Algoritmos de Machine Learning

Análise Descritiva - Mostra o passado
Análise Preditiva - Prever o futuro

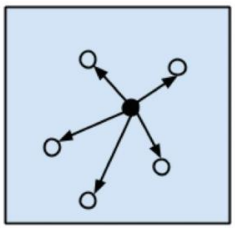
O que é exatamente o aprendizado de máquina? É buscar uma função matemática dentro de um espaço de hipóteses, ou seja encontrar uma função matemática dentre algumas possíveis funções matemáticas.

Modelos

Generativos - Naive Bayes
Discriminativos - Arvores de Decisão, Redes Neurais, KNN

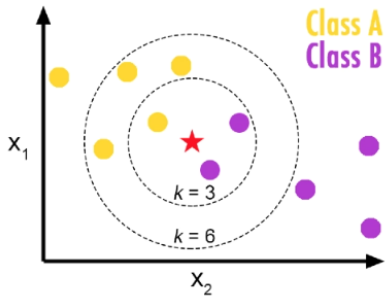
Métodos de Aprendizagem

Métodos Baseados em Instancia



Armazena os exemplos de treinamento e os utiliza quando tiver que classificar um novo dado por comparação direta. A computação ocorre no momento da classificação, ou seja, não ocorre a construção de um modelo de classificação explícito.
Para conjuntos de dados discretos, utiliza como cálculo a distância Euclidiana. Para dados continuos atribui 0 para mesmo valor e 1 para valores diferentes, para esses casos temos que normalizar os dados.

Ex.:



KNN
K – Nearest Neighbours
(Lazy)

K = 1 → 1-NN = seleciona apenas o primeiro elemento mais próximo
K = 5 → 5-NN = seleciona os 5 elementos mais próximos

Métodos Probabilísticos - Baseado no Teorema de Bayes: Assume que a probabililidade de um evento A, que pode ser uma classe, dado em um evento B, poder o conjunto dos valores dos aatributos de entrada, não depender apenas da relação entre A e B, mas também da probablilidade de observar A independemente de B.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

O classificador Naive Bayes é baseado na suposição simplificadora de que os valores dos atriubutos são condicionalmente independentes dado o valor alvo.

Métodos Baseados em Procura - Árvores de Decisão. Ex. Algoritmo ID3 (1986) e C4.5 (1993).

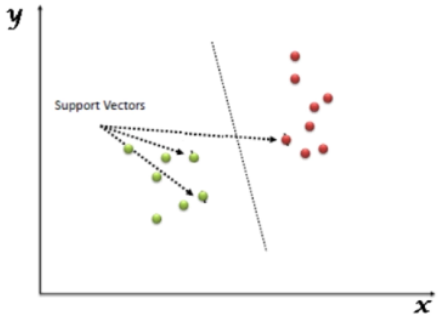
Modelos Baseados por Reforço

- **Modelos Conexionistas**
  - **Perceptrons**
  - **Perceptrons de Multiplas Camadas**
  - **Redes Neurias Convulocionais**
  - **Redes Neurais Recorrentes**
  - **Rede de Kohonen**
  - **Rede de Hopfield**

Métodos baseados em Otimização

Máquinas de Vetores de Suporte (SVM) - Uilizado para classificação binária.

Support Vector Machine



Clustering

- **Normalização**
  - **Min-Max**
  - **Z-score**
  - **Desvio absoluto médio**
- **Grupos**
  - **Hard Clustering**
  - **Soft Clustering -** Atribuido por probabilidade

Principais Algoritmos

Modelos de Conectividade

Modelos Centroide

Modelos de Distribuição

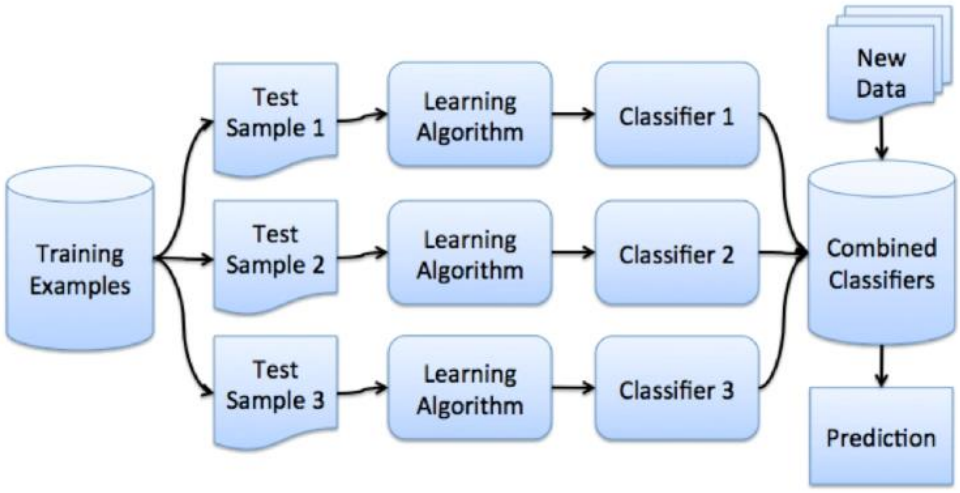
Modelos de Densidade

Métodos Utilizados para Clusterização

Métodos Hierárquicos

Metodos de Particionamento

Métodos Ensemble - Combinação de preditores. Une a saída de vários modelos e pesa através de votos de cada um deles.



Métodos

Os algoritmos seguem duas abordagens principais para criar seu próprio ensemble.

Bootstrap Aggregation ou Bagging

Nessa abordagem treina-se vários modelos, geralmente do mesmo tipo, em diferentes amostras de dados. Principais algoritmos:

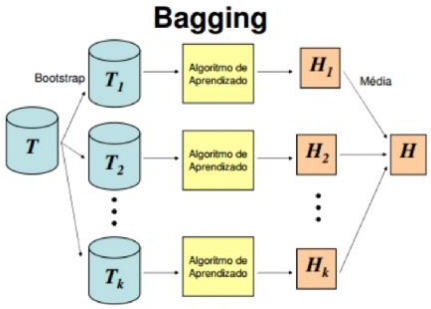
- Bagged CART
- Random Forest

Boosting

Nessa abordagem treinamos vários modelos, geralmente do mesmo tipo, sendo que cada um tenta corrigir o erro do outro. Os principais algoritmos são:

- C5.0
- Stochastic Gradient Boosting
- AdaBoost

Bootsrap Aggregating (Bagging)



Bootstraps (amostras diferentes da base de dados que são usadas para aprender hipóteses diferentes)

Gradiating Boosting
Voting

Redução de Dimensionalidade

Extração de Atributos - Principal Component Analysis, Multidimensional Scaling e o FastMap.
Seleção de Atributos - Algoritmos de Aprendizado de Máquina, cálculo de dimensão fractal e wrapper.

7 Técnicas de Redução de Dimensionalidade

Missing Values Ratio
Low Variance Filter
High Correlation Filter
Random Forests / Ensemble Trees
Forward Feature Construction
Backward Feature Elimination
Principal Component Analysis (PCA) - Inventado em 1901 por Carl Pearson. Precisa ser alimentado com dados normalizados.

Simulação

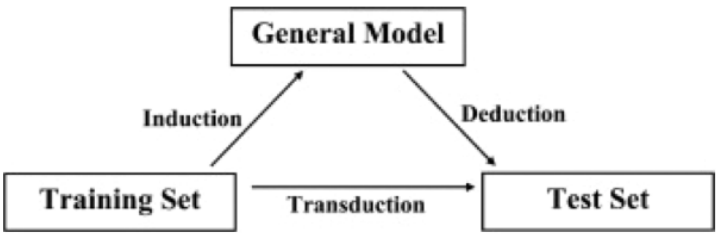
Modelos Determinísticos - Segue uma lei definida pré-determinada em função dos dados de entrada. Ex.: Se uma pessoa tem 16 anos ela pode tirar carteira de motorista. Se tiver menos de 16, não pode.
Modelos Estocásticos - Segue a lei das probabilidades. Ex: Modelo para prever a reação das pessoas em um shopping numa situação de emergência. Um modelo probabilístico tenta descrever o comportamento "aleatório" das entidades. Método Montecarlo.

Aprendizado

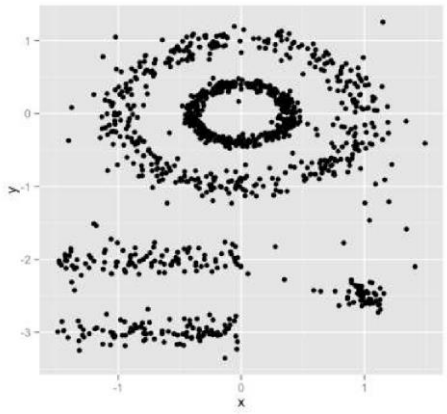
Aprendizado = Representação + Avaliação + Otimização

Processo de Aprendizagem de Máquina

As técnicas de Machine Learning empregam o princípio de inferência denomindo indução, no qual se obtém uma formulação genérica a partir de um grupo particular de exemplos (dados de treinamento).
O aprendizado pode ser divido em supervisionado e não supervisionado.

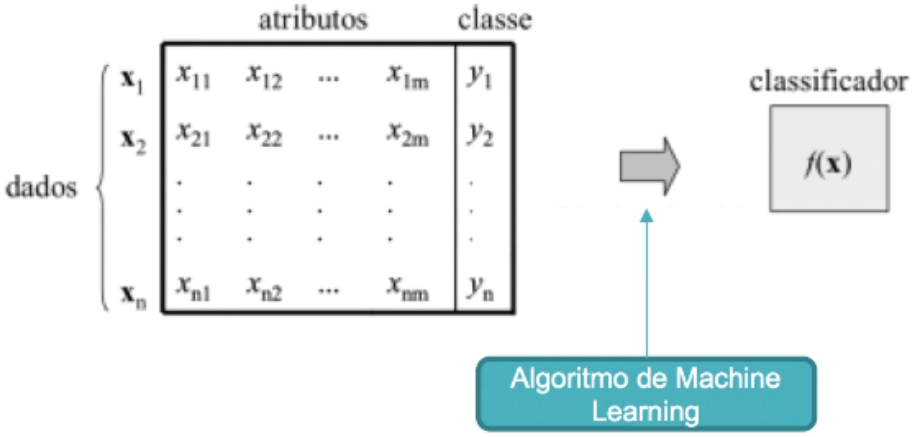


Um requisito importante para as técnicas de Machine Learning é que elas sejam capazes de lidar com dados imperfeitos, denominados ruídos.



Uma maneira de alcaçar esse objetivo é não fixar a obtenção dos classificadores sobre esse tipo de caso específico. Deve-se também aliminar a presença de outlier durante o processo de indução.

Os conceitos referentes na geração de um classificador a partir do aprendizado supervisionado são representados de maneira simplificada na figura abaixo:



Temos nessa figura um conjunto com n dados. Cada dado possui m atributos enquanto que as variáveis y representam as classes, ou seja, as variáveis de saída ou preditiva.

A partir dos exemplos e suas respectivas classes o algoritmo extrai um classificador. Podemos dizer que o modelo gerado fornece uma descrição compacta dos dados fornecidos. A obtenção de um classificador para um algoritmo de ML a partir de uma amostra de dados, também pode ser considerado um processo de busca. Procura-se entre todas as hipóteses que o algoritmo é capaz de gerar a partir dos dados, àquela com maior capacidade de descrever o dominio em que ocorre o aprendizado.

Para estimar a taxa de predições corretas ou incorretas, divide-se o dataset em dois sub conjuntos: treinamento e teste.

Outro importante conceito empregado em ML é a generalização do classificador, definida como a sua capacidade de prever corretamente a classe de novos conjuntos de dados.

Generalization