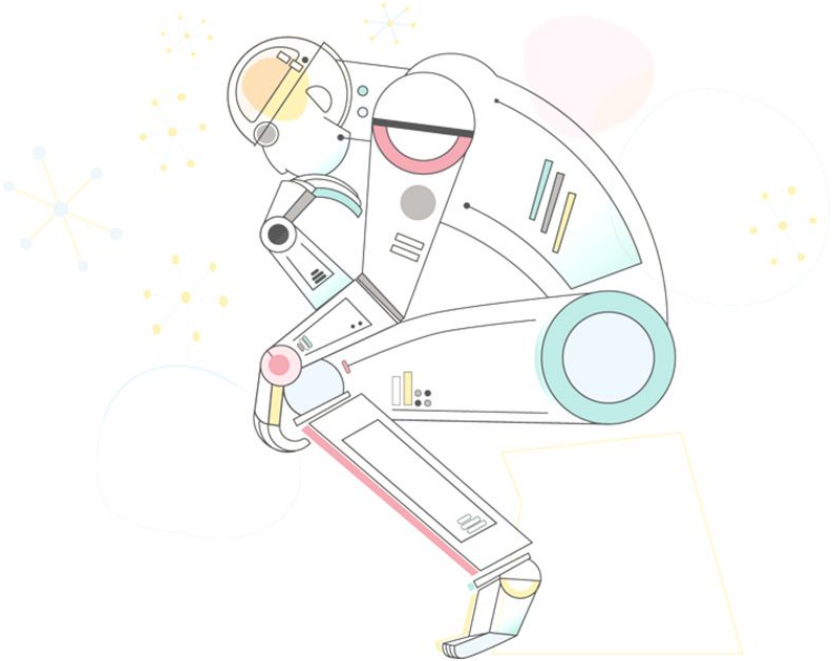
Introdução Machine Learning



Daniela Gomes

Sobre Mim

Daniela Ap. Pires Gomes

Pós Graduada em Big Data

Graduada em Analista De Sistemas

Apaixonada por Tecnologia

Adoro Viajar

O Exterminador do futuro

O Homem Bicentenário

"Machine Learning (ML) é a ciência (e a arte)da programação de computadores para que eles possam aprender com os dados"

Fonte: Introduction to Machine Learning with Python . Autor : Andreas C. Müller, Sarah Guido

“Machine Learning é sobre extrair conhecimento dos dados”

Fonte: Introduction to Machine Learning with Python . Autor : Andreas C. Müller, Sarah Guido

Como extrair conhecimento dos dados?

Utilizando alguns tipos de algoritmos que podem ser classificados como supervisionados e não supervisionados

Aprendizado Supervisionado

- O conjunto de dados já possui um rótulo(label). Portanto, temos o conhecimento da saída correta.
- Os problemas de aprendizado Supervisionado são classificados como Regressão e Classificação
- **Exemplo de algoritmos:**
  - **Regressão Linear**
  - **Regressão Logística**
  - **Árvores de Decisão**
  - **KNN**

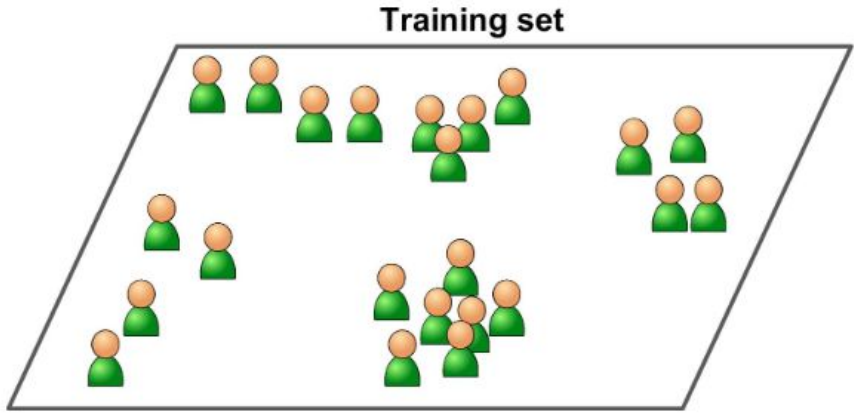
- Ex: Classificação de Spam

Fonte: Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow Autor: Aurélien Géron

Aprendizado Não Supervisionado

- O conjunto de dados já possui ***não*** um rótulo (label). Portanto, temos muito pouco ou nenhum conhecimento de como o resultado vai se apresentar.
- Os problemas de aprendizado Não Supervisionado são Clusterização, regras de associação, visualização e redução de dimensão
- **Exemplo de algoritmos:**
  - **Clusterização : K=means, Cluster hierárquicos**
  - **Regras de Associação: Apriori**
  - **Redução de Dimensão : Análise de componentes principais**

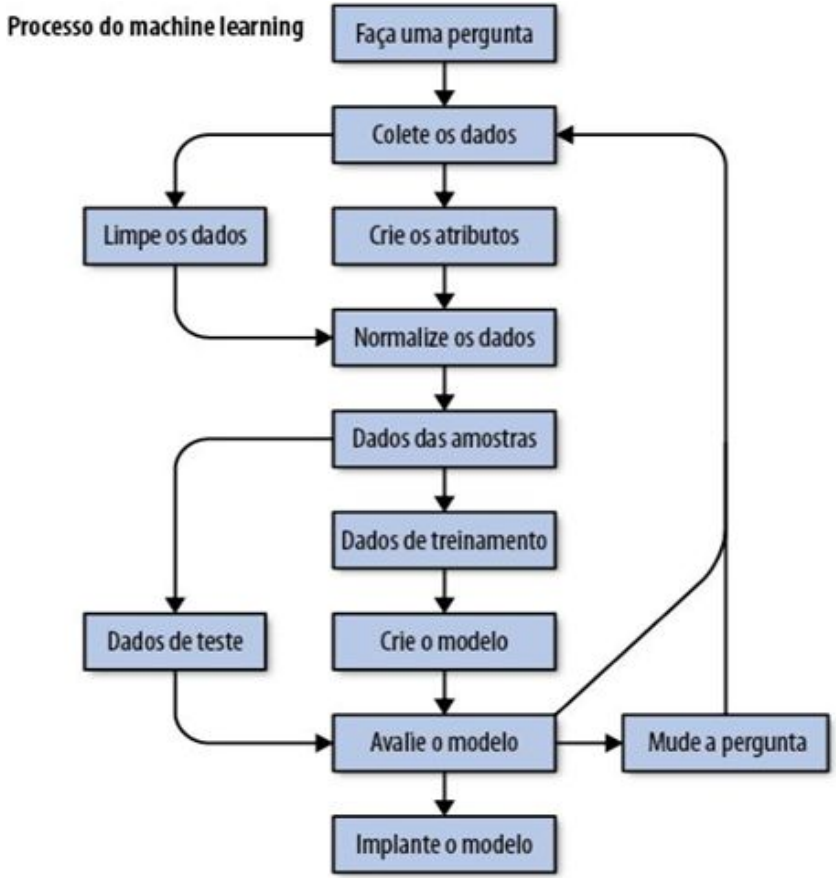
- Ex: Clusterização



Fonte: Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems Autor: Aurelien Geron

Processo de Machine Learning

CRISP-DM é a abreviação de **Cross Industry Standard Process for Data Mining –** Processo genérico de análise de dados que pode ser aplicado em diversos área(financeiras, comerciais, industriais. Além disso é independente de ferramenta.



Fonte: Machine Learning - Guia de Referência Rápida. Autor: Matt Harrison

Análise Exploratória de Dados (AED)

"A análise exploratória de dados busca informações para estabelecer alguma forma de regularidade , padrão ou ainda um modelo presente nas observações"

Fonte: Estatística básica . Autor: Pedro A. Morettin, Wilton de O. Bussab

Tipos de Dados

- Quantitativos
  - Numéricos
    - Discretos — Ex: número de filhos
    - Contínuos — Ex: Altura, Peso, Hora
- Qualitativos
  - Categóricos
    - Nominal — Ex: Fumante/Não Fumante
    - Ordinal — Ex: Nível Escolaridade

Conhecer o tipo de dado ao qual vamos trabalhar é importante pois nos ajuda a determinar o tipo de análise ou modelo que vamos utilizar e o tipo de visualização que usaremos para esse dados

Fonte: Estatística Prática para Cientistas de Dados. Autor Andrew Bruce , Peter Bruce

Medidas de Tendência Central

Moda

- Em um conjunto de dados é o valor que aparece com maior freqüência
- Pode ser classificada com ***Unimodal*** (uma **moda**), bimodal (duas **modas**), multimodal (várias **modas**) e A**modal** (nenhuma **moda**)

Mediana

- Em um conjunto de dados ordenados é o ponto que divide os dados no meio
- Se o número de elementos for par, então a mediana é a média dos dois valores centrais. Soma os dois valores centrais e divide o resultado por 2: **(a + b)/2**.
- Se o número de elementos for ímpar, então a mediana é o valor central.

Média

- Média Aritmética Simples - Soma das observações divida pela quantidade delas

Fonte: Estatística (Coleção Schaum). Autor: Murray R. Spiegel, Larry J. Stephens

www.estudematematica.com.br

O FAMOSO TRIO MMM DA ESTATÍSTICA

Parta sempre de uma lista ORDENADA!

Por exemplo,

4 - 2 - 5 - 2 - 2 - 1 - 2 - 3 - 6 - 5

O ORDENADOR!

1 - 2 - 2 - 2 - 2 - 3 - 4 - 5 - 5 - 6

1) MODA: O valor mais frequente! 2

2) MÉDIA: Razão entre a soma e o número de observações $\frac{32}{10} = 3,2$

PRESUMA MÉDIA ARITMÉTICA

3) MEDIANA: Depende do número de observações!

- Número ÍMPAR? → É quem está BEM NO MEIO!
- Número PAR? → Média dos 2 CENTRAIS! $\frac{2+3}{2} = 2,5$

Medidas de Dispersão

Amplitude	• Em um conjunto de dados é a diferença entre o menor e o maior valor observado no conjunto de dados
Vâriancia	• É a media dos quadrados dos desvios
Desvio Padrão	• Determina o quão distante as observações em um conjunto de dados se encontram da média. • Quanto mais próximo o valor está de zero mais uniforme são os dados

Fonte: Estatística (Coleção Schaum). Autor: Murray R. Spiegel, Larry J. Stephens

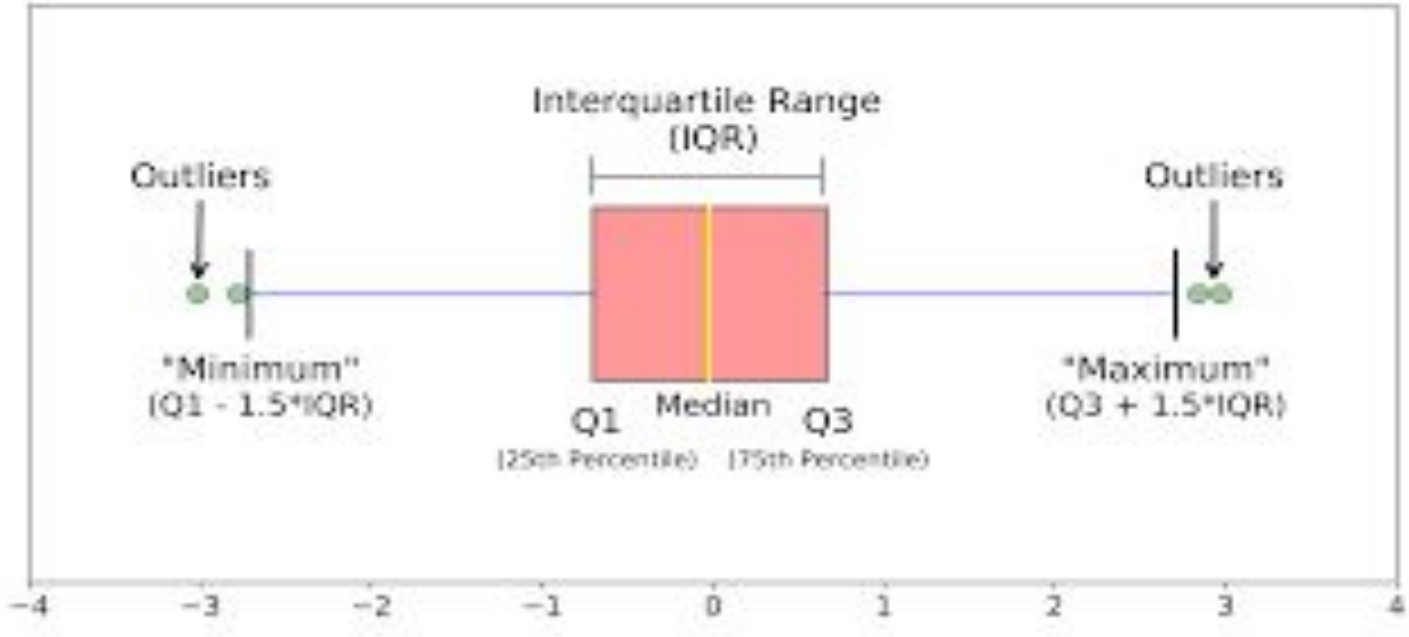
www.estudematematica.com.br

Outliers

- Outlier é qualquer valor que esteja muito distante dos outros valores em um conjunto de dados.
- Um outlier não significa que o dado está errado ou é inválido
- Quando um outlier é resultante de dados ruins , a média aritmética pode ser mal estimada
- De qualquer modo os outliers devem ser identificados, removidos, porém devem ser melhor investigados.
- Uma maneira de identificar outliers e utilizando o Boxplot

Fonte: Estatística básica . Autor: Pedro A. Morettin, Wilton de O. Bussab

Identificando outliers



Fonte: Estatística básica . Autor: Pedro A. Morettin, Wilton de O. Bussab

Desafios Machine Learning

Escassez de especialistas

Qualidade dos dados

Ética

Privacidade

Segurança - Restrições e acesso aos dados

Mudança do Midset Organizacional

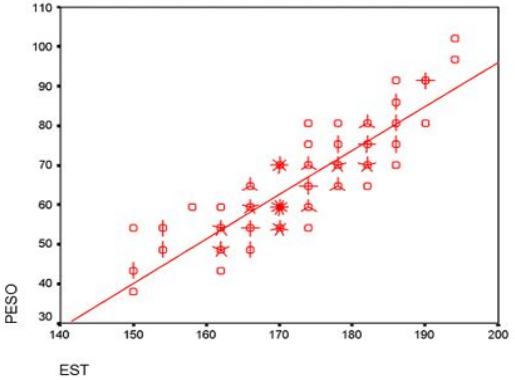
Correlação

O objetivo da Correlação é descrever a associação entre duas variáveis, isto é, desejamos saber o grau de *dependência* entre elas.

Lembrando que correlação não significa causalidade

Correlação - Diagrama de Dispersão

Uma forma de de visualizarmos a correlação de duas variáveis quantitativas é utilizando o diagrama de dispersão.



Coeficiente de Correlação de Pearson

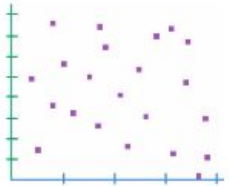
O coeficiente de correlação de Pearson (r) de Pearson mede o grau da correlação linear(forte ou fraca e a direção (positiva ou negativa) entre duas variáveis quantitativas.

Os valores do coeficiente (r) assumem valores entre -1 e 1. Se r= 1 ou r= -1 dizemos que é uma correlação perfeita positiva ou negativa respectivamente. Se r= 0 não existe correlação linear entre as variáveis.

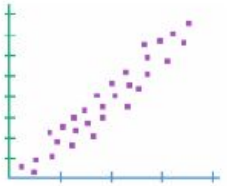
Correlação	Valor de r
Positiva Perfeita	+ 1
Positiva Forte	+ 0,75
Positiva Média	+ 0,50
Positiva Fraca	+ 025
Inexistente	0
Negativa Fraca	- 0,25
Negativa Média	- 0,50
Negativa Forte	- 0,75
Negativa Perfeita	- 1

Morettin e Bussab-Estatística Básica 6 ed

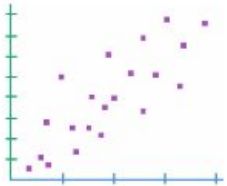
Coeficiente de Correlação de Pearson



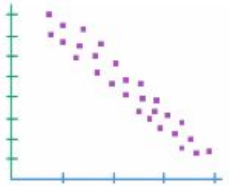
Sem correlação



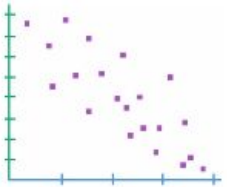
Correlação positiva forte



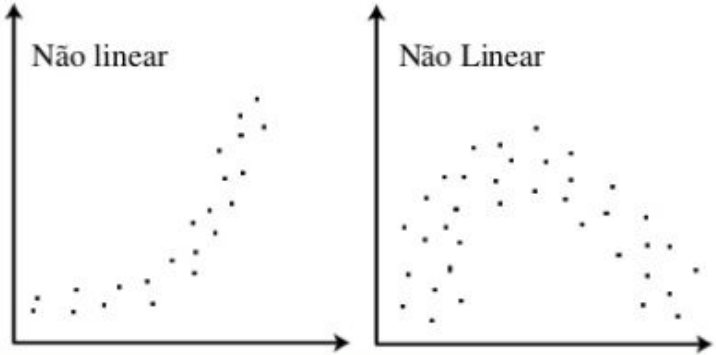
Correlação positiva média



Correlação negativa forte



Correlação negativa média



fonte: https://caetreinamentos.com.br/blog/ferramentas/aprenda-a-montar-um-diagrama-de-dispersao-com-o-excel/

Regressão Linear

A análise de regressão linear gera uma equação matemática que descreva o relacionamento entre uma ou mais variáveis independente e a variável dependente.

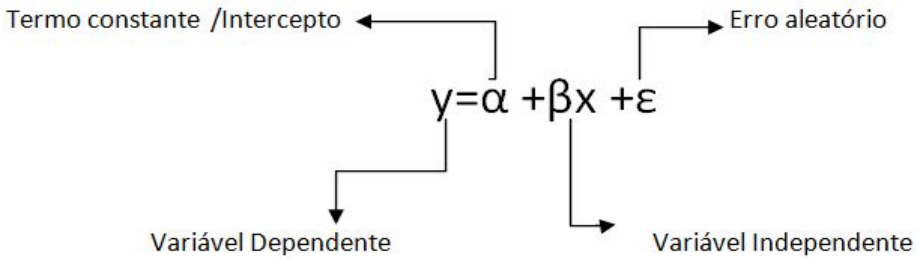
Utiliza dados históricos para prever dados futuros.

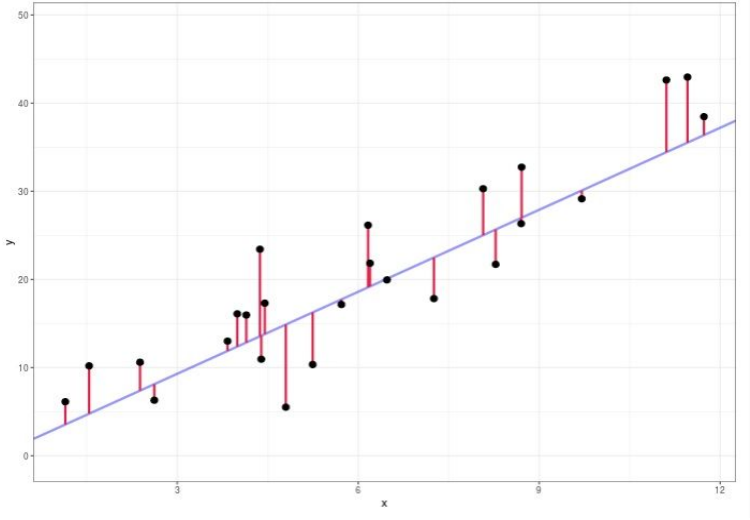
Regressão Linear Simples

Regressão linear simples (MRLS) - define a relação linear entre uma variável dependente com uma variável independente.

Variável dependente - é aquela que está sendo explicada.
Variáveis Independentes - é usada para explicar a variação na variável dependente

Regressão Linear Simples





Y: Variável explicada (dependente); representa o que o modelo tentará prever;

α: É uma constante, que representa a interceptação da reta com o eixo vertical;

β: Representa a inclinação (coeficiente angular) em relação à variável explicativa (independente);

ε :Representa todos os fatores residuais mais os possíveis erros de medição.

MRLS- Método dos Mínimos Quadrados

É uma técnica de otimização matemática que procura encontrar o melhor ajuste para um conjunto de dados tentando minimizar a soma dos quadrados das diferenças entre o valor estimado e os dados observados (tais diferenças são chamadas resíduos)

Método utilizado para obter a reta que melhor se ajusta ao conjunto de dados que estamos estudando

fonte:https://eml.berkeley.edu/sst/regression.html

MRLS- Avaliando o Modelo

- **Teste F de Significância global:** afirma se ao menos uma variável do meu modelo está relacionada com a variável alvo. Para isso, o valor-p desta estatística precisa ser **menor que 0.05**
- **Teste de Significância individuais ou p-values dos coeficientes:** diz o quanto das variáveis preditoras explicam a variável alvo. A métrica padrão é o p-value ser **menor que 0.05**.
- **Coeficiente R^2**: uma medida descritiva da qualidade do ajuste obtido. É um valor entre 0 e 1. Quanto mais próximo de 1, melhor.

fonte:https://medium.com/data-hackers/implementando-regress%C3%A3o-linear-simples-em-python-91df53b920a8l

Regressão Linear - Exemplos de Utilização

Variável dependente	*Variável independente*
Nota da Prova	Tempo de Estudo
Satisfação Cliente	Qualidade do produto, tempo de entrega
Investimento em Marketing	Quantidade de vendas
Salário	Anos de estudo, tempo de experiência
Câncer de pulmão	Ser fumante ou Não

Links Úteis

- http://www.portalaction.com.br/estatistica-basica/estatisticas-descritivashttp://www.portalaction.com.br/estatistica-basica/estatisticas-descritivas
- https://operdata.com.br/blog/desvio-padrao-e-erro-padrao/
- https://pt.khanacademy.org/math/probability/data-distributions-a1/summarizing-spread-distributions/a/introduction-to-standard-deviation

http://ecologia.ib.usp.br/bie5782/doku.php?id=bie5782:02_tutoriais:tutorial2:start

Fonte: Estatística básica . Autor: Pedro A. Morettin, Wilton de O. Bussab

Obrigada!!!

daniela-ap-p-gomes-56580924

daniela.dapg@yahoo.com.br