Clusterização

Alunos: Anderson Santos, Izaquiel Queiroz
Professor: Paulo Neto
Disciplina: Reconhecimento de Padrões e Aprendizado de Máquina

Aprendizado não supervisionado

Aprendizado não supervisionado

A set of statistical tools intended for the setting in which we have only a set of features X1, X2, ..., Xp measured on n observations. The goal is to discover interesting things about the measurements on X1, X2, ..., Xp.

Is there an informative way to visualize the data? Can we discover subgroups among the variables or among the observations?

(James, Gareth, et al., 2013)

Clusterização

Clusterização

Clustering refers to a very broad set of techniques for finding subgroups, or clustering clusters, in a data set. (James, Gareth, et al., 2013)

Essa divisão é feita de modo que os grupos sejam diferentes uns dos outros, mas os elementos pertencentes ao mesmo grupo sejam parecidos entre si. (Izbicki, Rafael, and Tiago Mendonça dos Santos, 2020)

Medidas de (dis)similaridade

Medidas

Distância Euclidiana

Distância Manhattan

Distância de Mahalanobis

Distância Cosseno

Distância de Jaccard

Medidas

→ Distância Euclidiana

Distância Manhattan

Distância de Mahalanobis

Distância Cosseno

Distância de Jaccard

$$d^2(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2;$$

Medidas

Distância Euclidiana

➔ Distância Manhattan

Distância de Mahalanobis

Distância Cosseno

Distância de Jaccard

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Medidas

Distância Euclidiana

Distância Manhattan

➔ Distância de Mahalanobis

Distância Cosseno

Distância de Jaccard

$$d\left(x^T, y^T\right) = \sqrt{(x-y)S^{-1}(x-y)^T}$$

$$S = matriz\left(s_{kj}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(x_{ik} - \overline{x_k}\right)\left(x_{ij} - \overline{x_j}\right)$$

Medidas

Distância Euclidiana

Distância Manhattan

Distância de Mahalanobis

➔ Distância Cosseno

Distância de Jaccard

$$d(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{||\mathbf{x}_i|| \, ||\mathbf{x}_j||};$$

Medidas

Distância Euclidiana

Distância Manhattan

Distância de Mahalanobis

Distância cosseno

➔ Distância de Jaccard

$$d(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{\sum_{k=1}^{d} x_{i,k} x_{j,k}}{\sum_{k=1}^{d} x_{i,k} + \sum_{k=1}^{d} x_{j,k} - \sum_{k=1}^{d} x_{i,k} x_{j,k}}$$

Algoritmos

K-Means

K-Means

K-means clustering is a simple and elegant approach for partitioning a data set into K distinct, non-overlapping clusters. We seek to partition the observations into a pre-specified number of clusters. (James, Gareth, et al., 2013)

The goal is to partition the data set into some number K of clusters, where we shall suppose for the moment that the value of K is given. A cluster as comprising a group of data points whose inter-point distances are small compared with the distances to points outside of the cluster. (Bishop, Christopher M., 2006)

Pseudocódigo

Algorithm 10.1 *K-Means Clustering*

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:

   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

Aplicações

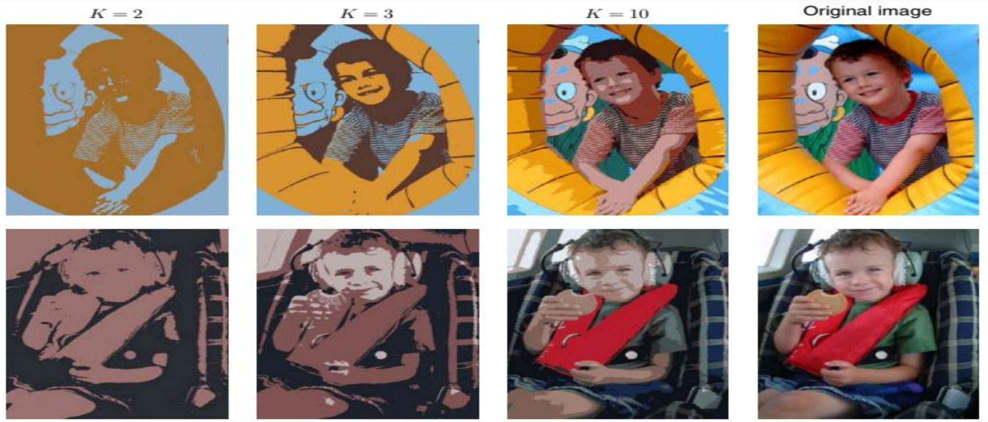


Figure 9.3 Two examples of the application of the K-means clustering algorithm to image segmentation showing the initial images together with their K-means segmentations obtained using various values of K. This also illustrates of the use of vector quantization for data compression, in which smaller values of K give higher compression at the expense of poorer image quality.

Exemplo

Self-Organizing Map (SOM)

Introdução

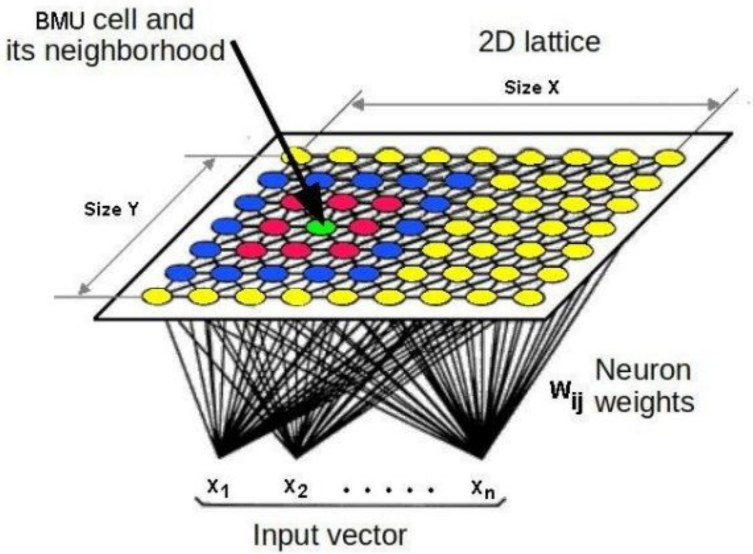
Foi desenvolvida por Kohonen na década de 80

São baseadas no mapa topológico presente no córtex cerebral. Neurônios topologicamente próximos tendem a responder a padrões ou estímulos semelhantes

A cada iteração, os neurônios se especializam em agrupamentos de padrões similares

Utiliza o paradigma da aprendizagem competitiva

Arquitetura



Algoritmo

1. Inicialização: geralmente aleatória
2. Competição: para cada padrão de entrada, calcula-se a resposta dos neurônios de saída. O neurônio com a maior resposta é o vencedor da competição. Denomina-se tal neurônio como BMU (Best Matching Unit)
3. Cooperação: o neurônio vencedor define uma vizinhança topológica (em função da grade) de neurônios excitados
4. Adaptação Sináptica: aprendizado em relação ao padrão de entrada. Os pesos do neurônio vencedor, e de sua vizinhança, ficam mais próximos do padrão de entrada

Competição

Obtenção do BMU por:

$$v = \arg\min_{j} \left\| \mathbf{x} - \mathbf{w}_j \right\|, \quad j = 1,2,\ldots,N$$

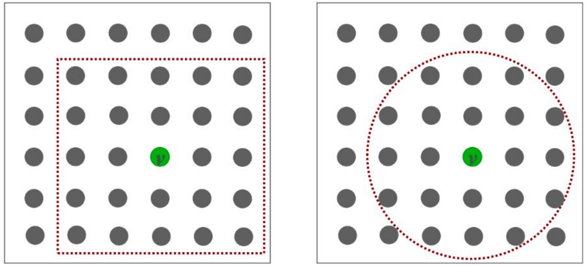
É baseado na maximização do produto interno. Matematicamente equivalente a minimizar a distância euclidiana entre x e w

Processo Cooperativo

Compreende a definição de uma função de vizinhança centrada no neurônio vencedor

Define uma região de neurônios cooperativos, que terão seus pesos ajustados juntamente com o neurônio vencedor

Há diversas formas de implementar a função de vizinhança. A função mais utilizada como função de vizinhança é a Gaussiana.



Adaptação Sináptica

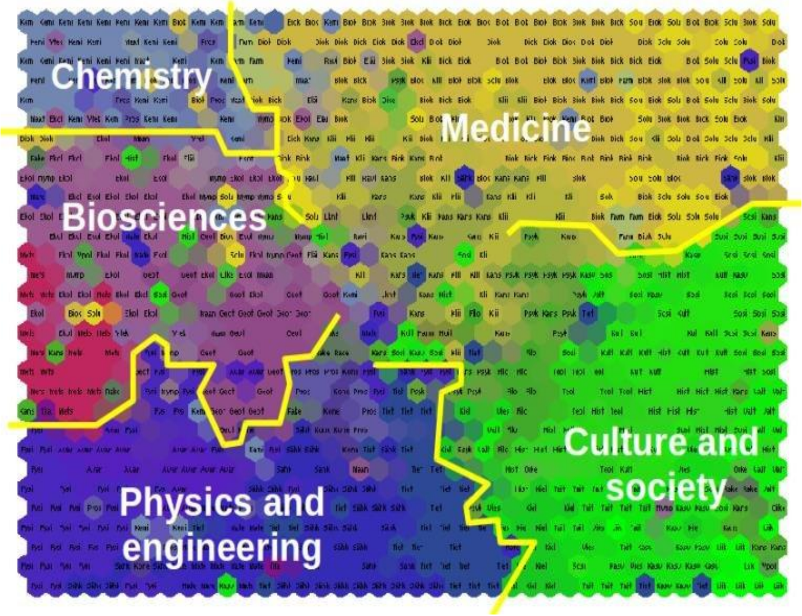
Modificação dos pesos em relação à entrada, de forma iterativa

A equação abaixo é aplicada a todos os neurônios da grade dentro da região de vizinhança

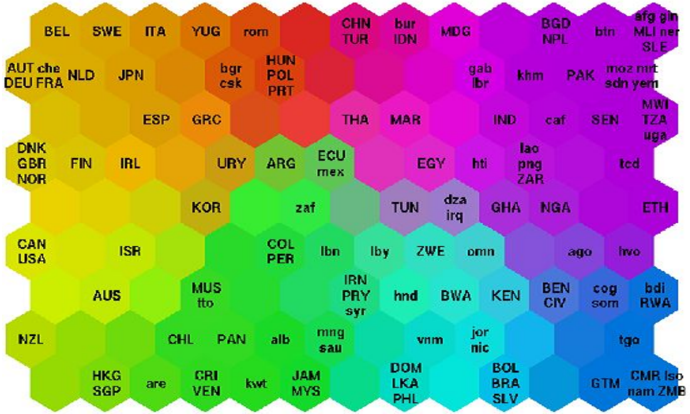
$$\mathbf{w}_j(t+1) = \mathbf{w}_j(t) + \eta(t)h_{vj}(t)[\mathbf{x} - \mathbf{w}_j(t)]$$

Vetor peso atualizado — Vetor peso anterior — Taxa de aprendizagem — Vizinhança — Adaptação

Aplicações


Chemistry
Medicine
Biosciences
Physics and engineering
Culture and society

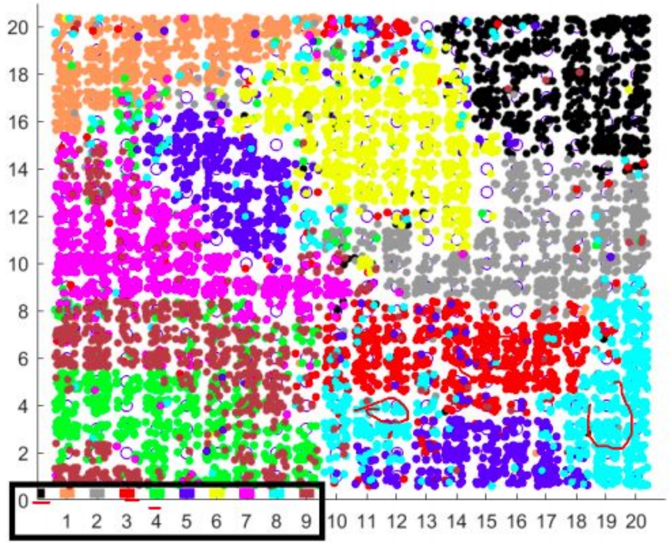
Aplicações



The Country Names

AFG	Afghanistan	GTM	Guatemala	NZL	New Zealand
AGO	Angola	HKG	Hong Kong	OAN	Taiwan, China
ALB	Albania	HND	Honduras	OMN	Oman
ARE	United Arab Emirates	HTI	Haiti	PAK	Pakistan
ARG	Argentina	HUN	Hungary	PAN	Panama
AUS	Australia	HVO	Burkina Faso	PER	Peru
AUT	Austria	IDN	Indonesia	PHL	Philippines
BDI	Burundi	IND	India	PNG	Papua New Guinea
BEL	Belgium	IRL	Ireland	POL	Poland
BEN	Benin	IRN	Iran, Islamic Rep.	PRT	Portugal
BGD	Bangladesh	IRQ	Iraq	PRY	Paraguay
BGR	Bulgaria	ISR	Israel	ROM	Romania
BOL	Bolivia	ITA	Italy	RWA	Rwanda
BRA	Brazil	JAM	Jamaica	SAU	Saudi Arabia
BTN	Bhutan	JOR	Jordan	SDN	Sudan
BUR	Myanmar	JPN	Japan	SEN	Senegal
BWA	Botswana	KEN	Kenya	SGP	Singapore
CAF	Central African Rep.	KHM	Cambodia	SLE	Sierra Leone
CAN	Canada	KOR	Korea, Rep.	SLV	El Salvador
CHE	Switzerland	KWT	Kuwait	SOM	Somalia
CHL	Chile	LAO	Lao PDR	SWE	Sweden
CHN	China	LBN	Lebanon	SYR	Syrian Arab Rep.
CIV	Cote d'Ivoire	LBR	Liberia	TCD	Chad
CMR	Cameroon	LBY	Libya	TGO	Togo
COG	Congo	LKA	Sri Lanka	THA	Thailand
COL	Colombia	LSO	Lesotho	TTO	Trinidad and Tobago
CRI	Costa Rica	MAR	Morocco	TUN	Tunisia
CSK	Czechoslovakia	MDG	Madagascar	TUR	Turkey
DEU	Germany	MEX	Mexico	TZA	Tanzania
DNK	Denmark	MLI	Mali	UGA	Uganda
DOM	Dominican Rep.	MNG	Mongolia	URY	Uruguay
DZA	Algeria	MOZ	Mozambique	USA	United States
ECU	Ecuador	MRT	Mauritania	VEN	Venezuela
EGY	Egypt, Arab Rep.	MUS	Mauritius	VNM	Viet Nam
ESP	Spain	MWI	Malawi	YEM	Yemen, Rep.
ETH	Ethiopia	MYS	Malaysia	YUG	Yugoslavia
FIN	Finland	NAM	Namibia	ZAF	South Africa
FRA	France	NER	Niger	ZAR	Zaire
GAB	Gabon	NGA	Nigeria	ZMB	Zambia
GBR	United Kingdom	NIC	Nicaragua	ZWE	Zimbabwe
GHA	Ghana	NLD	Netherlands		
GIN	Guinea	NOR	Norway		
GRC	Greece	NPL	Nepal		

Aplicações



Exemplo

Outras abordagens

Abordagem hierárquica

Métodos hierárquicos são uma alternativa que evitam o problema de especificar de antemão o valor do K

São representados visualmente por meio de dendrogramas

Pseudocódigo base:

1. Atribua cada observação a um cluster diferente. Calcule cada uma das $\binom{n}{2}$ distâncias entre esses clusters
2. Para i=n, n-1, . . . , 2:
   a. Procure entre todos os pares formados por dois dos *i* clusters aqueles mais parecidos. Junte esses dois clusters em um só. A dissimilaridade entre esses dois clusters indica a altura do dendrograma em que a junção será feita
   b. Calcule cada uma das distâncias entre os novos i-1 clusters

Linkage

Para fazer o agrupamento é necessário definir a distância entre dois clusters. Há várias formas de se definir essas distâncias chamadas de linkage

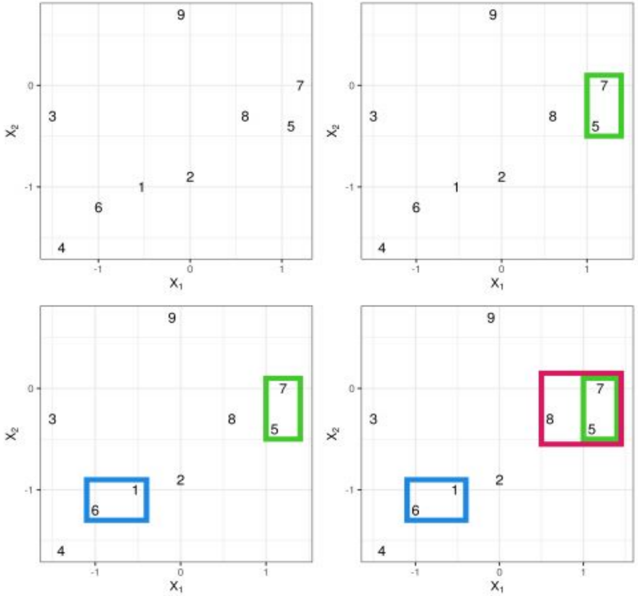
Complete: a maior das distâncias entre todos os pares de observações pertencentes aos dois clusters.

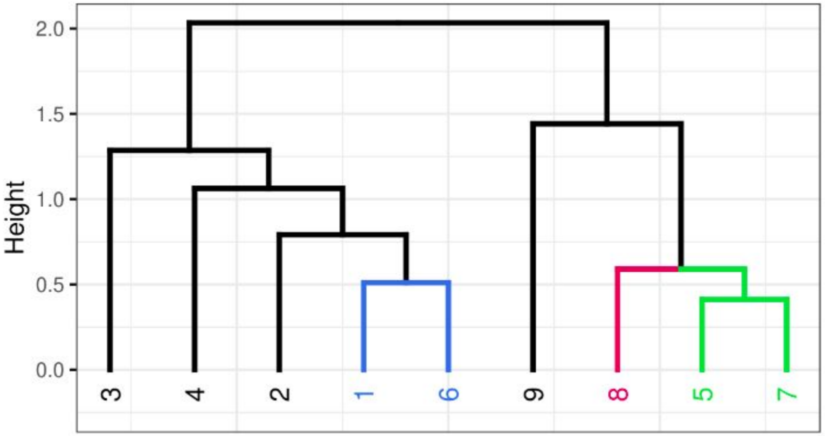
Single: a menor das distâncias entre todos os pares de observações pertencentes aos dois clusters.

Average: a média das distâncias entre todos os pares de observações pertencentes aos dois clusters.

Centroid: a distância entre os centróides dos dois clusters

Exemplo





Referências

Referências

1. Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.
2. Izbicki, Rafael, and Tiago Mendonça dos Santos. Aprendizado de máquina: uma abordagem estatística. Rafael Izbicki, 2020.
3. James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: Springer, 2013.

Complementar:

4. Stork, D. G., Duda, R. O., Hart, P. E., & Stork, D. (2001). Pattern classification. A Wiley-Interscience Publication.
5. De Sa, JP Marques. Pattern recognition: concepts, methods and applications. Springer Science & Business Media, 2012.
6. Murty, M. Narasimha, and V. Susheela Devi. Introduction to pattern recognition and machine learning. Vol. 5. World Scientific, 2015.