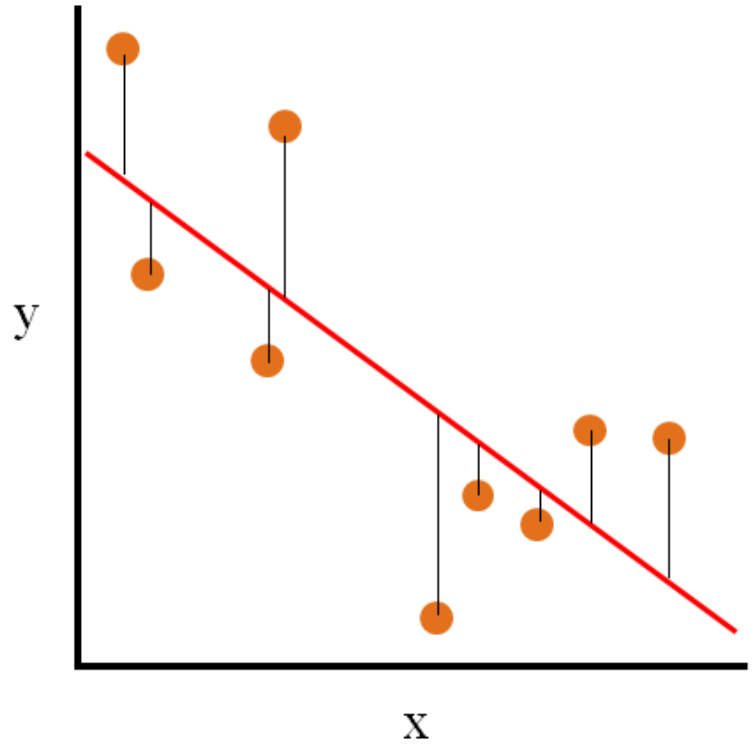
TIPOS DE ALGORITMOS DE REGRESSÃO

- 1. Regressão Simples: Um valor de entrada.

- 2. Regressão Múltipla: Múltiplos valores de entrada.

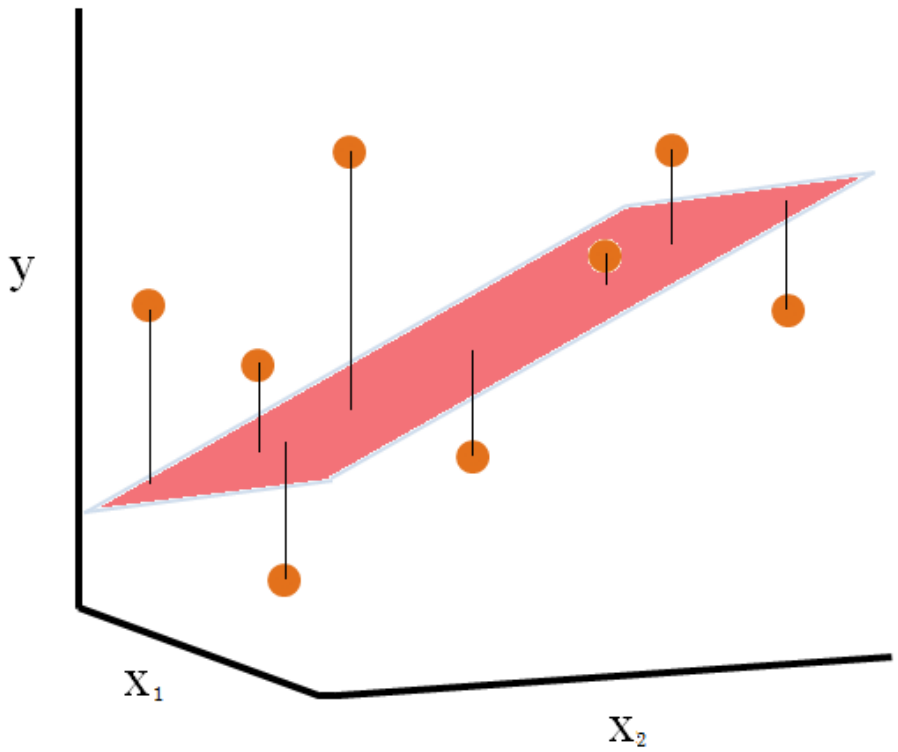


Simple Linear Regression



Multiple Linear Regression

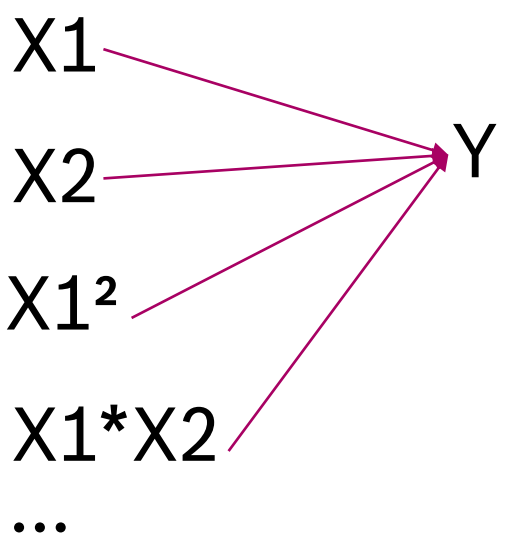
(2 Independent Variables (x_1 , x_2))



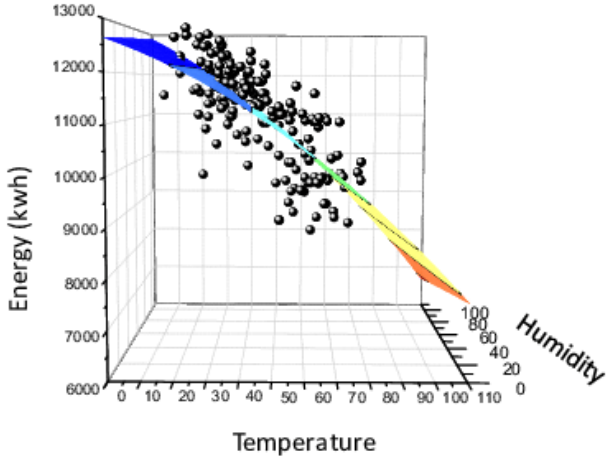
Fonte: https://assets.website-files.com/5e6f9b297ef3941db2593ba1/5f3a434b0444d964f1005ce5_3.1.1.1.1-Linear-Regression.png

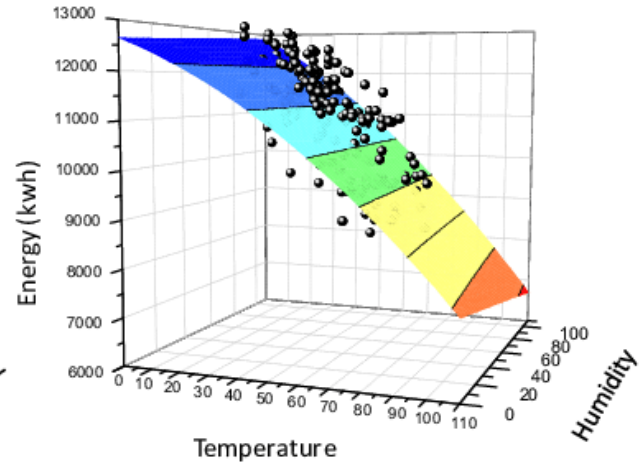
TIPOS DE ALGORITMOS DE REGRESSÃO

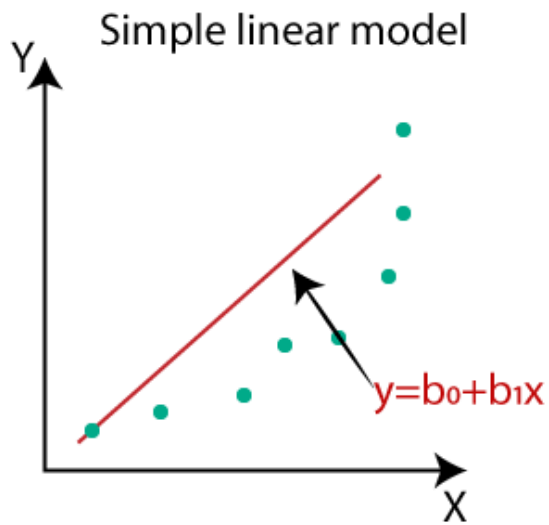
- 3. Regressão Polinomial: Um ou múltiplos valores de entrada, que podem ser combinados para assumir formas mais complexas. Por exemplo, uma equação de resposta pode apresentar termos elevados à ordens de potência menores ou maiores que 1.

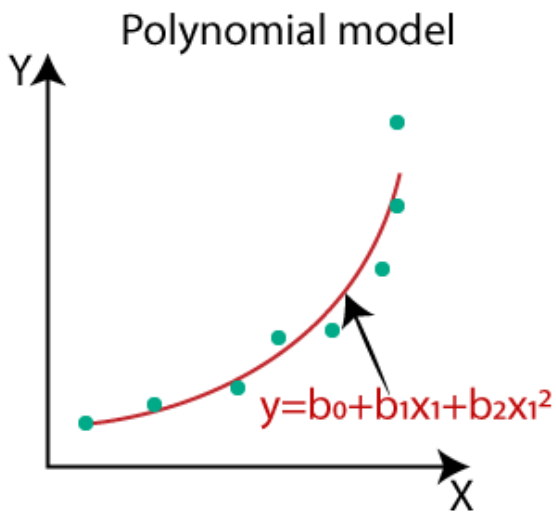


Ilustrações para alguns resultados de Regressão Linear com múltiplas variáveis e Regressão Linear Polinomial









EQUAÇÕES RESULTANTES

O objetivo dos algoritmos de regressão é encontrar os **pesos** (vetor b0...bn) que melhor explicam **a relação** das entradas X para as saídas Y, **reduzindo o erro** de Y_real – Y_previsto.

Regressão Linear Simples

$$y = b_0 + b_1 x_1$$

Regressão Linear Múltipla

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

Regressão Linear Polinomial

$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$$

Métricas de Desempenho
Métricas de Erro - Regressão

- **R²** é chamado de mínimos quadrados e mede o grau de dependência que uma ou mais características do modelo pode(m) ser explicada(s) ou relacionada(s) com outra(s) característica(s) do sistema em estudo. É a medida de ajuste de um modelo estatístico variando de 0 a 1 (0% a 100%).

$$R^2 = \sum_i \frac{(y_i - \bar{y})^2 - (y_i - f_i)^2}{(y_i - \bar{y})^2} = \frac{SS(média) - SS(fit)}{SS(média)}$$

onde y_i, $\bar{y}$ e f_i correspondem respectivamente ao valor real da observação, valor médio das observações e valor estimado pela curva de regressão.

SS = Sum of Squres (Soma dos Quadrados)

De forma resumida, R^2 indica quanto o modelo foi capaz de explicar os dados coletados. Quanto mais próximo de 1 (100%) for o coeficiente, melhor é a capacidade do modelo de representar os dados coletados.

Métricas de Desempenho
Métricas de Erro - Regressão

- **MSE** (Mean Squared Error) é o erro quadrático médio, referente à variância entre os valores gerados pela função de ajuste e o valores reais. Em outras palavras, mede quanto os resultados preditos pelo modelo desviam dos dados reais.

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - f_i)^2$$

RMSE (Root Mean Squared Error) é a raiz quadrada do erro quadrático médio, ou seja, é o desvio padrão entre os valores gerados pela função de ajuste e os valores reais. Geralmente usamos o **RMSE** para representar adequadamente os erros com as mesmas unidades de medida que as variáveis originais do sistema em estudo.

$$RMSE = \sqrt{MSE}$$

Métricas de Desempenho
Métricas de Erro - Regressão

- **MAE** (Mean Absolute Error) é o erro absoluto médio, parecido com o MSE, porém ao invés da soma dos quadrados dos erros, o MAE utiliza a soma dos valores absolutos dos erros.

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - f_i|$$

Métricas de Desempenho
Regressão Linear

Em resumo das métricas de erro de regressão:

- O **R^2** é uma métrica mais simples para explicar a variância do modelo, é a única técnica apresentada aqui que leva em consideração **overfitting (super ajuste)**.
- **MSE** e **RMSE** apresentam respectivamente apenas a variância e o desvio padrão do ajuste em relação aos valores reais ou experimentais do problema em estudo.
- **MAE** resulta no módulo da diferença entre os valores preditos pela função de ajuste e os valores reais ou experimentais do problema em estudo.

Métricas de Desempenho Para Regressão Linear
Exemplos

- Exemplos de regressão linear simples, polinomial e regressão linear múltipla com Python observando o comportamento de R^2.

- Suponha que desejamos saber se o número de horas gasto estudando e o número de simulados tiveram efeito na nota que determinado estudante consegue obter nas provas. Para explorar essa relação, podemos aplicar o método de regressão linear com múltiplas variáveis em Python.

Horas de Estudo: `1, 2, 2, 4, 2, 1, 5, 4, 2, 4, 4, 3, 6, 5, 3, 4, 6, 2, 1, 2`

Número de Simulados: `1, 3, 3, 5, 2, 2, 1, 1, 0, 3, 4, 3, 2, 4, 4, 4, 5, 1, 0, 1`

Notas das Provas: `76, 78, 85, 88, 72, 69, 94, 94, 88, 92, 90, 75, 96, 90, 82, 85, 99, 83, 62, 76`

Métricas de Desempenho Para Regressão Linear
Exemplos

- Podemos prever a emissão de CO_2 de um carro com base no volume do motor (em cm^3), mas com a regressão múltipla podemos incluir mais variáveis, como o peso do carro (em kg), para tornar a previsão mais precisa. Considerando a base de dados do arquivo CO2_&_CARROS_1.csv, que reúne informações sobre algumas marcas e modelos de carros, desenvolva um script em Python para realizar uma regressão linear de múltiplas variáveis e verificar a relação entre a emissão de CO_2 com o peso do carro e o volume do motor.

Marca	Mode	Volume_Motor	Peso_Carro	CO2
Toyota	Aygo	1000	790	99
Mitsubishi	Space Star	1200	1160	95
Skoda	Citigo	1000	929	95
Fiat	500	900	865	90
Mini	Cooper	1500	1140	105
VW	Up!	1000	929	105
Skoda	Fabia	1400	1109	90

Adequação ou Correção de Escalas
Padronização de Escalas

- Quando o conjunto de dados possuí valores muito diferentes ente si, ou ainda, quando temos valores com unidades de medidas distintas pra várias grandezas, pode ser difícil realizar uma comparação entre as mesmas.

- A solução para esse problema é dada através de uma adequação ou correção de escala, permitindo que os dados sejam transformados sem perder as características originais dentro de um intervalo que simplifica a comparação entre eles.

- Usaremos o método conhecido como **Padronização de Escalas**. No entanto, existem vários métodos de adequação de escalas que podem ser empregados.

Adequação ou Correção de Escalas
Padronização de Escalas

- Na Padronização de Escalas usamos a equação a seguir:

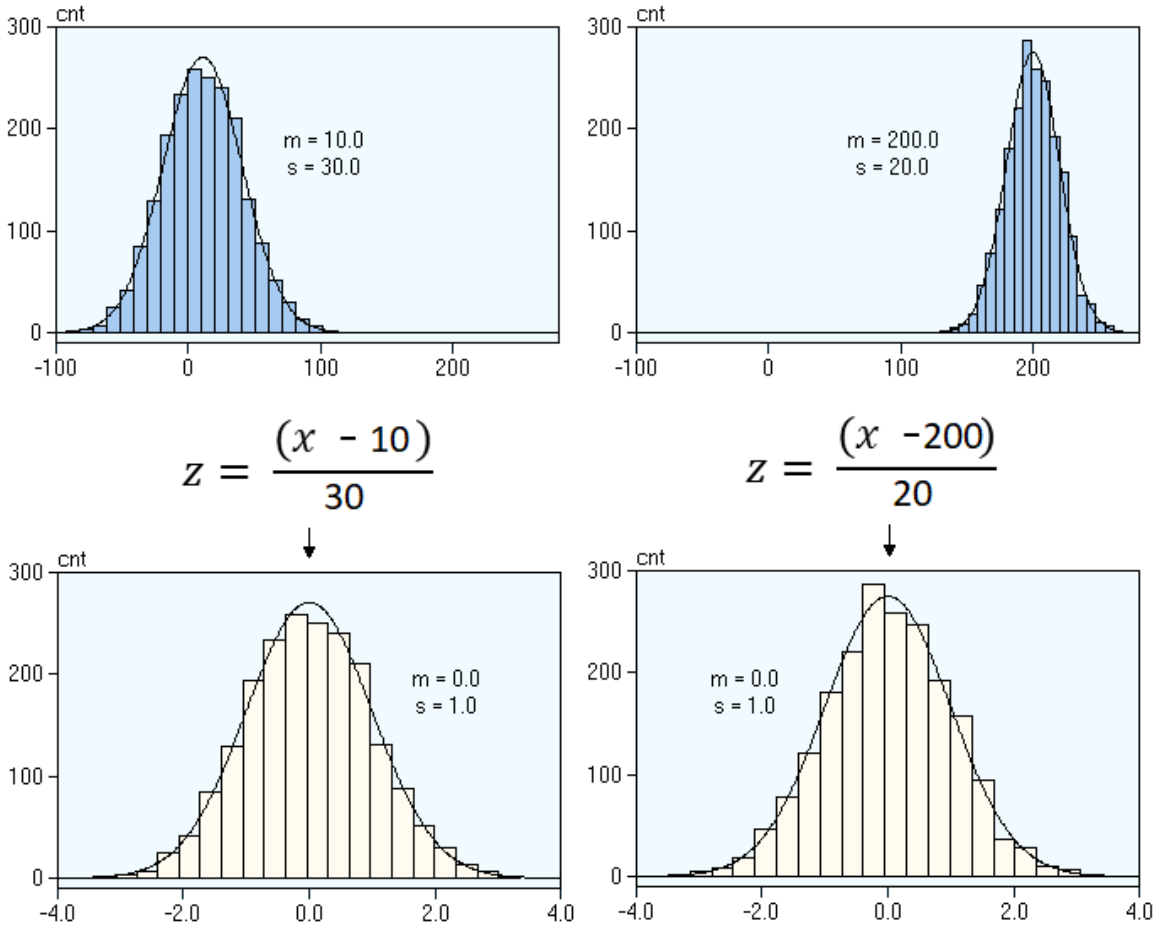
$$z = \frac{(x - \bar{x})}{\sigma}$$

$z = Nova\ Escala$

$x = Valores\ na\ Escala\ Original$

$\bar{x} = Valor\ Médio\ na\ Escala\ Original$

$\sigma = Desvio\ Padrão\ na\ Escala\ Original$



Adequação ou Correção de Escalas:
Exemplo

- A partir do conjunto de dados do arquivo CO2_&_CARROS_2.csv, temos que as escalas de volume são dadas em litros ao invés de centímetros cúbicos. Realize a padronização dos dados usando um script em Python. Usando regressão linear, mostre que a correção de escalas não afeta a relação entre as variáveis preditoras e a variável de resposta.

Marca	Modelo	Volume_Motor	Peso_Carro	CO2
Toyota	Aygo	1	790	99
Mitsubishi	Space Star	1.2	1160	95
Skoda	Citigo	1	929	95
Fiat	500	0.9	865	90
Mini	Cooper	1.5	1140	105
VW	Up!	1	929	105
Skoda	Fabia	1.4	1109	90