Introdução a Machine Learning

Profº Esp. Paulo de Assis
pan2@cin.ufpe.br
paulodeassis.nascimento@gmail.com

Agenda

1 - Sobre mim ;)

2 - Introdução.

3 -Base de Dados.

4 - Classificação.

5 - Regressão.

6 - Clusterização.

7 - Prática (Classificação, Regressão e Clusterização).

Sobre mim

Estudante de Mestrado em Ciência da Computação - UFPE
Análise de Sentimentos - Processamento de Linguagem Natural
Tema da Pesquisa: Redes Neurais Artificiais aplicadas a Análise de Sentimentos baseada em contexto nas Redes Sociais.

Especialização em Engenharia de Software - UNIBRATEC
Tema da Pesquisa: Programação Orientada a Aspectos no Android Studio - Um Estudo de Caso.

Licenciatura em Computação - UFRPE

Desenvolvedor de Sistemas - Policentro Tecnologia.

Software Engineer - Avanade do Brasil

Analista Desenvolvedor Júnior - Supermercados Arco-Mix

Introdução

O que é Machine Learning?

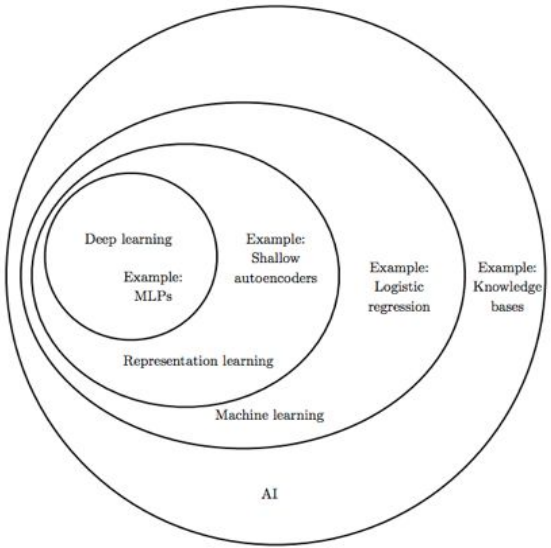
É uma área da Inteligência Artificial que estuda as técnicas computacionais para aprendizado automático por meio de métodos e técnicas de aprendizagem de máquina.

Preocupação

Construção de softwares que melhoram seu desempenho por meio da experiência.

Aprendem automaticamente a partir de grande volume de dados e geram hipóteses a partir dos dados.

Separando as coisas



Mas e Data Mining?

Formalmente

“Um programa aprende a partir da Experiência E, em relação a uma classe de tarefas T, com medida de desempenho P, se seu desempenho em T, medido por P, melhora com E”

Mitchell, 1997

Base de Dados

As bases de dados, também chamadas de Dataset, contém dados de diversas áreas para serem aplicados na Ciência dos Dados (Machine Learning).

Essas bases servem para treinar e testar os algoritmos que são desenvolvidos e estão disponíveis em diversos repositórios.

Normalmente estão disponíveis no formato .csv

Há uma prática quanto ao treinamento dos modelos sobre as bases de dados de modo que elas são divididas em 75% para treinamento e 25% para teste.

kaggle

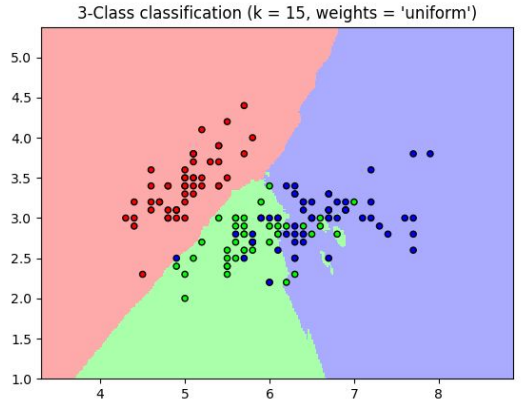
Aplicação da Aprendizagem de Máquina

Classification

Classification (Classificação) - O computador procura especificar a qual das categorias k uma entrada pertence.

Aplicações: Detecção de Spam, Reconhecimento de Imagem.

Algoritmos: SVM (Support Vector Machine), Nearest neighbors, Random forest, Naive Bayes, Decision Tree

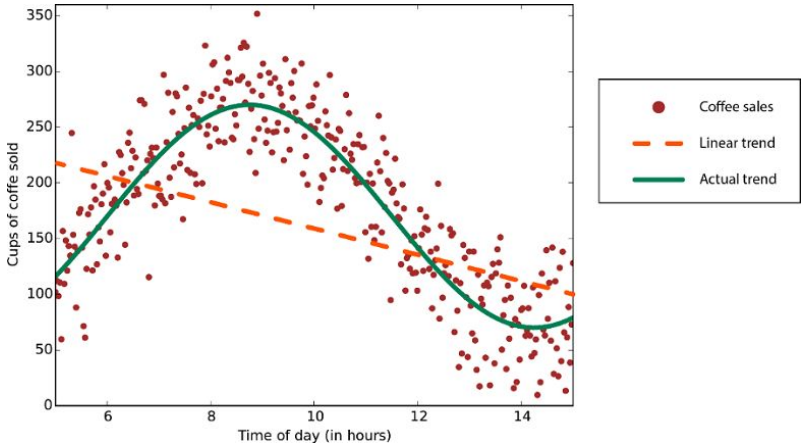


Regression

Nessa atividade, o computador procura uma descrição dos dados sempre em busca de prever valores a partir de uma entrada.

Aplicações: Resposta de drogas (efeito de remédios), curso de bolsa… ações...

Algoritmos: SVR(Support Vector Regression), Ridge Regression, Lasso

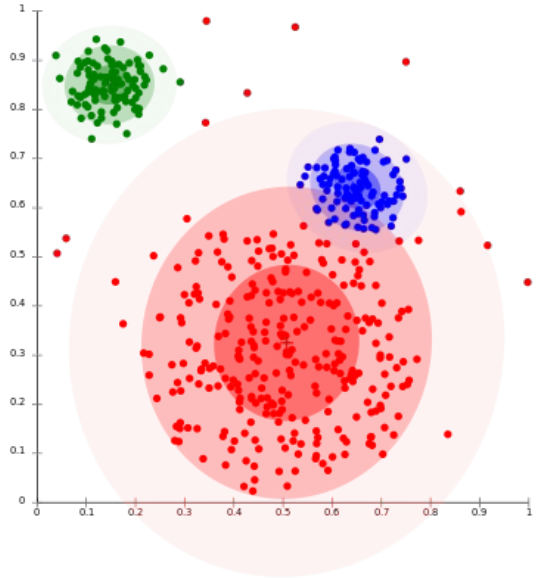


Clustering

Identifica e agrupa objetos que tem semelhança em meio a uma grande quantidade de dados.

Aplicações: Segmentação de clientes, agrupamento de resultados da experiência

Algoritmos: k-Means, spectral clustering,mean-shift



Como tudo isso é feito?

Todas as atividades de Aprendizagem de Máquina são provadas por meio de algoritmos e para cada tipo de atividade há um algoritmo específico.

$$P(w_j|x) = \frac{p(x|w_j).P(w_j)}{p(x)} \qquad p(x) = \sum_{j=1}^{2} p(x|w_j).P(w_j)$$

Algoritmo Bayesiano

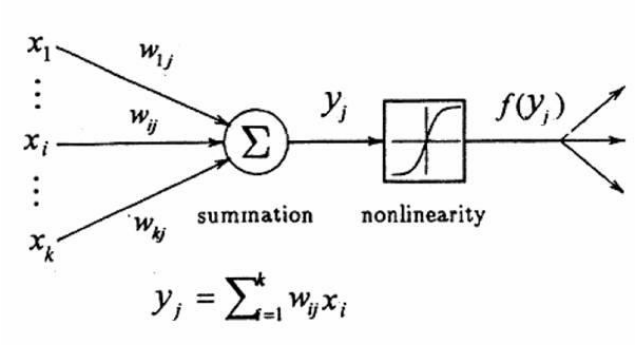
Euclidean $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$

Manhattan $\sum_{i=1}^{k}|x_i - y_i|$

Minkowski $\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$

KNN

Como tudo isso é feito?



Rede Neural Artificial - Multilayer Perceptron (Base do Deep Learning)

Tipos de Aprendizagem

Supervisionada - Os dados disponíveis para aprendizagem contém labels para categorização dos exemplos de treinamento.

Não supervisionada - Não há labels de identificação das classes de modo que o algoritmo tende a clusterizar os objetos por por semelhança.

Configuração do Ambiente.

Python, scikit-learn, numpy, matplotlib etc

1 - Baixar Python da URL https://github.com/paulodeassis/jornada_academica_unibratec

2 - Instalar Python 3.6

3 - Verificar se o pip está instalado: Executar o comando

C:\users\meuUsusario>pip list

4 - Instalar as bibliotecas para Análise de Dados scikit-learn

c:\users\meuUsuario>pip install -U scikit-learn

Prática Classification

Algoritmo KNN

Iris Dataset

5.1	3.5	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor

Attribute Information:

1. sepal length in cm, 2. sepal width in cm, 3. petal length in cm, 4. petal width in cm
5. class: Iris Setosa, Iris Versicolour, Iris Virginica.

Missing Attribute Values: None

Uma dessas classes é linearmente separável das demais, porém duas delas não são.

K-NN / K Nearest Neighbor

1 - Acessar: https://github.com/paulodeassis/jornada_academica_unibratec

2 - Copiar a URL para clonar

3 - Abrir o CMD e efetuar o comando, de preferência, na pasta Documents.
git clone URL que foi copiada

5 - no CMD executar o comando more iris.data

4 - python plot_classification.py

Prática Regression

Algoritmo SVM - Support Vector Machine

SVM / Support Vector Machine

1 - Acessar: https://github.com/paulodeassis/jornada_academica_unibratec

2 - Copiar a URL para clonar

3 - Abrir o CMD e efetuar o comando, de preferência, na pasta Documents.
git clone URL que foi copiada

4 - python plot_svm_regression.py

(os dados para este caso são aleatórios)

Prática Clustering

Algoritmo K-Means

K-Means

1 - Acessar: https://github.com/paulodeassis/jornada_academica_unibratec

2 - Copiar a URL para clonar

3 - Abrir o CMD e efetuar o comando, de preferência, na pasta Documents.
git clone URL que foi copiada

4 - python plot_kmeans_assumptions.py

(os dados para este caso são aleatórios)

Referências

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Duda, Richard O., Peter E. Hart, and David G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

Demonstration of k-means assumptions disponivel em <http://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_assumptions.html#sphx-glr-auto-examples-cluster-plot-kmeans-assumptions-py> acessado em 30.08.2017

Nearest Neighbors Classification Disponível em <http://scikit-learn.org/stable/auto_examples/neighbors/plot_classification.html#sphx-glr-auto-examples-neighbors-plot-classification-py> acessado em 30.08.2017

Support Vector Regression (SVR) using linear and non-linear kernels disponível em <http://scikit-learn.org/stable/auto_examples/svm/plot_svm_regression.html> acessado em 30.08.2017