Curso: Aprendizagem de Máquina em Inteligência Artificial

Disciplina: Aprendizado Não Supervisionado

Prof. Marcelo Novaes de Rezende

Revisão

Aprendizado Supervisionado x Não Supervisionado

Aprendizado Supervisionado :

Há as amostras (X) e os resultados corretos para elas (Y). Queremos descobrir uma função que mapeie X a Y e possa prever, para novos X', novos Y'. O nome decorre do fato que, após o aprendizado, podemos verificar se o algoritmos aprendeu (supervisionar o aprendizado) usando novas amostras (X) para as quais conhecemos as repostas (Y).

Exemplos : Regressão e Classificação

Aprendizado Supervisionado x Não Supervisionado

Aprendizado Não Supervisionado :

Há as amostras (X) .Queremos descobrir alguma estrutura ou distribuição de dados entre elas. Não há resposta certa ou errada verificável, como no caso dos algoritmos de aprendizado supervisionado.

Clustering : o objetivo é descobrir possíveis agrupamentos entre os dados. Exemplo : K-Means

Association : o objetivo é descobrir tendências do tipo : comprou A então comprou B (para usar em Cross –Sell, por exemplo). Exemplo : Apriori

Clustering

Faculdade de Computação e Informática
Mackenzie

2013Q2

Introdução

Clustering, considered as the most important question of unsupervised learning, deals with the data structure partition in unknown area and is the basis for further learning. The complete definition for clustering, however, isn't come to an agreement, and a classic one is described as follows:

(1) Instances, in the same cluster, must be similar as much as possible;

(2) Instances, in the different clusters, must be different as much as possible;

(3) Measurement for similarity and dissimilarity must be clear and have the practical Meaning;

Fonte:A Comprehensive Survey of Clustering Algorithms
Dongkuan Xu· Yingjie Tian 2015

Discussão Inicial

Em que problemas podemos utilizar Clustering?

Faculdade de Computação e Informática
Mackenzie

Processo de Clustering

(1) Feature extraction and selection: extract and select the most representative features from the original data set;

(2) Clustering algorithm design: design the clustering algorithm according to the characteristics of the problem;

(3) Result evaluation: evaluate the clustering result and judge the validity of algorithm;

(4) Result explanation: give a practical explanation for the clustering result;

Fonte:A Comprehensive Survey of Clustering Algorithms
Dongkuan Xu· Yingjie Tian 2015

Clustering (Classic Methods)

Conglomerados (Clusters)

Métodos Hierárquicos → Aglomerativos

Métodos Hierárquicos → Divisivo

Métodos não hierárquicos : Ex: **K-Means**

Algoritmos de Clustering (Classificação usual)

Algoritmos Hierárquicos

Criam uma hierarquia de conjuntos de classes por fusão de classes menores em classes maiores (ascendente) ou por divisão de classes maiores em classes menores (descendente).

O resultado de um algoritmo hierárquico é uma árvore ou dendrograma.

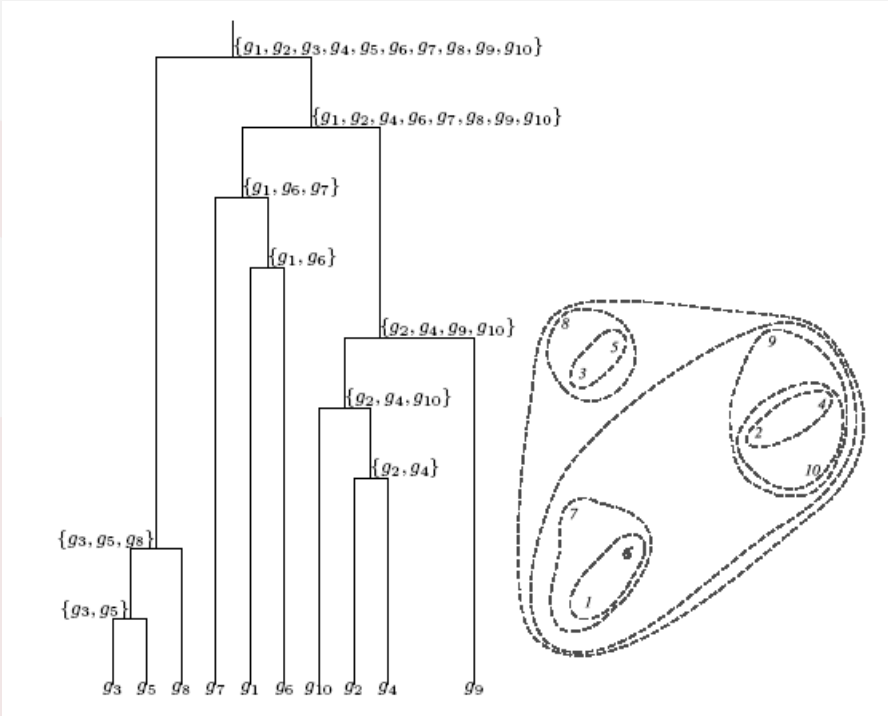
Cortando a árvore num determinado nível é obtida uma partição dos indivíduos em k classes.

Hierárquicos aglomerativos: Partem de **n** individuos agrupados em n classes, cada classe com 1 indivíduo. Agrupam as classes sucessivamente até se obter uma única classe.

Hierárquicos divisivos: Partem de uma única classe que inclui os n indivíduos. As classes são sucessivamente divididas em classes menores até se obterem n classes, cada uma com um indivíduo.

Clustering

Exemplo : Cluster Hierárquico Aglomerativo



Fonte : An Introduction to Bioinformatics Algorithms

K-Means

The core idea of **K-means** is to **update the center of cluster** which is represented by the center of data points, by iterative computation and the iterative process will be continued **until some criteria for convergence is met**. **K-mediods** is an improvement of K-means to deal with discrete data, which takes the data point, most near the center of data points, as the Representative of the corresponding cluster.

Fonte: A Comprehensive Survey of Clustering Algorithms
Dongkuan Xu· Yingjie Tian 2015

K-Means Advantages/Disadvantages

Advantages : relatively low time complexity and high computing efficiency in general;

Disadvantages : not suitable for non-convex data, relatively sensitive to the outliers, easily drawn into local optimal, the number of clusters needed to be preset, and the clustering result sensitive to the number of clusters

Fonte: A Comprehensive Survey of Clustering Algorithms
Dongkuan Xu· Yingjie Tian 2015

Kmeans- Pseudocode

Como é o K-Means em “pseudocódigo”

Defina o número de clusters (k)

Defina os centróides iniciais dos k clusters

Faça

Forme os k clusters associando cada objeto a seu centróide mais próximo

Recompute o centróide de cada cluster

Enquanto mudarem os objetos dos clusters

Clustering

Prática com duas dimensões e norma *euclidiana :

No Excel : Crie 6 pontos no plano

Definiremos 3 clusters

Com centroides iniciais os pontos 1,2 e 3

(usar gráfico dispersão)

Ponto	x	y
1	1	1
2	2	2
3	5	5
4	5	6
5	1	5
6	2	6

*$d_{1,2}=\sqrt{(x1-x2)^2 +(y1-y2)^2}$

Faculdade de Computação e Informática
Mackenzie

Clustering

Atividade 1

Calcule a distância euclidiana de cada ponto aos centróides e verifique o centróide mais próximo

Calcule o novo centróide do cluster

Monte os novos clusters..até que não haja mudança de clusters...

Use uma iteração em uma aba do excel

Partir de K-means.xls

Clustering

Atividade 2: Repetir a atividade do Excel (K-Means) com Pandas no Python

Kmeans.ipynb

Faculdade de Computação e Informática
Mackenzie

Clustering

Atividade 3: Repetir a atividade do Excel (K-Means) com sklearn

Kmeans-scikit.ipynb

Faculdade de Computação e Informática
Mackenzie

Clustering

K-Means consegue perceber o agrupamento das flores por espécie?

Atividade 3 : Partindo de kmeans-íris.ipynb. Verificar se o kmeans encontra 3 padrões como são os targets

Qual métrica usaremos?

Clustering

Custo de uma partição em “n” clusters

$$\sum_{i=1}^{n} d(x_i, cx_i)^2$$

n:número de pontos

d=distância

Cx_i : cluster associado ao ponto x_i

Atividade 4 : estime o custo para n=2 até 7 clusters

Faça o gráfico..qual é o melhor k?..discussão

(base : kmeans-iris-cost.ipynb)

Até a próxima aula

OBRIGADO!

Prof Marcelo Rezende
email rezendemn@gmail.com