Engenharia Web
IoT Aula 04
Modelos

Professor Mestre: Adilson Lopes Khouri

10 de novembro de 2019

Modelagem

Cronograma

Aula	Conteúdo
Aula 01	Introdução IoT
Aula 02	Método Científico
Aula 03	EDA
Aula 04	Modelos
Aula 05	Seminários
Aula 06	Seminários
Aula 07	Seminários
Aula 08	Seminários

Amostras viesadas

- Precisamos de informação precisa e sem viés para tomarmos boas decisões.
- Se você “cria conhecimento’ ou “toma decisões” usando informação viesada você não está sendo # datadriven
- A probabilidade de tomar decisões ruins aumenta... e decisões ruins costumam ser caras...

processo de KDD

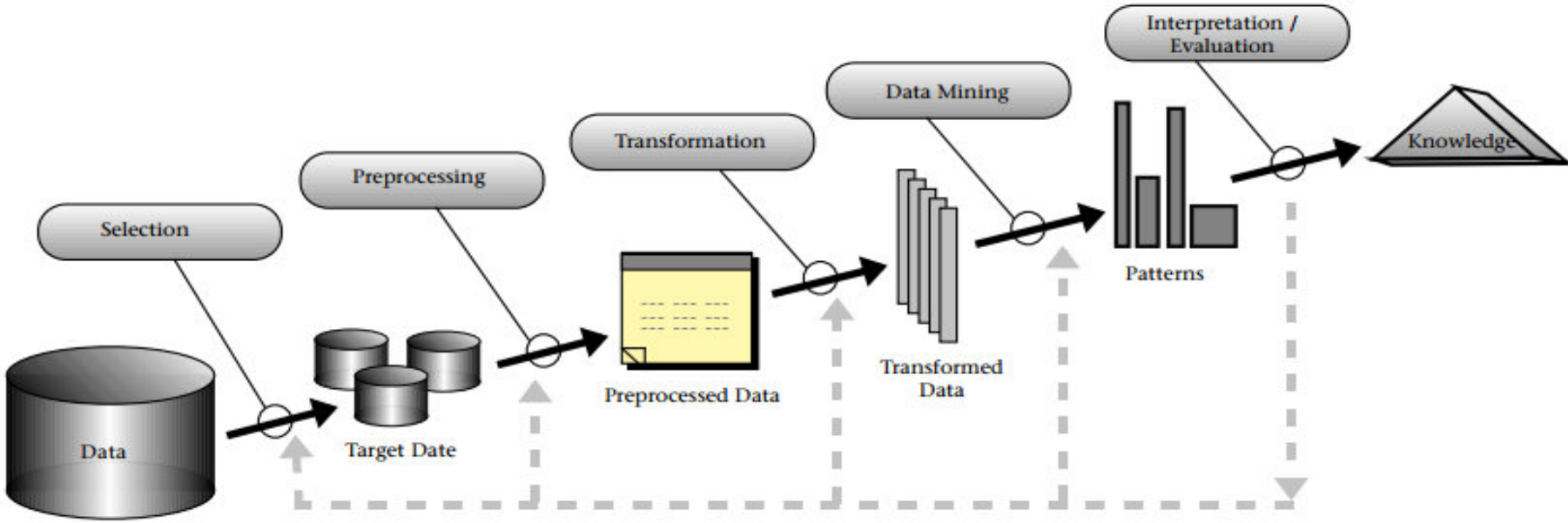


Figura: Processo de KDD

- Se você cometer um erro durante a etapa de: “seleção” os passos seguintes e suas conclusões estarão erradas.

Amostragem 101

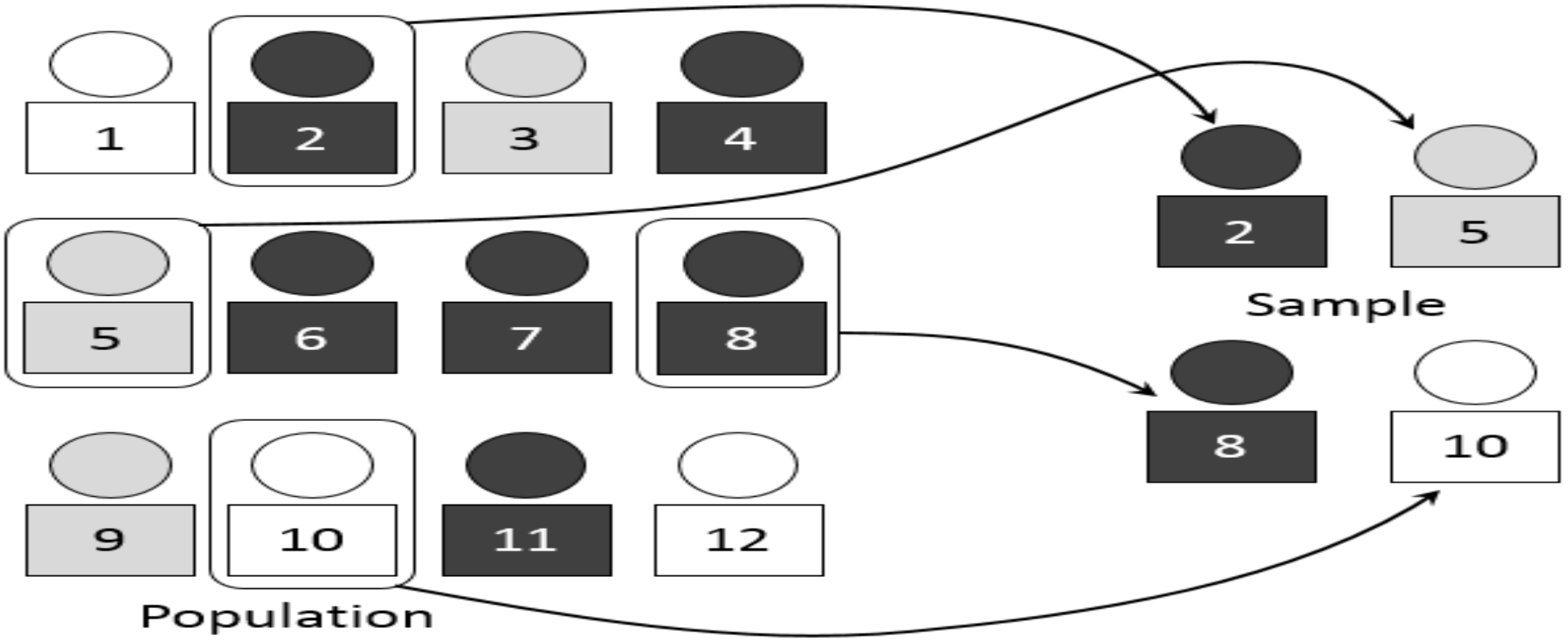


Figura: Overview de amostragem

- O subconjunto (amostra) de elementos deve ser representativa da população.

Bias de auto seleção

- Suponha um estudo estatístico sobre detalhes íntimos da sexualidade de estudantes em universidades. Algumas pesssoas provavelmente vão mentir.
- Uma pesquisa online sobre quem gosta de usar computador.
- Em ambos as pessoas selecionadas vão terão seus comportamentos diferentes da população geral.

Undercoverage Bias

- Digest em 1936 fez uma pesquisa eleitoral que previa vitória larga do candidato Lando em relação ao candidato Roosevelt. Roosevelt ganhou com uma margem larga, a pesquisa era feita por telefone, na época pessoas pobres (maioria da população que era a favor de Roosevelt) não tinha telefone. Essa foi uma das causas do erro estatístico.

Survivorship Bias

- Ocorre quando as observações estudadas no fim da investigação são não aleatórias em comparação as presentes no começo da observação.

Survivorship Bias

- Exemplo da segunda guerra mundial (tiros em avião)

Engenharia de features

- Modelos usam muitas variáveis para tomar decisões
- Encontrar boas variáveis é parte fundamental para um modelo
- Citar exemplo de variáveis de transações financeiras
- Citar exemplo de variáveis de pagamento de assinaturas
- Citar exemplo de um classificador de brasileiros e peruanos

Modelos

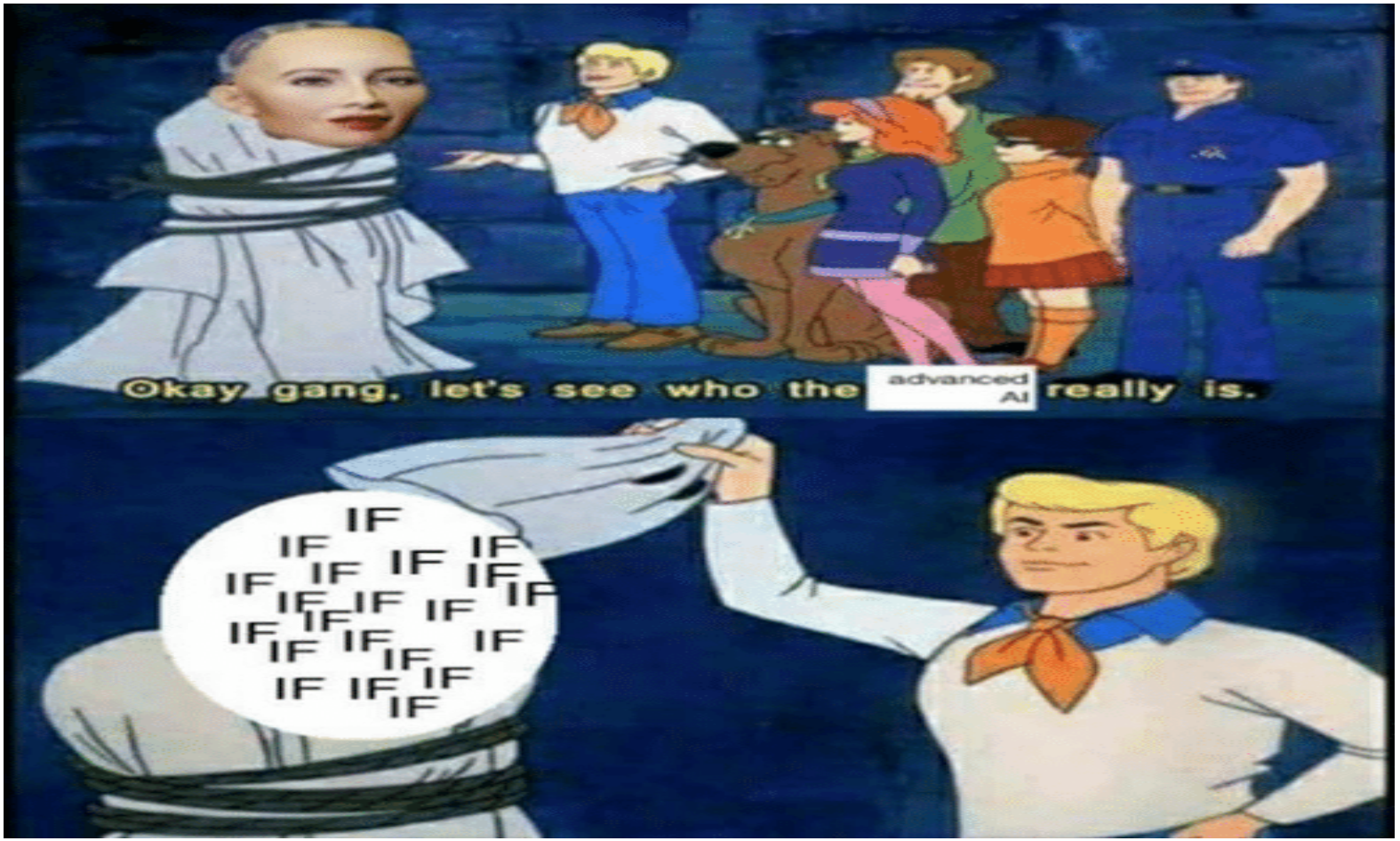


Figura: Brincadeira, cada modelo trabalha internamente de uma forma distinta!

Modelos

- Modelos tomam decisões baseados em diversas variáveis para, entre outras coisas, classificar dados
- Quem são peruanos e quem são brasileiros nessa sala?
- Há modelos para classificar em duas classes ou mais.

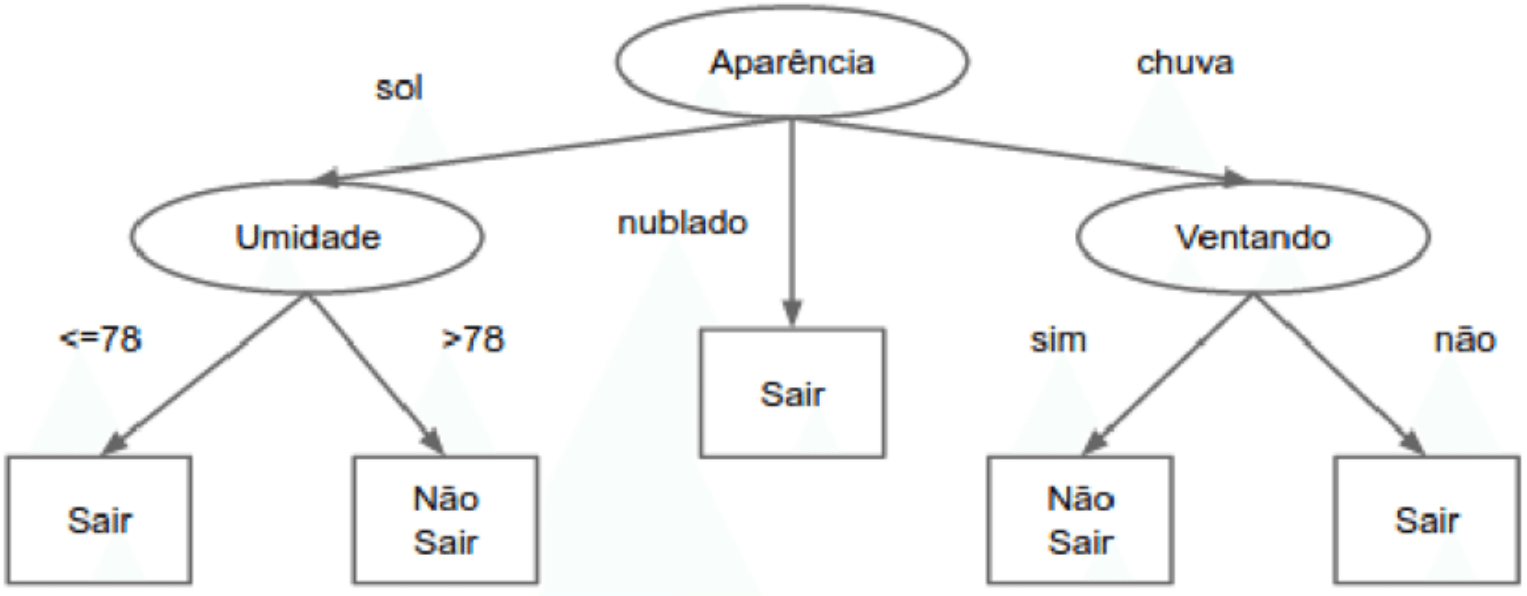


Figura: Exemplo de árvore de decisão para sair de casa

Ferramentas

- Na teoria pode-se usar qualquer linguagem de programação para trabalhar com Data Science
- Na prática usa-se, majoritariamente, a plataforma R e a linguagem python (com alguns pacotes científicos)
- http://scikit-learn.org/stable/ (biblioteca Python)
- https://www.r-bloggers.com (blog de plataforma científica)

Treinamento

- O processo de treinamento é único para cada modelo mas a forma como se treina um modelo é parecida
- Os dados são divididos em treino (70%) e teste (30%)
- O conjunto de treino é apresentado ao modelo com os rótulos de cada observação
- Tipicamente usa-se uma validação cruzada para treinar o modelo

Validação

- O modelo é validado com o conjunto de teste, o qual não deve exibir os rótulos para o modelo
- Alguma métrica de validação de modelos é usada, por exemplo, precisão $\frac{VP}{VP + FP}$

		PREDITO	
		Classe A	**Classe B**
VERDADEIRO	**Classe A**	VP	FN
	Classe B	FP	VN

Figura: Obtido no link: Scielo

Métricas

$$\text{Precision} = \frac{TP}{TP + FP}$$

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

$$\text{F1} = 2\left(\frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}\right)$$

Métricas de uma forma visual

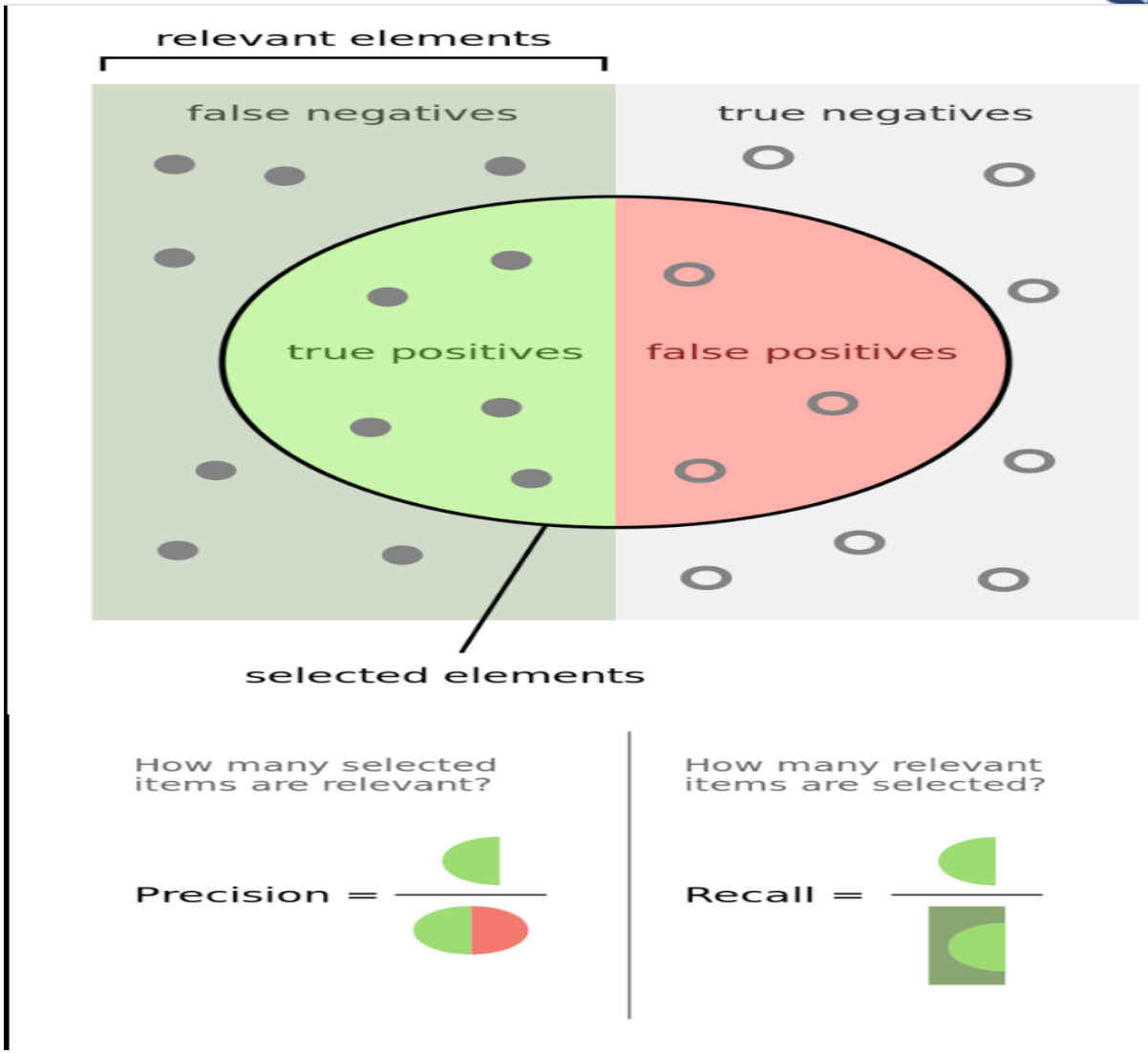


Figura: Métricas de uma forma visual

Gráficos

- Curva ROC (“True Positive Rate vs. False Positive Rate”)
- AuC (Área sobre a curva ROC)

Exemplo ROC e AUC

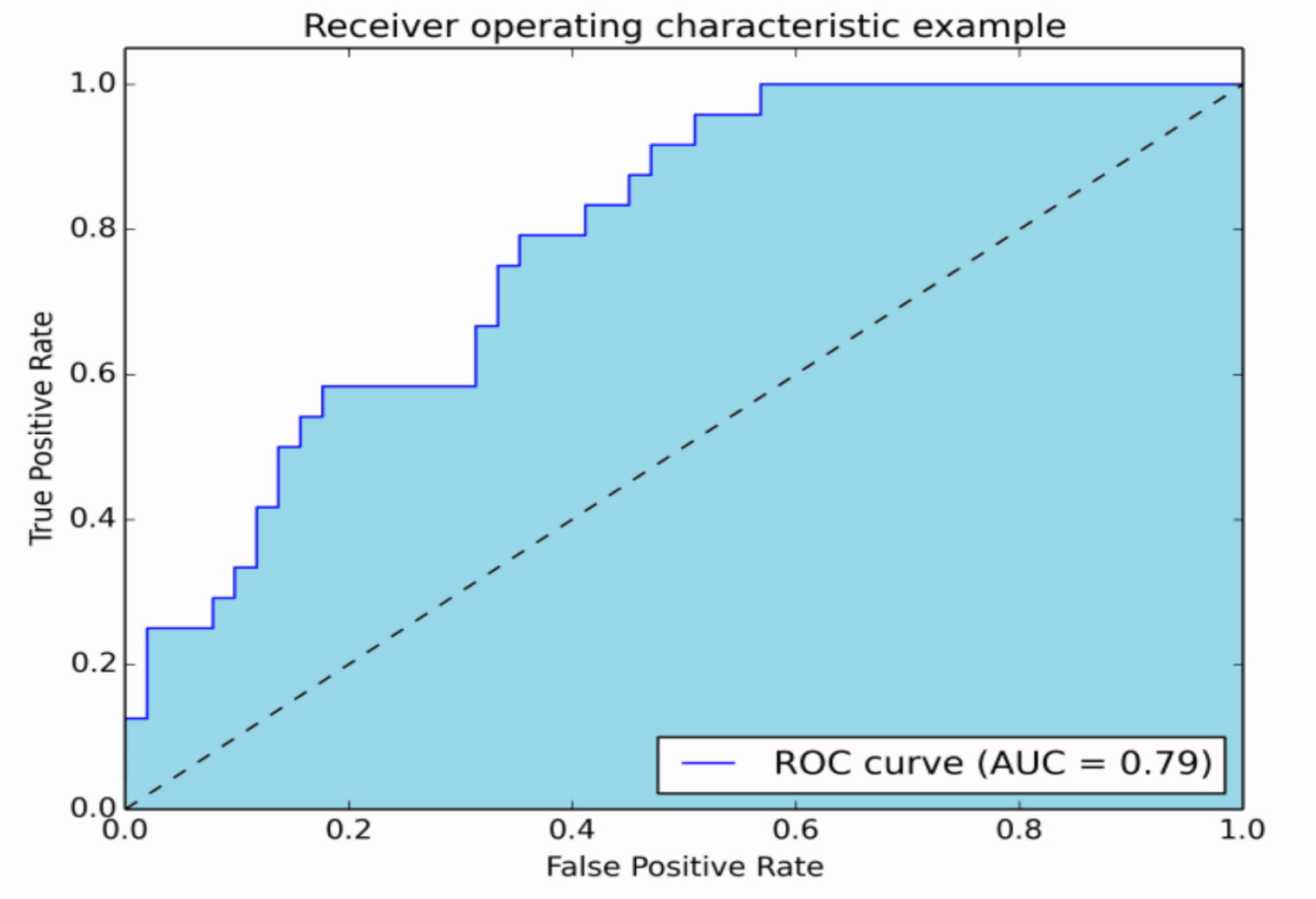
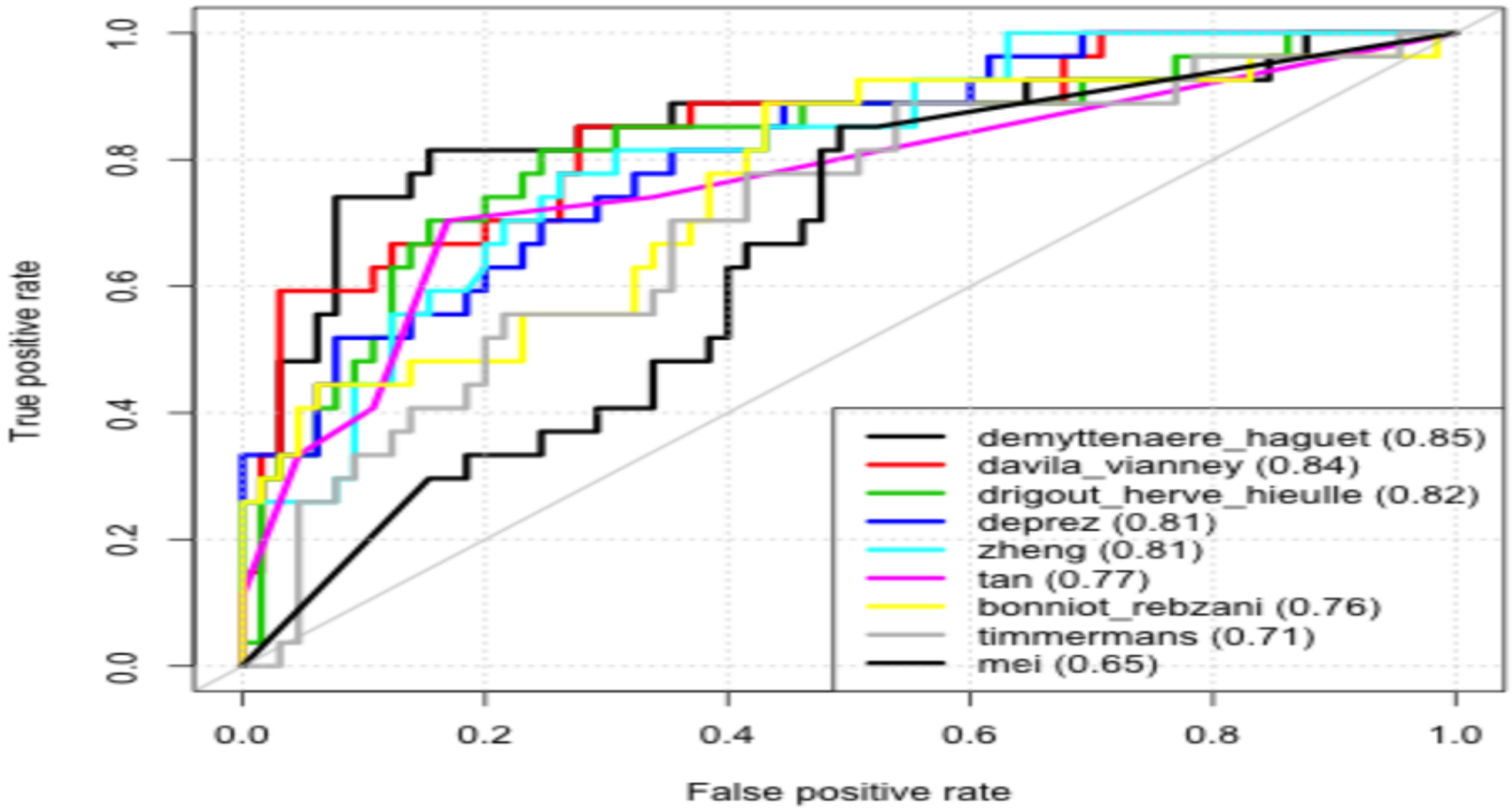


Figura: Exemplo ROC e AUC

Comparação de modelos



Referência da figura.

Figura: Comparação de modelos


Machine Learning Algorithms
Bayesian
Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Gaussian Naive Bayes
Multinomial Naive Bayes
Bayesian Network (BN)
Decision Tree
Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5
C5.0
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
Conditional Decision Trees
M5
Dimensionality Reduction
Principal Component Analysis (PCA)
Partial Least Squares Regression (PLSR
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Principal Component Regression (PCR)
Partial Least Squares Discriminant Analysis
Mixture Discriminant Analysis (MDA)
Quadratic Discriminant Analysis (QDA)
Regularized Discriminant Analysis (RDA)
Flexible Discriminant Analysis (FDA)
Linear Discriminant Analysis (LDA)
Instance Based
k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)
Clustering
k-Means
k-Medians
Expectation Maximization
Hierarchical Clustering
Deep Learning
Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders
Ensemble
Random Forest
Gradient Boosting Machines (GBM)
Boosting
Bootstrapped Aggregation (Bagging)
AdaBoost
Stacked Generalization (Blending)
Gradient Boosted Regression Trees (GBRT)
Neural Networks
Radial Basis Function Network (RBFN)
Perceptron
Back-Propagation
Hopfield Network
Regularization
Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least Angle Regression (LARS)
Rule System
Cubist
One Rule (OneR)
Zero Rule (ZeroR)
Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
Regression
Linear Regression
Ordinary Least Squares Regression (OLSR)
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Logistic Regression

Exemplos de modelos

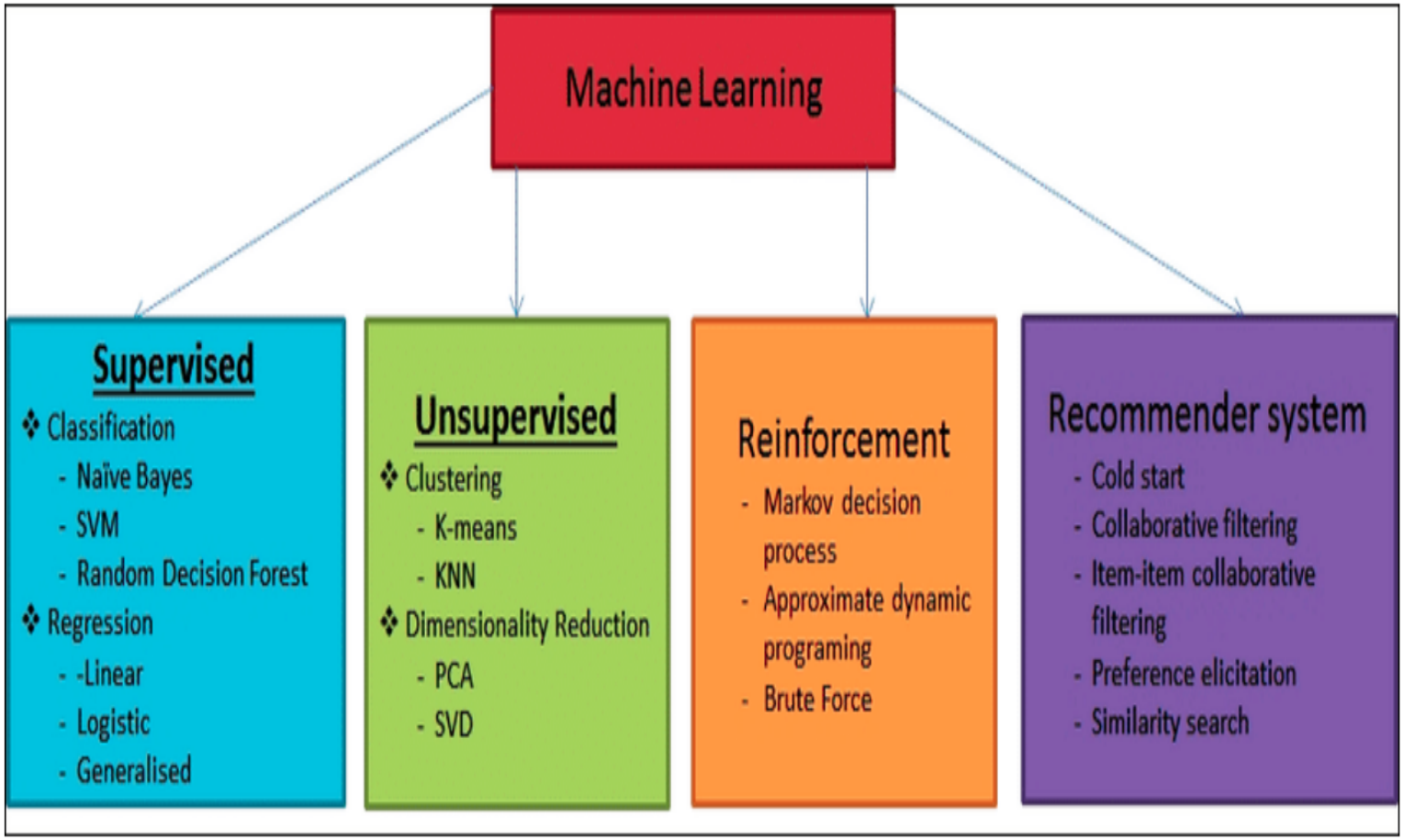


Figura: Exemplos de modelos 2

Exemplos de modelos

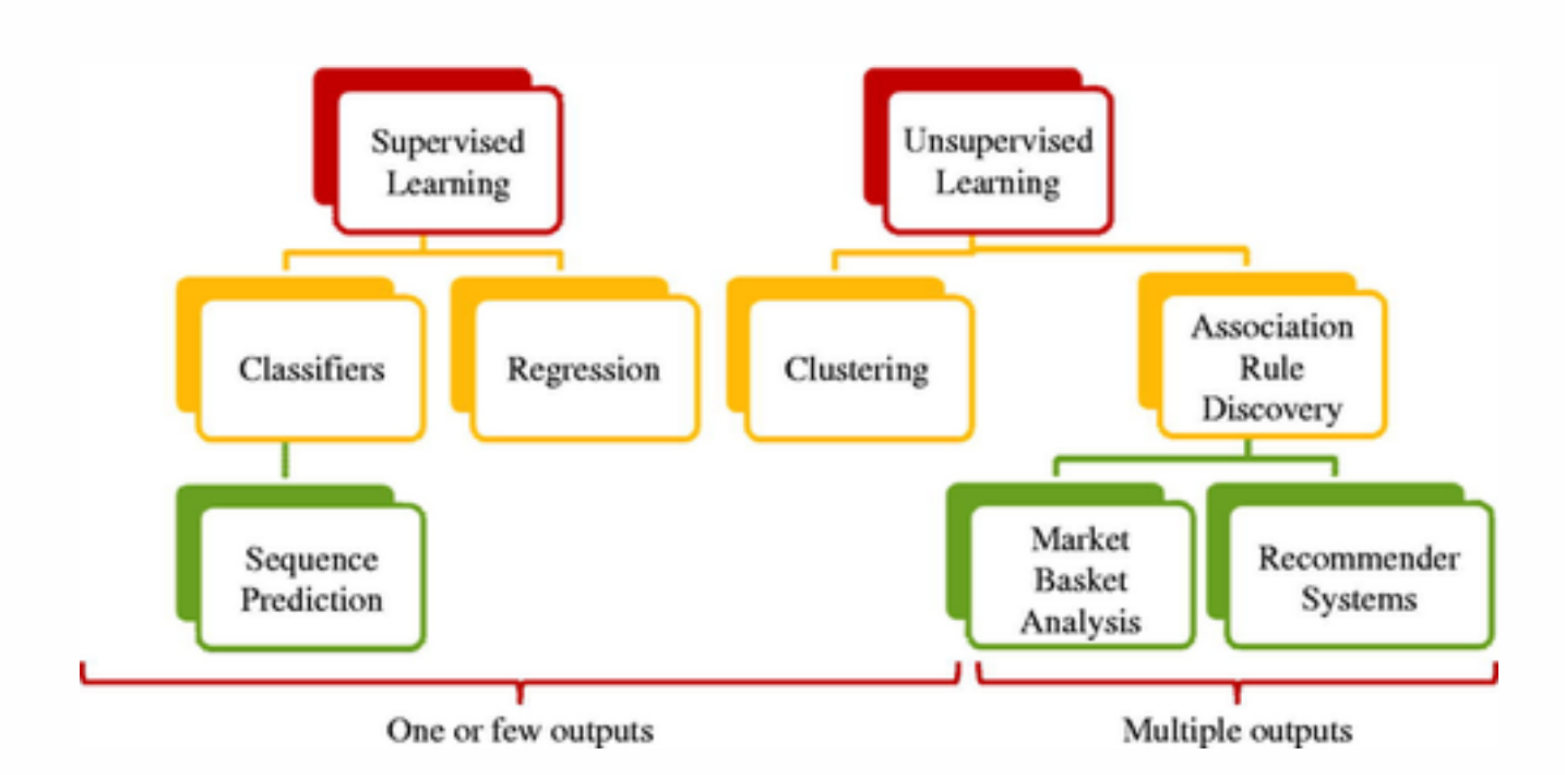


Figura: Exemplos de modelos 3

Modelos baseados em árvore

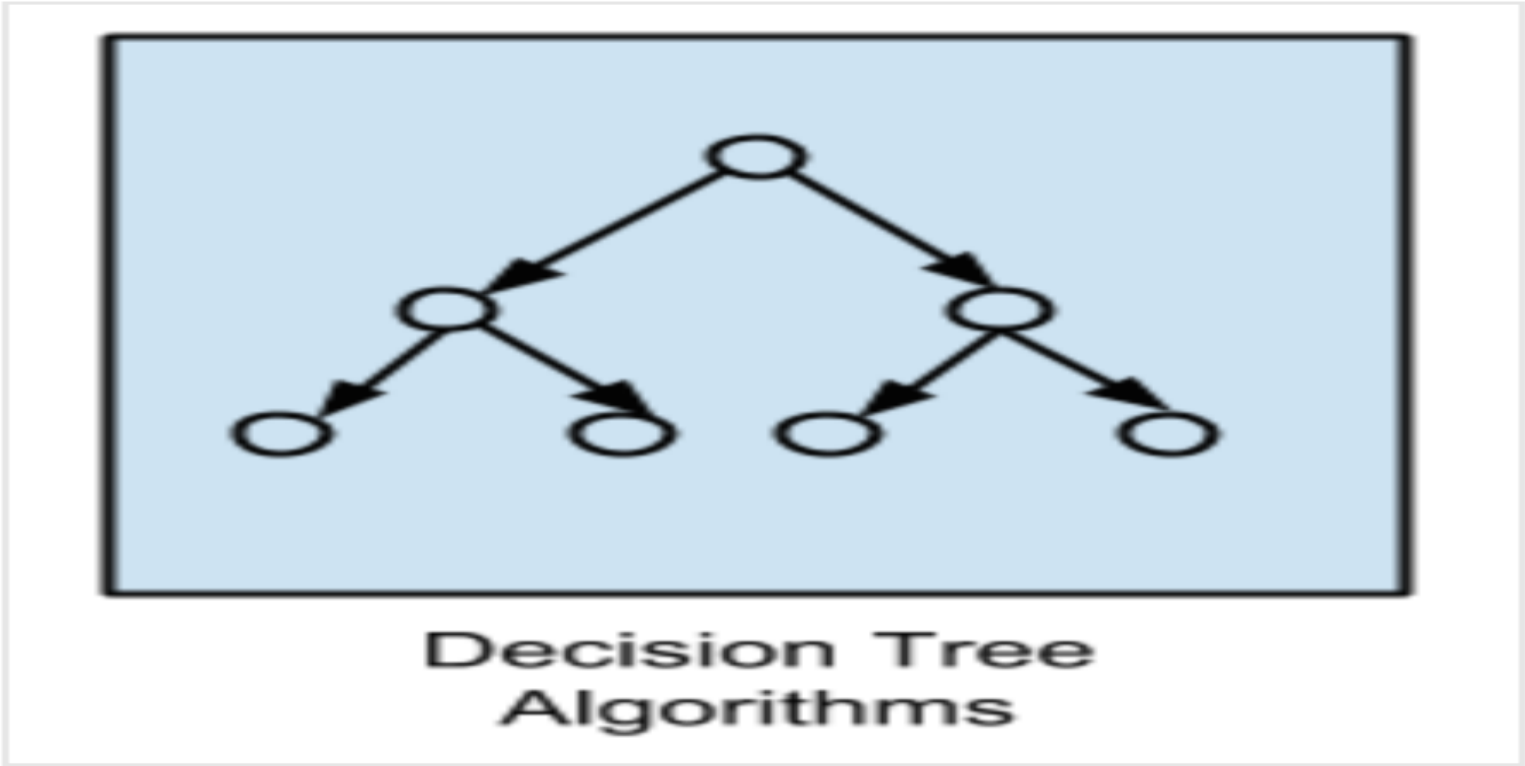


Figura: Modelos baseados em árvore

Sobre esses modelos

- Treina o modelo usando os valores dos dados apresentados
- Cada nós da árvore é uma decisão tomada, encontrar os nós é o trabalho do treinamento
- CART, *C*4.5, *C*5.0, *CHAID* são exemplos de algoritmos baseados em árvore

Modelos baseados em regressão

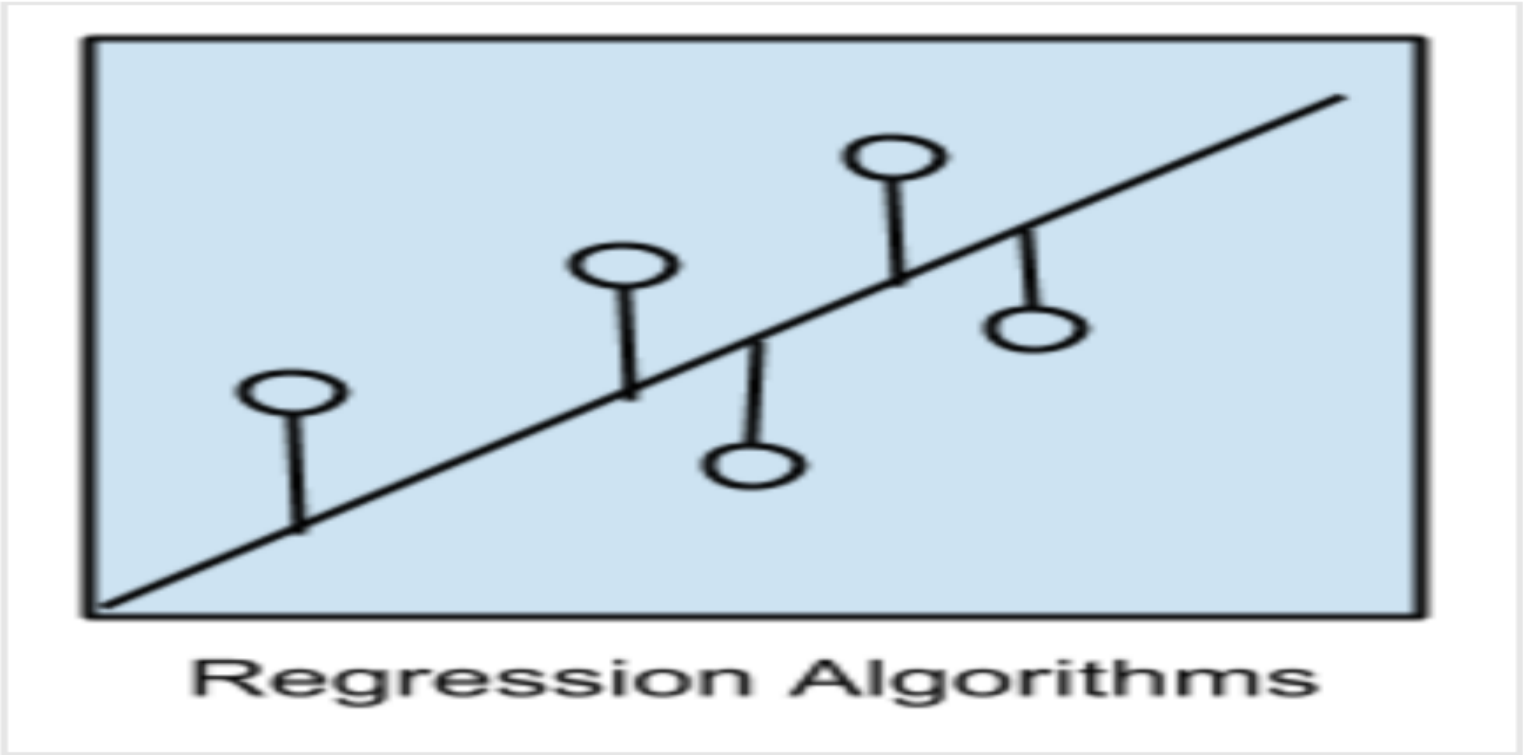


Figura: Modelos baseados em regressão

Sobre esses modelos

- Regressão consiste em treinar um modelo que se adapta numéricamente aos dados de forma iterativa
- A cada iteração reduzimos o erro
- Reg. Linear, Logistica, OLSR, MARS são exemplos de algoritmos de regressão.

Modelos baseados em Instância

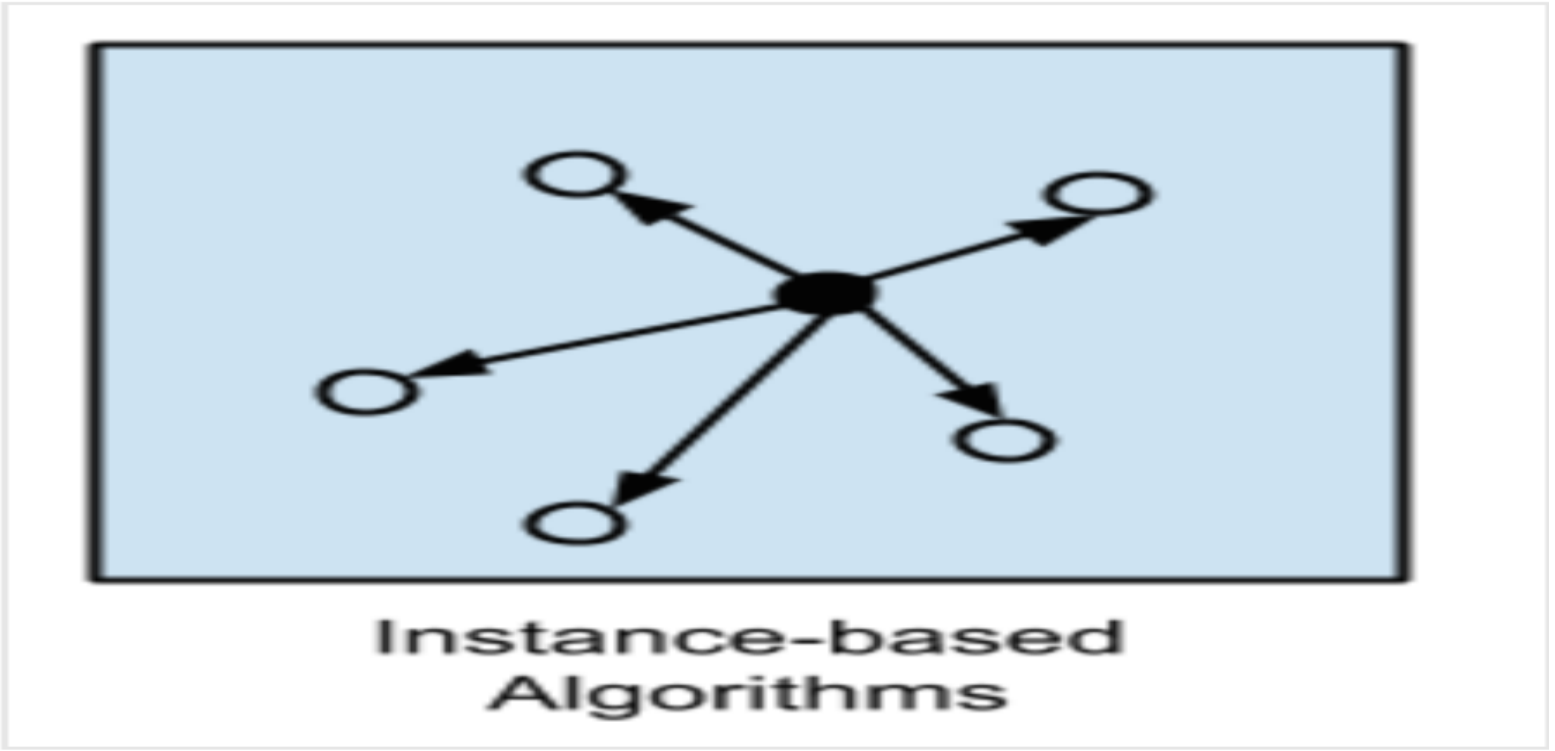


Figura: Modelos baseados em Instância

Sobre esses modelos

- Constroem uma base de conhecimento sobre os pontos dos dados
- Novos pontos são comparados com a base usando uma métrica de distância
- A menor distância é usada para classificar esses pontos
- KNN, LVQ, SOM, SVM são exemplos de algoritmos

Modelos baseados em regularização

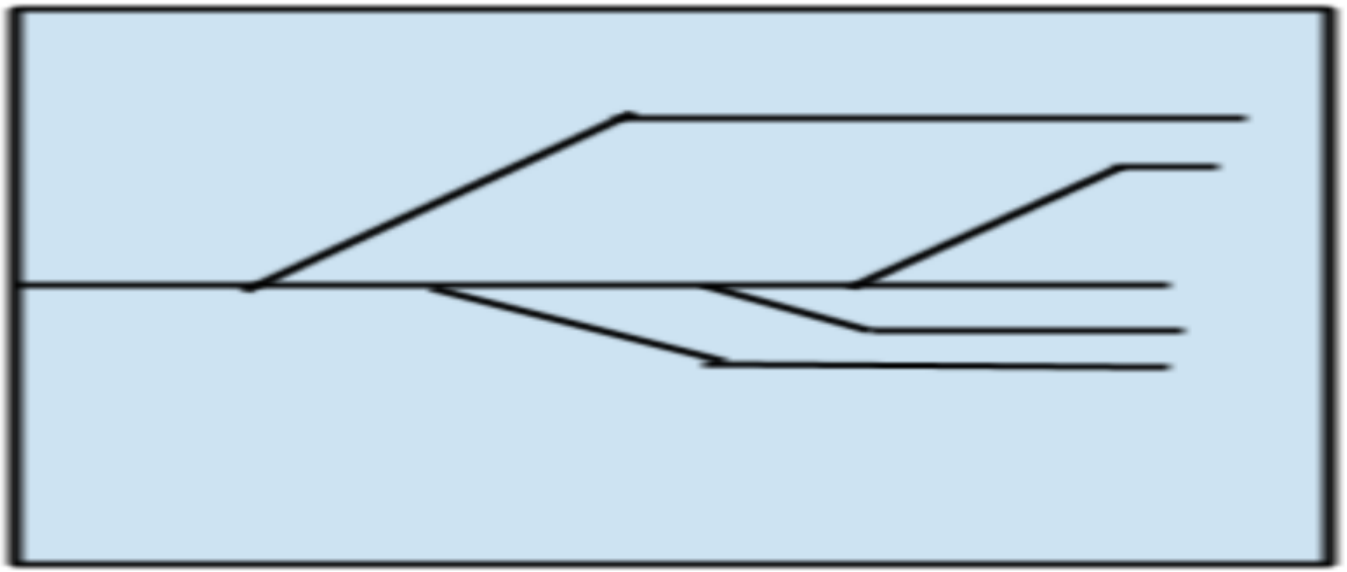


Figura: Modelos baseados em regularização

Sobre esses modelos

- Extensão de técnicas de regressão
- Usa um fator de regularização (modelos mais complexos são penalizados)
- Tenta evitar o overfitting
- Reg. Ridge, LASSO são exemplos de algoritmos

Modelos baseados em Bayes

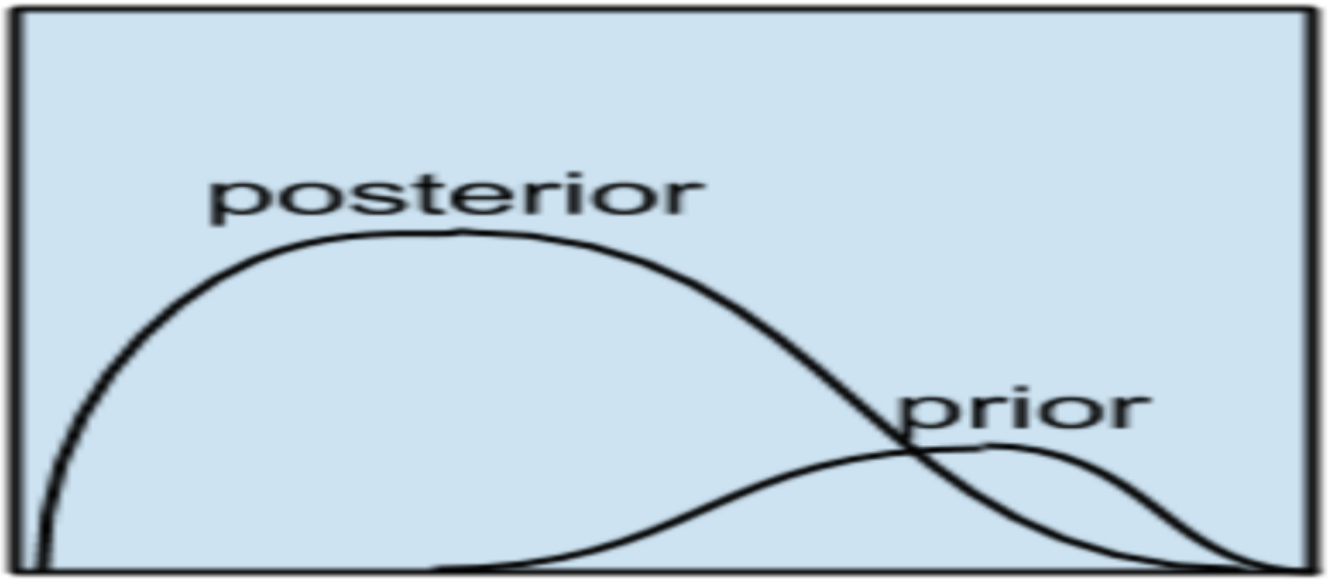


Figura: Modelos baseados em Bayes

Sobre esses modelos

- Aplicam o teorema de Bayes para predizer
- Tipicamente necessitam conhecimento prévio das probabilidades da classe target
- Naive Bayes, Bayesian Network (BN), Multinomial Naive Bayes são exemplos de algoritmos

Modelos baseados em Agrupamento

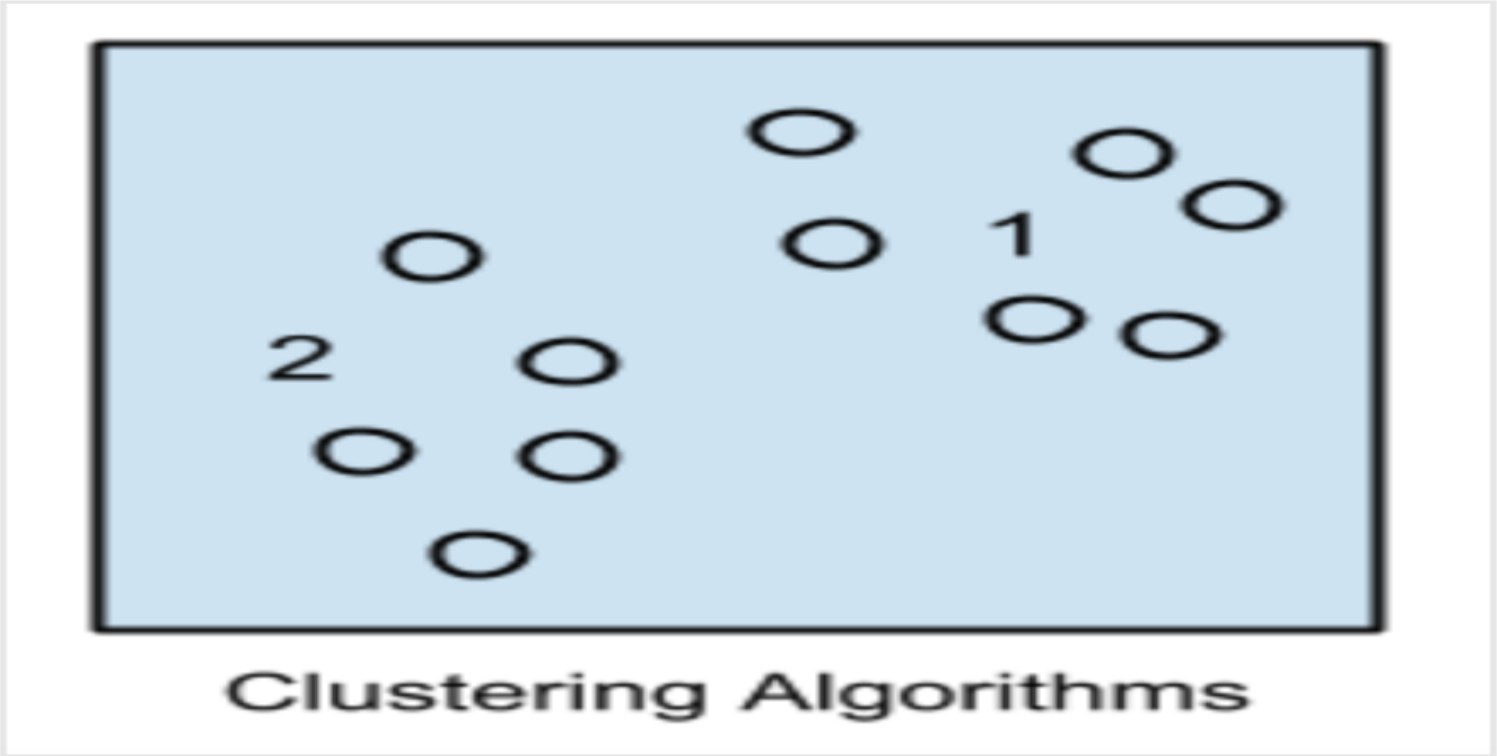


Figura: Modelos baseados em Agrupamento

Sobre esses modelos

- Usam alguma métrica de distância para tentar organizar os dados em agrupamentos
- Há técnicas baseadas em centróides, hierárquicas, densidade entre outras
- k-Means, k-Medians, CLARA, CLARANS, AGNES, DIANA, DBSCAN são exemplos de algoritmos

Modelos baseados em regras de associação

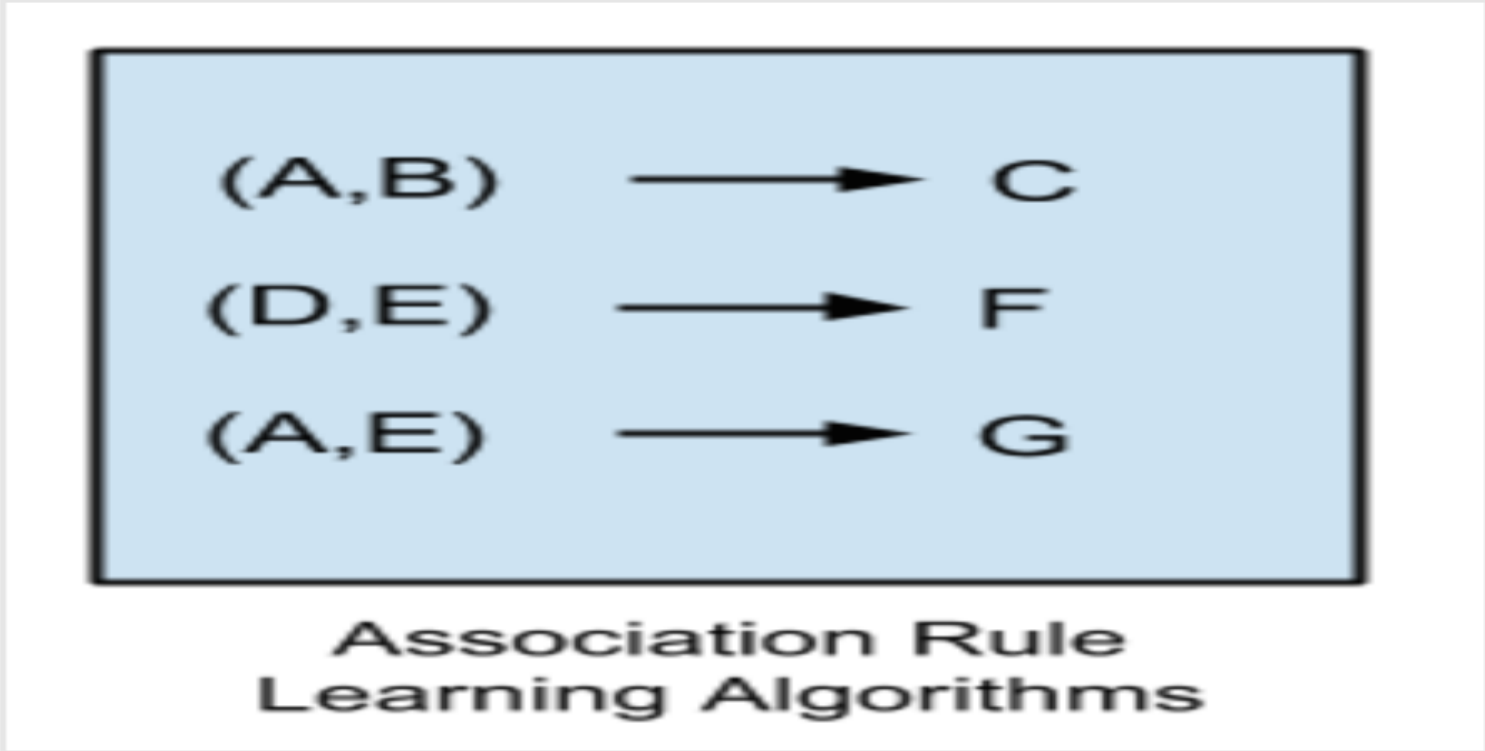


Figura: Modelos baseados em associação

Sobre esses modelos

- Extraem regras que expliquem melhor as relações entre variáveis
- A famosa “análise de carrinho de compra”
- Apriori, Eclat são exemplos de algoritmos

Modelos baseados em redes neurais

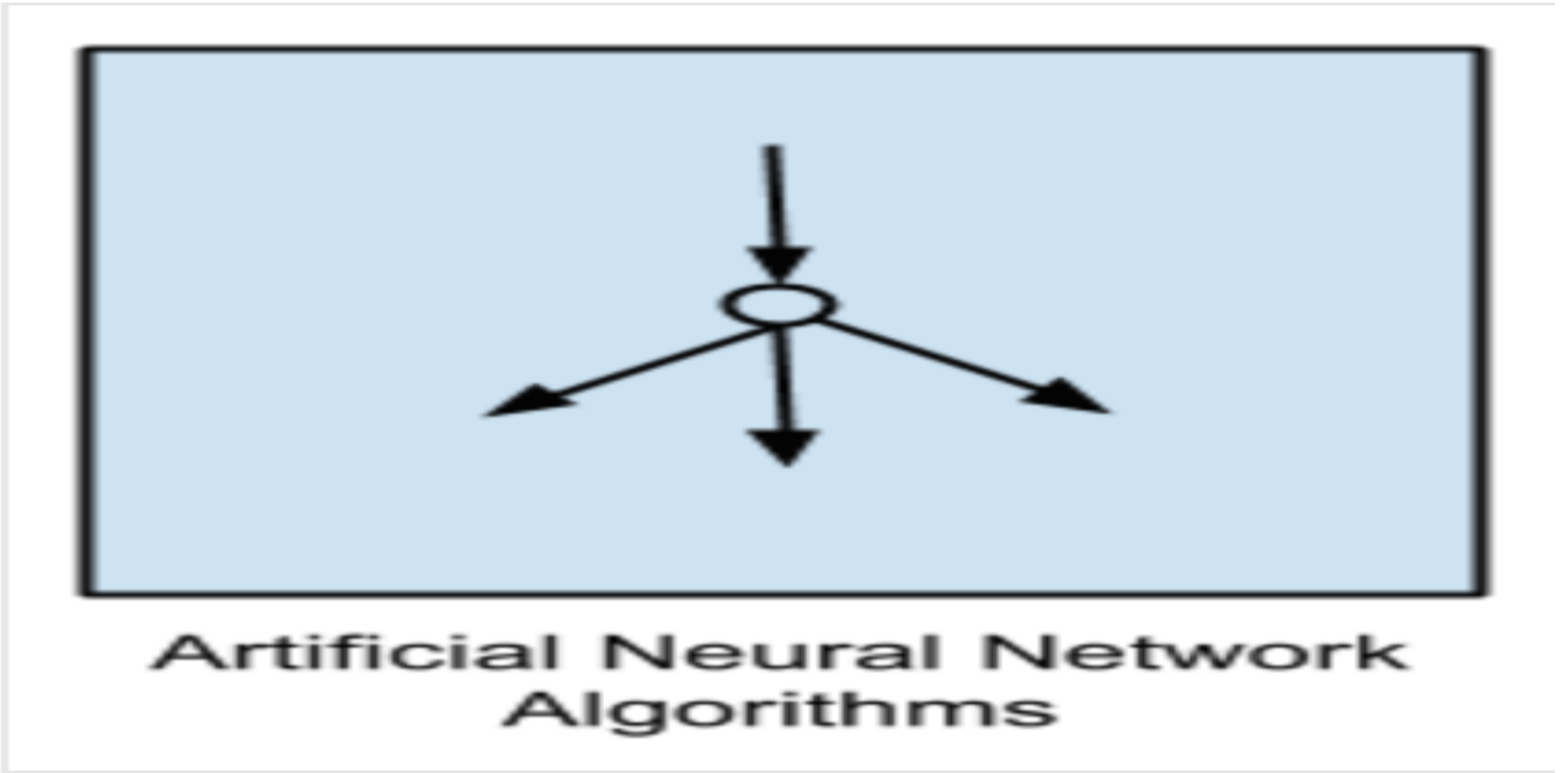


Figura: Modelos baseados em redes neurais

Sobre esses modelos

- Inspirados em neurônios biológicos
- Consistem em encontrar pesos para a rede neural
- Perceptron, MLP, Hopfield Network são exemplos de algoritmos

Modelos baseados em Deep Learning

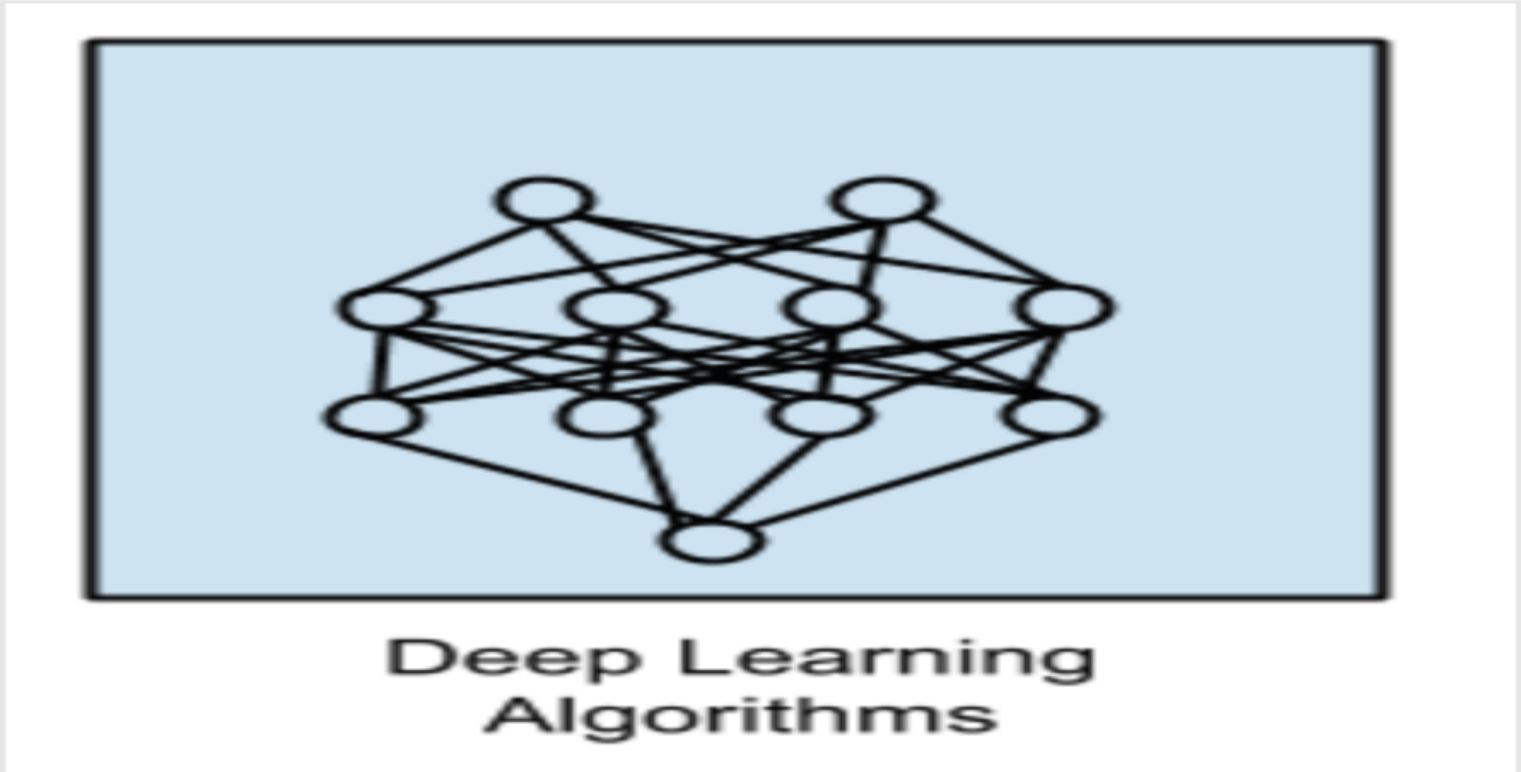


Figura: Modelos baseados em Deep Learning

Sobre esses modelos

- Evolução das redes neurais
- Convolutional Neural Network (CNN), Long Short-Term Memory Networks (LSTMs) são exemplos de algoritmos

Modelos baseados em redução de dimensionalidade

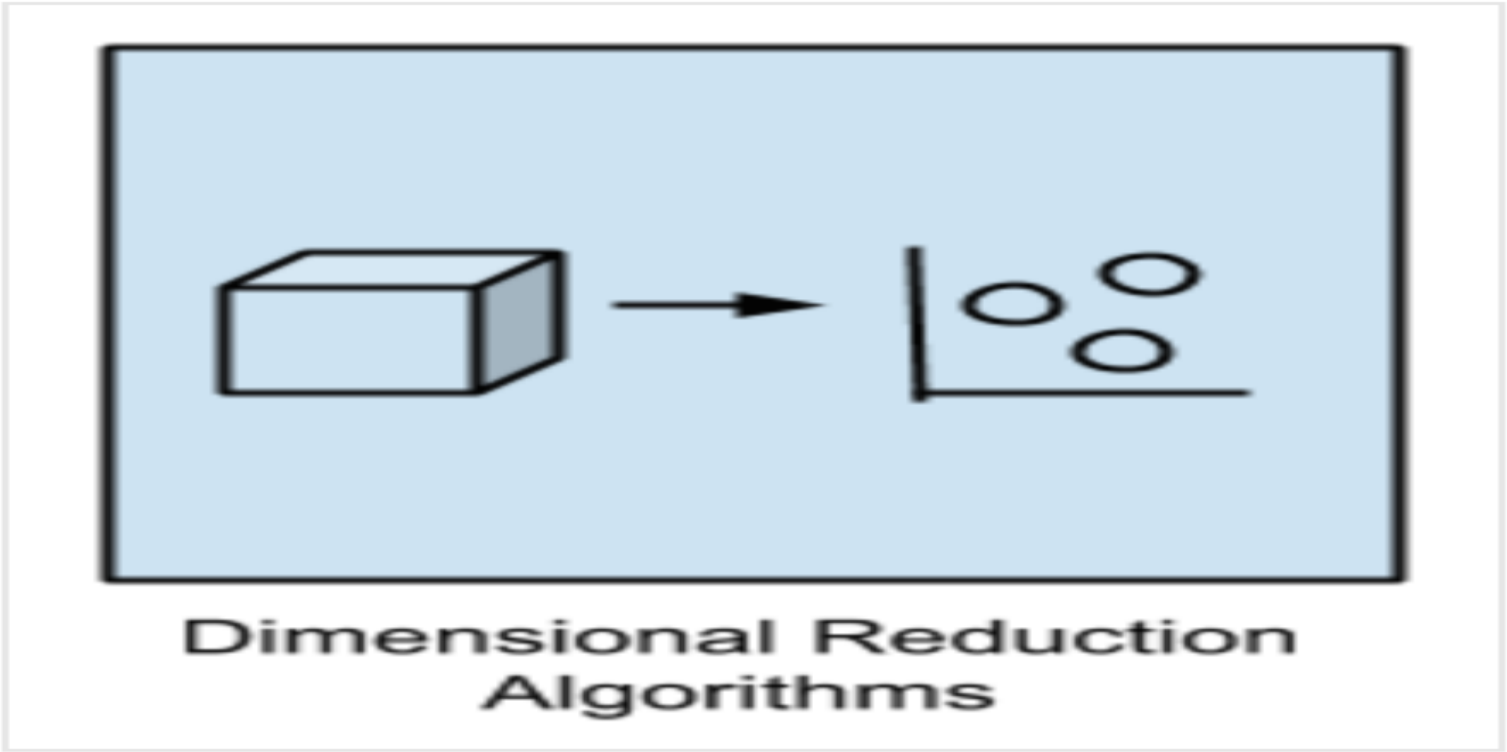


Figura: Modelos baseados em redução de dimensionalidade

Sobre esses modelos

- Explora a estrutura dos dados para tentar remover features que não são importantes
- São não supervisionados
- PCA, SVD, LDA, FDA são exemplos de algoritmos

Modelos baseados em ensemble

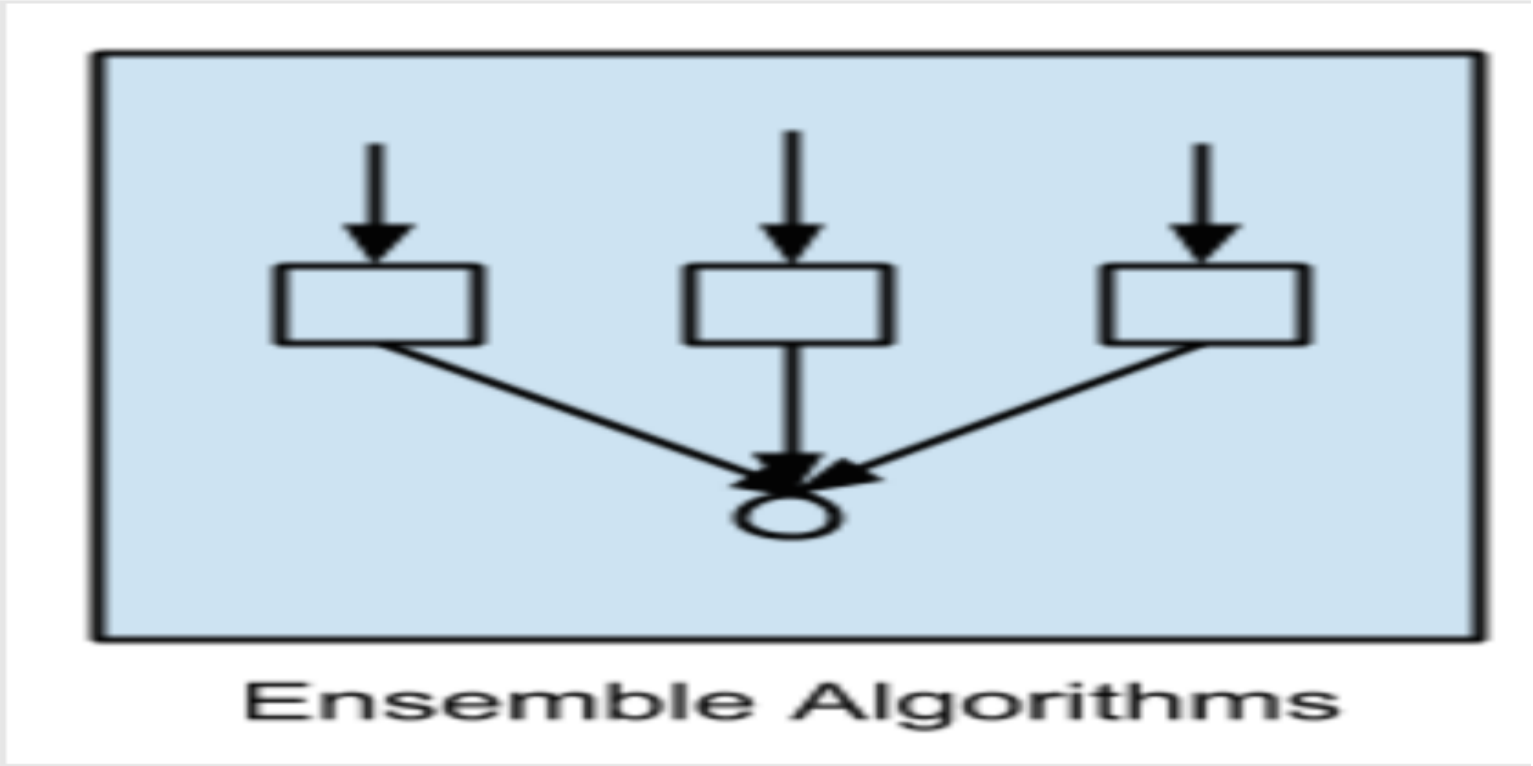


Figura: Modelos baseados em ensemble

Sobre esses modelos

- Modelos compostos de outros modelos treinados de forma independente
- As predições são combinadas de alguma maneira
- Boost, Random Forest, Rotation Forest, Xgboost são exemplos de algoritmos

Dúvidas...

Alguma dúvida?

Contato

- E-mail: 0800*dirso@gmail.com* (alunos SENAC)
- E-mail: *adilson.khouri.usp@gmail.com*
- Phone: +55119444 – 26191
- Linkedin do professor
- Lattes do professor
- Slides no GitHub do professor