10 Algoritmos de Aprendizagem de Máquinas (Machine Learning) que você precisa saber

Fonte: <https://towardsdatascience.com/10-machine-learning-algorithms-you-need-to-know-77fb0055fe0>

Vivemos no início da era revolucionada devido ao desenvolvimento de análise de dados, grande poder de computação e computação em nuvem. A aprendizagem de máquinas definitivamente terá um papel importante lá e os cérebros por trás da Aprendizagem de Máquinas são baseados em algoritmos.Este artigo abrange 10 algoritmos de aprendizado de máquinas mais populares que usa atualmente.

Esses algoritmos podem ser categorizados em 3 categorias principais.

1. **Algoritmos supervisionados:** o conjunto de dados de treinamento tem entradas, bem como a saída desejada. Durante a sessão de treinamento, o modelo ajustará suas variáveis ​​para mapear as entradas para a saída correspondente.
2. **Algoritmos não supervisionados:** nesta categoria, não há um resultado alvo. Os algoritmos agruparão o conjunto de dados para diferentes grupos.
3. **Algoritmos de reforço:** esses algoritmos são treinados para tomar decisões. Portanto, com base nessas decisões, o algoritmo irá treinar com base no sucesso / erro de saída. Eventualmente, o algoritmo de experiência poderá dar boas previsões.

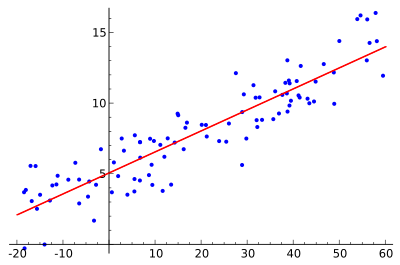
Os seguintes algoritmos serão abordados neste artigo.

* Regressão linear
* SVM (Support Vector Machine)
* KNN (K-vizinhos mais próximos)
* Regressão Logística
* Árvore de decisão
* K-Means
* Floresta aleatória
* Baías ingénuas
* Algoritmos de redução dimensional
* Algoritmos de aumento de gradiente

**1. Regressão linear**

O algoritmo de Regressão linear usará os pontos de dados para encontrar a melhor linha de ajuste para modelar os dados. Uma linha pode ser representada pela equação, **y = m \* x + c** onde **y** é a variável dependente e **x**é a variável independente. As teorias básicas de cálculo são aplicadas para encontrar os valores para **m** e **c** usando o conjunto de dados fornecido.

A Regressão linear tem 2 tipos como **Regressão linear simples,** onde são utilizadas apenas 1 variável independente e **Regressão linear** múltipla, onde múltiplas variáveis ​​independentes são definidas.

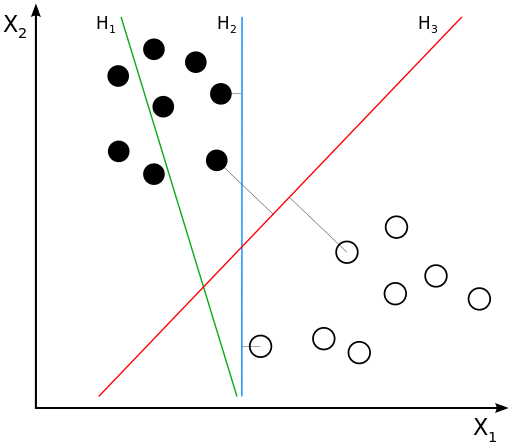


<https://upload.wikimedia.org/wikipedia/commons/thumb/3/3a/Linear_regression.svg/400px-Linear_regression.svg.png>

“ [Scikit-learn](http://scikit-learn.org/" \t "_blank) “ é uma ferramenta simples e eficiente para a aprendizagem de máquinas em python. A seguir, uma implementação da Regressão Linear usando scikit-learn.

**2. SVM (Support Vector Machine)**

Isso pertence ao algoritmo do tipo de classificação. O algoritmo irá separar os pontos de dados usando uma linha. Esta linha é escolhida de tal forma que será mais importante dos pontos de dados mais próximos em 2 categorias.

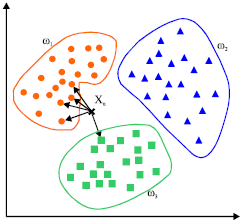


<https://upload.wikimedia.org/wikipedia/commons/thumb/b/b5/Svm_separating_hyperplanes_%28SVG%29.svg/512px-Svm_separating_hyperplanes_%28SVG%29.svg.png>

No diagrama acima, a linha vermelha é a melhor linha, pois tem a maior distância dos pontos mais próximos. Com base nesta linha, os pontos de dados são classificados em 2 grupos.

**3. KNN (K-vizinhos mais próximos)**

Este é um algoritmo simples que prevê pontos de dados desconhecidos com os seus vizinhos mais próximos. O valor de k é um fator crítico aqui quanto à precisão da predição. Ele determina o mais próximo ao calcular a distância usando funções básicas de distância como Euclidean.

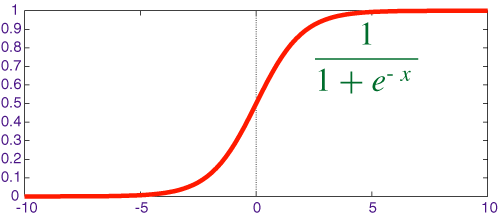


<https://au.mathworks.com/matlabcentral/mlc-downloads/downloads/submissions/63621/versions/1/screenshot.gif>

No entanto, esse algoritmo precisa de alto poder de computação e precisamos normalizar os dados inicialmente para trazer todos os pontos de dados para o mesmo intervalo

**4. Regressão Logística**

A regressão logística é usada onde uma saída discreta é esperada, como a ocorrência de algum evento (ex. Prever se a chuva irá ocorrer ou não).Normalmente, a regressão logística usa alguma função para espremer valores para um determinado intervalo.



Função logistica ( <https://qph.ec.quoracdn.net/main-qimg-05edc1873d0103e36064862a45566dba> )

“Sigmoid” (função logística) é uma das funções que possui a curva de forma “S” utilizada para a classificação binária. Converte valores para o intervalo de 0, 1 que interpretou como uma probabilidade de ocorrer algum evento.

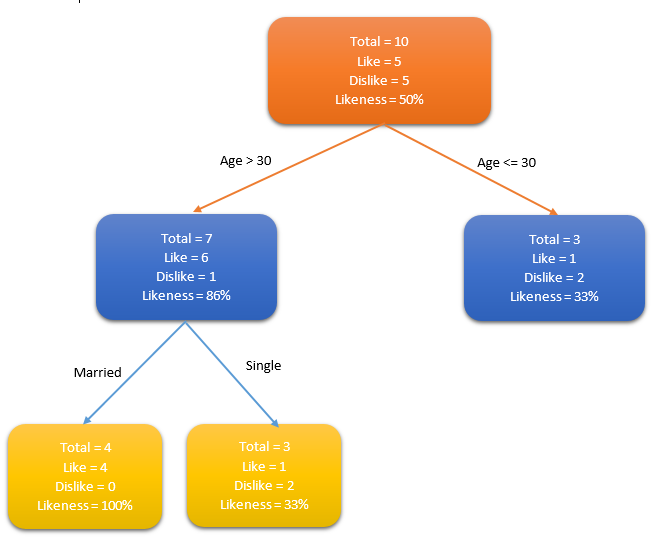
**y = e ^ (b0 + b1 \* x) / (1 + e ^ (b0 + b1 \* x))**

Acima é uma equação de regressão logística simples onde b0, b1 são constantes. Enquanto os valores de treinamento para estes serão calculados de tal forma que o erro entre a predição eo valor real se torne mínimo.

**5. Árvore de decisão**

Este algoritmo classifica a população para vários conjuntos com base em algumas propriedades escolhidas (variáveis ​​independentes) de uma população. Geralmente, esse algoritmo é usado para resolver problemas de classificação. A categorização é feita usando algumas técnicas como Gini, Qui-quadrado, entropia etc.

Vamos considerar uma população de pessoas e usar algoritmo de árvore de decisão para identificar quem gosta de ter um cartão de crédito. Por exemplo, considere a idade eo estado civil como propriedades da população. Se idade> 30 ou uma pessoa é casada, as pessoas tendem a preferir cartões de crédito muito e menos de outra forma.



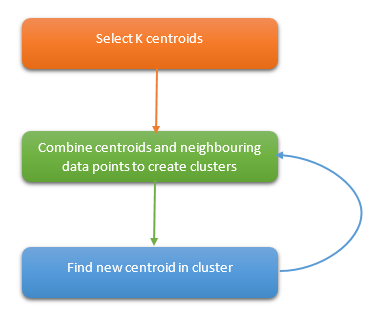
Árvore de decisão simples

Esta árvore de decisão pode ser ampliada por meio da identificação de propriedades adequadas para definir mais categorias. Neste exemplo, se uma pessoa é casada e tem mais de 30 anos, é mais provável que tenham cartões de crédito (100% de preferência). Testar dados é usado para gerar esta árvore de decisão.

**6. Meios K**

Este é um algoritmo sem supervisão que fornece uma solução para o problema de agrupamento. O algoritmo segue um procedimento para formar clusters que contêm pontos de dados homogêneos.

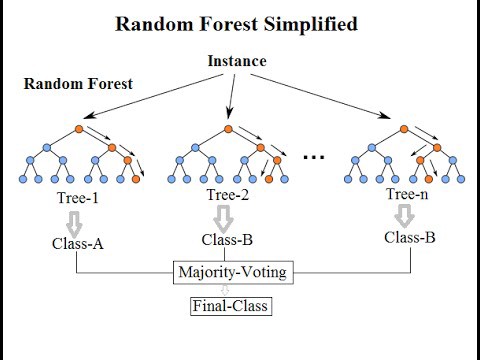
O valor de k é uma entrada para o algoritmo. Com base nisso, o algoritmo seleciona k número de centroides. Em seguida, os pontos de dados vizinhos para um centróide se combinam com o centroide e criam um cluster. Mais tarde, um novo centróide é criado dentro de cada cluster. Em seguida, os pontos de dados próximos ao novo centróide serão combinados novamente para expandir o cluster. Esse processo é continuado até que os centroides não mudem.



Processo de formação de cluster

**7. Floresta aleatória**

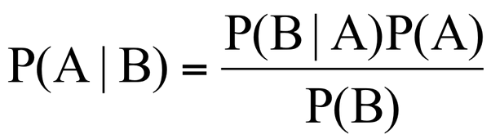
A floresta aleatória pode ser identificada como uma coleção de árvores de decisão como o próprio nome diz. Cada árvore tenta estimar uma classificação e isso é chamado como “voto”. Idealmente, consideramos cada voto de cada árvore e escolhemos a classificação mais votada.



<https://www.kdnuggets.com/wp-content/uploads/rand-forest-1.jpg>

**8. Naive Bayes**

Este algoritmo baseia-se no “Teorema de Bayes” em probabilidade. Devido a isso, Naive Bayes só pode ser aplicado se os recursos forem independentes um do outro, pois é um requisito no teorema de Bayes. Se tentarmos prever um tipo de flor por seu comprimento e largura de pétala, podemos usar a abordagem Naive Bayes uma vez que ambos os recursos são independentes.



Equação de Bayes

O algoritmo Naive Bayes também cai no tipo de classificação. Este algoritmo é usado principalmente quando existem muitas classes no problema.

**9. Algoritmos de redução dimensional**

Alguns conjuntos de dados podem conter muitas variáveis ​​que podem ser muito difíceis de manipular. Especialmente hoje, a coleta de dados em sistemas ocorre em um nível muito detalhado, devido à existência de recursos que sejam suficientes. Nesses casos, os conjuntos de dados podem conter milhares de variáveis ​​e a maioria delas pode ser desnecessária também.

Neste caso, é quase impossível identificar as variáveis ​​que têm o maior impacto em nossa previsão. Algoritmos de redução dimensional são utilizados neste tipo de situações. Ele utiliza outros algoritmos como Random Forest, Decision Tree para identificar as variáveis ​​mais importantes.

**10. Algoritmos de aumento de gradiente**

Gradient Boosting Algorithm usa vários algoritmos fracos para criar um algoritmo preciso mais poderoso. Em vez de usar um único estimador, ter múltiplos criará um algoritmo mais estável e robusto.

Existem vários Algoritmos de Gradient Boosting.

* XGBoost — usa algoritmos de linha e árvore
* LightGBM — usa apenas algoritmos baseados em árvores

A especialidade dos Algoritmos de Gradient Boosting é a sua maior precisão.Além disso, algoritmos como o LightGBM também apresentam alto desempenho incrível.