PERGUNTAS E RESPOSTAS – MBA EM DATA SCIENCE E ANALYTICS

Disciplina: Unsupervised Machine Learning: Clustering II

Data: 15/06/2021

Nelida Elizabet Quiñonez Silvero

Eu vi em alguns exemplos que as variáveis categóricas, depois de convertidas para “dummy”, se calculam o PCA e se seleciona os PCs que explicaram a maior variabilidade possível. Isso procede?

Tecnicamente, a clusterização usa somente variáveis métricas ou binárias. Para variáveis categóricas é aconselhável usar análise de correspondência.

Cleverson de Souza

E a função "daisy" do pacote CLUSTER. estava lendo que ela faz isso com factors, é isso?

A função daisy permite encontrar a distância entre as linhas quando as variáveis não estão no mesmo formato. Tecnicamente, a clusterização usa somente variáveis métricas ou binárias. Para variáveis categóricas é aconselhável usar análise de correspondência.

Sintia Silva De Almeida

Vai ser trabalhado algum exemplo de clusterizacao para sequencias genéticas?

Oi Sintia. Acho que foge do escopo do curso, por tratar de tema muito específico. Contudo, você pode utilizar análise de cluster para fazer tal análise

Isadora Salvador Rocco

O que acontece quando a distância euclidiana de A e B eh igual a e C, mas a de B e C eh maior? Quem agrupa com A primeiro?

Vai depender das demais variáveis do modelo. Vale destacar que na análise de cluster um indivíduo pode ficar na interseção de 2 clusters. Geralmente quando há muitas interseções a clusterização não foi satisfatória.

Raphael Ribeiro Da Silva Albino

Em relação a análise de grupos. Caso eu tenha uma variável qualitativa (cor). Eu poderia ver cada cor como um comprimento de onda em um espectro e daí fazer a análise de cluster?

Não conheço nenhuma função no R, mas no Phyton existem algumas ferramentas neste sentido: https://acertbr.com.br/extracao-de-paleta-de-cores-com-k-means-clustering/

Igor De Oliveira Cardoso Nobre Potengy

Ao tentar fazer uma análise de cluster com 3 variáveis padronizadas, dificilmente consigo "clusters" formados quando tenho variáveis igual a zero (forma só 1 grupo + outliers). Devo fazer algo mais?

Pode tentar outros tipos de medida de distância. Medida de Jaccard é bastante utilizada para este tipo de dados. Essa medida não leva em conta a frequência do par de respostas 0-0, considerada irrelevante. Entretanto, é possível que ocorra uma situação em que todas as variáveis sejam iguais a O para duas determinadas observações, ou seja, somente exista frequência na célula d da Tabela 9.7. Nesse caso, softwares como o Stata apresentam medida de Jaccard igual a 1, o que faz sentido do ponto de vista de similaridade.

Marcus Costa

Professora como faz para plotar os valores das intersecções no dendograma?

O gráfico do dendograma já plota de forma automática os valores das distâncias no gráfico.

Henrique Pinto Caria Oliveira

As linhas ficaram com os nomes dos alunos, mas a variável alunos continua lá

Verifique de novo a aula por volta dos 50min. Tente roda dos códigos iniciais novamente, especialmente aqueles das linhas 23 até 26

Erika Aparecida Gava

Caso eu tenha dois alunos com o mesmo nome, por exemplo Bru, como se comportaria como nome da linha?

Seria interessante renomear um dos alunos, tipo Bru1 e Bru2

Hugo Hideo Tanaka

Adriana, a primeira variavel com nome continha aparecendo mesmo executando rownames(alunos_pap)

Verifique de novo a aula por volta dos 50min. Tente roda dos códigos iniciais novamente, especialmente aqueles das linhas 23 até 26

Artur Pires De Jesus

Quando utilizo o 'rect.hclust' no dendograma single, as linhas ficam paralelas, não marcam o dendograma com uma linha horizontal como em outros casos. Por quê?

Deve ser o padrão do código. É necessário verificar a documentação dó código para saber qual motivo.

Alex Jornada Queiroz

Qual a diferença das funções rect.hclust pro cutree?

O rect.hclust é o método tradicional de clusterização hierárquica. Enquanto o cutree você estabelece o número de clusters. Neste vídeo você pode verificar bem como isso funciona: https://www.youtube.com/watch?v=GPOUGpF-Sno

Renata Maria Marè Gogliano

Professora, você falou dos traços ou perninhas dos gráficos: quando são longos demais, pela sua experiência você entende que é preciso ter mais grupos. Mas como o leigo pode fazer essa interpretação?

Renata, o interesse é rodar vários modelos com número de clusters e distâncias diferentes, para saber qual melhor responde sua pergunta de pesquisa.

Eliane Chinaglia

Adriana, qual a diferença em abrir direto um arquivo com extensão ".R" ou ".Rmd" de abrir um arquivo .Rproj?

.R ou .Rmd abre apenas os scripts. Enquanto o .Rproj abre todo projeto com os respectivos scripts.

Henrique Azevedo Marques Araujo

o que é o n = n() na função summarise?

Cria uma variável n que é o número de indivíduos que tem em cada grupo.

Alberto Prado

O que é a varivel N dentro da sumarizacao?

Cria uma variável n que é o número de indivíduos que tem em cada grupo.

Israel Luiz Harmendani Diniz

O Bind é similar ao Joiin no SQL?

Sou sincero. Não conheço SQL e foge um pouco do escopo do curso.

Antonio Piratelli Filho

Como converter um arquivo de codigo formato .R em formato ipynb, para usar no Júpiter notebook via navegador?

Olá Antônio. Não conheço como fazer essa conversão e essa pergunta foge um pouco do escopo do curso.

Andre Kenji Yai

É certo afirmar que nem sempre conseguiremos separar em um total de 1 a n (numero de elementos) distintos. Pois o corte em uma certa altura pode resultar em mais de uma separação.

O ideal é fazer vários modelos com vários parâmetros para verificar qual melhor responde sua pergunta de pesquisa.

Luciane Sebastião

o Valor energético depende diretamente da quantidade de carboidratos, gorduras, etc.... usar valor energético não representa uma dupla valoração?

Depende muito a teoria que você irá utilizar para conceituar suas variáveis. O analista de dados deve interpretar as variáveis e definir seu escopo. Em análise de dados não existe uma reposta certa ou errada, mas um conjunto de ferramentas para te ajudar a tomar decisões.

Roberta Mayumi Takenaka Granero

Adriana, mas os clientes teriam a percepção por exemplo de quantidade de ferro ou fibra ? Faria sentido usar isso como variável? Entendo que algumas pessoas conseguem perceber os carbs e proteínas.

Depende muito a teoria que você irá utilizar para conceituar suas variáveis. O analista de dados deve interpretar as variáveis e definir seu escopo. Em análise de dados não existe uma reposta certa ou errada, mas um conjunto de ferramentas para te ajudar a tomar decisões.

Wilians Pereira Dutra

mesmo no canal 3 deu uma travada aqui. como é atualizado o centroide?

Oi Willians. A aula fica gravada, assim você pode verificar sua dúvida caso haja interrupção da gravação.

Mauricio Eneas De Oliveira

Taina, na aba "Packages" flega o pacote factoexta caso não esteja selecionado.

Na aba “packages” você pode instalar os pacotes necessários para sua análise, mas não esqueça de chamar eles depois com “library”

Renata da Encarnação Onety

O que fazer com os dados de bordas dos clusteres? Tem um default para isso?

O ideal é fazer vários modelos com vários parâmetros para verificar qual melhor responde sua pergunta de pesquisa. Os dados de bordas podem trazer insights interessantes sobre os seus dados.

Murilo Marin Pechoto

Existe alguma funcção similar a cutree para usar fazer a mesma análise com o kmeans?

O rect.hclust é o método tradicional de clusterização hierárquica. Enquanto o cutree você estabelece o número de clusters. Neste vídeo você pode verificar bem como isso funciona: https://www.youtube.com/watch?v=GPOUGpF-Sno

Vitor Bruno da Silveira Guimarães

Minha pergunta não foi qual utilizar, mas como utilizar no R. Não ficou claro nessa parte.

Oi Vitor. Tente verificar na aula gravada se sua pergunta foi respondida. Caso não consiga, pode enviar um email com toda questão para tentarmos responder.

Augusto Ponso Alves

Pergunta para o professor João (IBS Americas): Seria possível termos o contato de alguém que fez esse módulo para conversar sobre como foi a experiência?

Oi Agusto. Acho que no site do IBS Americas e tentar verificar se eles podem te dar essa orientação. (https://ibs-americas.com/pt/)

Bruno César Moreira

na tabela "mediagrupo", como identificar qual é qual no caso dos grupos que só tinham 1 observação?

Oi Bruno, você pode usar a função filter() para criar uma tabela nova somente com os casos do grupo.

Lucas Roberto da Silva Dias

Há diferença entre os conceitos de [1] distância para os centróides e [2] as distâncias geodésicas [teoria de grafos]?

São algoritmos diferentes com objetivos diferentes que utilizam formas distintas para calcular as distâncias.

Monica Regina Mandari

Error in file(file, "rt") : não é possível abrir a conexão Além disso: Warning message: In file(file, "rt") : não foi possível abrir o arquivo 'dados/alunos_pap.csv': No such file or directory

Oi Monica. Espero que tenha conseguido abrir os arquivos. De toda forma, tente assistir a aula de novo. Caso não consiga, talvez esse tutorial pode te ajudar: https://pt.coredump.biz/questions/45022511/in-filefile-quotrtquot-cannot-open-the-connection-in-addition-warning-message-in-filefile-quotrtquot-cannot-open-file-no-such-file-or-directory

Alex Ananias Da Silva

Error in colMeans(x, na.rm = TRUE) : 'x' deve ser numérico

Oi Alex. Espero que tenha conseguido abrir os arquivos. De toda forma, tente assistir a aula de novo. Caso não consiga, tente largar a primeira coluna com `prcomp(d[,-1])`

Debora Duarte Pinheiro

Qual a diferença do método Elbow e do R^2?

O R2 é uma medida para verificar a variância dentro dos grupos, enquanto o método Elbow é um algoritmo usado para clusterizar.

ANTONIO THYRSO CORSINO PEREIRA de SOUZA

Sobre isso método elbo, vale a pena derivar a curva e encontrar o "melhor" números de clusters?

Não conheço essa análise. Talvez possa ser feito, mas é necessário saber se faria sentido, considerando que o algoritmo de cluster já é bem completo.

Rafael Viegas De Carvalho Carlos Gomes

As dimensoes sao as variaveis do McDoanld (calorias, carboidratos, proteinas...) ou sao outros parametros?

Não sei se entendi bem a pergunta, mas as dimensões são formada pela própria análise de cluster.

Laila Monte Neto Donni

Qual a diferença do método k-means para o SVM

Svm é utilizado para aprendizagem de máquina como classificador binário não probabilístico, enquanto o k-means serve para clusterizar observações.

Edward Jonathan Chirinos Peralta

cual e a differencia entre k-means e bisecting k-means?

Neste site tem um tutorial que pode te ajudar: https://spark.apache.org/docs/latest/ml-clustering.html#:~:text=Bisecting%20k-means.%20Bisecting%20k-means%20is%20a%20kind%20of,performed%20recursively%20as%20one%20moves%20down%20the%20hierarchy. Mas, basicamente, são algoritmos diferente que apresentar resultados distintos. Vale a pena rodar os dois modelos para definir qual deles responde melhor sua questão de pesquisa.

Leonardo Pacheco Silva

Que tipo de custo está vinculado ao numero de clusters?

Depende bastante da configuração dos seus dados.

Bruna Muller

professora, na clusterização não hierárquica, como que eu "descubro" quais são as variáveis da dim 1 e dim2? serão as mesmas independentemente do K?

As variáveis usadas para clusterização serão aquelas que você escolheu.

Marcio Adriano Nichimura

e se definir 7 para escolher 3 com menos variabilidade é uma forma de análise

Toda análise que responde sua pergunta de pesquisa é válida. O melhor é utilizar várias técnicas até você conseguir responder suas questões de pesquisa.

Fernando Brito

Não seria interessantes antes de uma análise de cluster verificar os market shares baixos e procurar talvez uma correlação entre eles? De forma a já associar produtos que talvez precisem ser treinados

Toda análise que responde sua pergunta de pesquisa é válida. O melhor é utilizar várias técnicas até você conseguir responder suas questões de pesquisa.

Adriano de Camargo Bisogni

Esse caso dos taxistas não é um exemplo que nem sempre usar somente algumas dimensões pode atrapalhar a analise?

Toda análise que responde sua pergunta de pesquisa é válida. O melhor é utilizar várias técnicas até você conseguir responder suas questões de pesquisa.

Vitor Hugo Miro Couto Silva

Adriano Bisogni, tem um função do Scipy chamada pdist. Veja aqui: https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.pdist.html

Obrigado pela dica Vitor.

Adriano de Camargo Bisogni

Uma nova variável com o número de produtos com Market share abaixo do esperado nao poderia ajudar na clusterizacao?

Toda análise que responde sua pergunta de pesquisa é válida. O melhor é utilizar várias técnicas até você conseguir responder suas questões de pesquisa.

Gustavo Murad

na entendi a ultima etapa do dbscan

Olá Gustavo. Você pode tentar assistir novamente a aula. Caso a dúvida persista entre em contato que podemos te indicar material complementar.

Eduardo Francisco Xavier Salles

Na aula passada foi falado sobre uma distância a ser evitada. Havia entendido que se chamava distância de Goura, mas procurei na internet e não encontrei resultados. Poderiam confirmar o nome?

Oi Eduardo. Poderia assistir de novo a aula e verificar qual seria esse nome. Não lembro de qualquer nome similar esse “Goura”

Yuri José de Santana Furtado

Por que no exemplo dos municípios São Paulo não ficou sozinho? Com um centroide só pra ele?

Depende muito das variáveis e dos dados. O algoritmo calcula isso de forma automática.

Eduardo Luís Hammes

Toda vez que tento rodar o comando fviz_nbclust, recebo a mensagem de erro the previous r session was abnormally terminated due to an unexpected crash

Oi Eduardo. Tente instalar os pacotes iniciais da aula e depois chamá-los no library.

Vitor Fernandes Jaguanharo Carvalho

Professora, não entendi muito bem a escala utilizada na plotagem dos gráficos de cluster utilizando a metodologia k-means. Por que temos duas dimensões, se utilizamos mais variáveis em nosso modelo?

Os indivíduos são agrupados a partir das variáveis semelhantes.

Eduardo da Silva Neto

Para quem usa linux o arquivo alunos_pap.csv está com o caractere especial no nome "Ze" e ocorre erro ao executar o plot

Tente mudar o nome Ze e retirar o caracter especial.

Lincoln Amaral Sotto

Por favor demonstrar na prática a classificação do nome do grupo nos dados, em resumo, criar uma coluna no final do dataset com o nome do grupo.

Oi Lincol. Tente usar a função filter() para criar uma nova tabela.

Alexandre Barros dos Santos

Prof. meu grafico k=3 está difetente, o cluster 1 (bolinha vermelha) está isolado

Tente repetir o exemplo. Não há como se diferente se os passos foram os mesmos.

Rafael Viegas De Carvalho Carlos Gomes

Poderia explicar novamente o que são as Dimensoes (dim1, dim2...) no grafico? Sao as varaiveis/colunas?

Oi Rafael. Você pode rever a aula gravada. Mas é basicamente o tamanho da dimensão do grupo.

Ingrid Harumi Miura

a base de dados municipio nao está carregando.

Oi Ingrid. Tente carregar os pacotes do início da aula e repetir o exemplo gravado.

Bruno Speria

Não ficou muito claro para mim as explicações sobre as dimensões no gráfico do fviz_cluster?

Oi Bruno. Tente assistir a aula gravada novamente para ter esse insight.

Eduardo Vinicius Ransolin Pigoso

Professora, eu não entendi a parte da Dim1 e Dim2, como eu defino ou descubro quais grandezas são essas?

Oi Rafael. Você pode rever a aula gravada. Mas é basicamente o tamanho da dimensão do grupo.

Fernando da Silva Razera

usa-se graus em ewbow para definir qual o melhor ponto?

Isso mesmo. Para definir o melhor número de clusters.

Israel Luiz Harmendani Diniz

Há alguma forma de analisar qual é o ganho em agregar ou não um grupo de forma quantitativa de forma análoga a custo variável??

Toda análise que responde sua pergunta de pesquisa é válida. O melhor é utilizar várias técnicas até você conseguir responder suas questões de pesquisa.

Rafael Bellotti Moreno

O que são as dimensões dos eixos do cluster não-hierárquico ("Dim2(20.7%)" e "Dim1(64%)")?

Você pode rever a aula gravada. Mas é basicamente o tamanho da dimensão do grupo.

Rafael Lopes De Souza

existe alguma função R que me dá a menor variação no método elbow?

Oi Rafael. Não conheço se existe alguma função ou teste neste sentido.

Ricardo José Pfitscher

Professora, e se na escala y do gráfico do cotovelo plotássemos em escala logaritmica?

Toda análise que responde sua pergunta de pesquisa é válida. O melhor é utilizar várias técnicas até você conseguir responder suas questões de pesquisa.

Marcus Costa

Poderia informar algum link com a fórmula matemática do método de elbow?

Desculpa, Marcus. Mas não conheço. Mas basicamente ele plota as variâncias dos grupos no gráfico.

Diego Barbosa Batista

Como extrair a diferença dos pontos no R?

Pela matriz distância acho que você consegue esses pontos.

Renata Maria Marè Gogliano

Mas não podia só DOIS LANCHES?? Como ela sugeriu 4 para o cliente??

Toda análise que responde sua pergunta de pesquisa é válida. O melhor é utilizar várias técnicas até você conseguir responder suas questões de pesquisa.

Adriano de Camargo Bisogni

Mas o K-Means calcula o elbow com o somatório das distancias, certo?

Usa a variância.

Hernandes Matias Junior

Como transformamos variáveis qualitativas em quantitativas para utilizarmos o KMeans?

Clusterização não usa variáveis qualitativas. Para isso use análise de correspondência.

Adriano de Camargo Bisogni

O grafico do elbow mostra um número ótimo sugerido! Ele deve ser usado como base para uma discussão com o Business, e nessa reunião será definido qual o melhor numero de grupos. certo?

Toda análise que responde sua pergunta de pesquisa é válida. O melhor é utilizar várias técnicas até você conseguir responder suas questões de pesquisa.

Rodrigo Peixoto J. C. Ferrão

Tem como levar a informação se é borda, centro ou oulier como uma coluna para a tabela final, alem da coluna com o numero do grupo?

Você pode usar a função filter() para isolar os elementos do grupo.

Fernando Brito

Antes de utilizar um método de Cluster, não seria interessante analisar se haveria alguma correlação entre market shares baixos? Vendo se algum dos cinco produtos se relacionava diretamente com outro.

Toda análise que responde sua pergunta de pesquisa é válida. O melhor é utilizar várias técnicas até você conseguir responder suas questões de pesquisa.

Rodrigo Pravalão

também usa holdout e baggins para os modelos preditivos não supervisionados?

Acho que pode fugir um pouco do escopo. Cada técnica tem objetivos diferentes.

Laila Monte Neto Donni

qual site ela falou?

Oi Laila. As aulas ficam gravadas, assim você pode tirar essa dúvida.

Américo Lopes Rodrigues Junior

Adriana, boa noite,uma opção para a solução seria entender se o market share do vendedor tá baixa pela falta de treinamento ou pq a região não se interessa pelo produto? Veria o impacto da região

Toda análise que responde sua pergunta de pesquisa é válida. O melhor é utilizar várias técnicas até você conseguir responder suas questões de pesquisa.

Adauto da Silva Teixeira

uma pergunta sobre ferramenta: no SAS seria mais simples?

Depende muito do conhecimento do cientista de dados. Mas com certeza o R é uma ótima ferramenta.

Fernando Gonçalves

Prof, só tem os metodos hybrid, raw e dist.

Existem vários métodos, mas focamos nos mais conhecidos.

Bruno Speria

exista alguma forma de rodarmos o r2 entre estes clusters?

Oi bruno. Acho que você pode isolar os grupos e calcular suas variâncias como explicado anteriormente.

Gabriela Werner Ceschini

Na tabela de municípios, o nome permaneceu na primeira coluna (variável categórica). Isso afeta o k-means?

A primeira coluna são os indivíduos que serão distribuídos nos clusters.

Adriana Melges Quintanilha Weingart

Ter essa "sensibilidade" do que faz sentido ou não (método a escolher, valores, grupos), não seria "adaptar a pergunta à resposta"? Esse conceito para melhor escolha ainda não está claro...

Toda análise que responde sua pergunta de pesquisa é válida. O melhor é utilizar várias técnicas até você conseguir responder suas questões de pesquisa.

Andrea da Costa Moreira de Oliveira

qual o mínimo de MInpts posso colocar como parâmetro?

Depende muito dos seus dados para fazer sentido. O melhor é utilizar várias técnicas com vários parâmetros, inclusive de MInpts, até você conseguir responder suas questões de pesquisa.

Carlos Eduardo Almeida Gomes

Prof, no caso do DBSCAN a forma que eu coloco os atributos poderia se tornar uma ponderação arbitraria?

Não necessariamente, considerando que o método já requer que você estipule os parâmetros de cálculo.

Ricardo Santana Feitosa

Quais critérios deveriamos adotar para definir estes parâmetros (EPS/MinPts) numa análise?

Depende muito dos seus dados. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Adriana Melges Quintanilha Weingart

A escolha dos parâmetros (Eps e MinPts) não é "regra". pode-se dizer que depende da experiência, ou tentativa-e-erro, ou? Só a experiência pode me dar os "valores" de início p/ testar os parâmetros?

Depende também muito dos seus dados. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Adauto da Silva Teixeira

escolher um valor para o dbscan não seria algo arbitrário?

Não necessariamente, considerando que o método já requer que você estipule os parâmetros de cálculo.

Alex De Lima Bassi

Professora, o Dbscan me parece uma ponderação arbitraria.

Não necessariamente, considerando que o método já requer que você estipule os parâmetros de cálculo.

William Henrique Stenico

Pode citar um exemplo de negócio em que o DBSCAN faça sentido?

Depende também muito dos seus dados. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Mariana Salustiano Rosa

Como podemos definir o eps a partir de uma base de dados?

Depende também muito dos seus dados. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Gabriel Mendroni De Souza

Como chegamos no valor do raio? EPS?

Depende também muito dos seus dados. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Maria Clara Barreiros Rodrigues

Como sugerir o valor do eps? Aleatorio?

Depende também muito dos seus dados. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Marcelo Patto

de onde veio o eps = 056? (linha 96)

Oi Marcelo. Foi respondido durante a aula à 3:15:16 horas

Rafael Da Silva Mello

Existem técnicas para determinar os valores de eps e MinPts do dbscan?

Não conheço técnicas específicas. Depende também muito dos seus dados. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Felipe de Souza Mendes e Silva

Como eu defino qual será o eps e o minPts do dbscan?

Não conheço técnicas específicas. Depende também muito dos seus dados. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Alexsandro Nicácio Siqueira

Existe algum método para sabermos qual eps e MinPts utilizar no Dbscan?

Não conheço técnicas específicas. Depende também muito dos seus dados. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Fernando Gonçalves

Prof, no meu código eu forcei que o DBSCAN encontrasse o mínimo de 5 grupos, mas ele ignorou e colocou 4 grupos. O que houve?

Talvez o algoritmo não tenha suportado 5 clusters, uma vez que ele faz de forma automática. Neste método você não estabelece o número de clusters.

Fernando Sarracini Júnior

como definir minpoints a eps do DBSCan para um dado caso? seria arbitrário e iterativo visualizando resultados?

Depende também muito dos seus dados. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa. Não considero arbitrário, considerando que o próprio algoritmo requer os parâmetros.

Hyago Marinho Reis

Como definir o valor de EPS ?

Depende também muito dos seus dados. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Felipe Francisco Nusda

podemos utillizar uma equação para chegar ao eps ou minpts?

Não conheço técnicas específicas. Depende também muito dos seus dados. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Juliano Santos

Dentro do conceito do Dbscan, poderiamos considerar a questão de "contiguidade" dos pontos?

Se fazer sentido para sua análise, sim. Não existe uma forma única de analisar. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Bruno Belintane

Professora, os parâmetros do Dbscan (Eps = 1, por exemplo) não são arbitrários?

Não necessariamente, considerando que o método já requer que você estipule os parâmetros de cálculo.

Alexandre Barros dos Santos

Prof, neste caso o raio é a dist. Euclidiana?

Não. É o simples tamanho do raio. Existe formula matemática específica para calcular o raio de uma circunferência.

Matheus Pontara

como defino as variaveis Eps e MinPts?

Depende também muito dos seus dados. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Mariana Rillo Otero

Dúvida: por que no método Dbscan não falamos em distância e sim densidade?

É uma especificidade do método, considerando que no referido método você estipula o tamanho do raio e a densidade que será utilizada dentro do círculo.

Ariston Farias

Boa noite, os pontos que se usa para calcular as distâncias estão padronizados?

Geralmente é de bom tom padronizar as variáveis.

Mauricio Matos da Silva Leal

não consegui padronizar a tabela de municípios. Comandei a linha 182 e não funcionou.

Oi Maurício. Tente repetir os comandos da aula. Não há motivo para dar erro. Verifique se instalou todos os pacotes e se chamou eles com “library”.

Marcelo Patto

Prof nao deveria levar em consideração o unit cost do produto? Talvez o problema não seja treinamento e sim o custo do produto e gross margin

Se fazer sentido para sua análise, sim. Não existe uma forma única de analisar. Vai muito do contexto e do seu objetivo. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Rodrigo Faria Soares

No método DBSCAN eu consigo selecionar o número exato de clusters a ser gerado? Ou ele sugere automaticamente baseado no raio e no número de pontos?

Conforme o código do algoritmo apresentado na aula não é possível estabelecer a quantidade de grupos.

Felipe Pereira Delage

Caso tenha cidades com o mesmo nome para estados diferentes, teria como concatenar por exemplo o nome da cidade + uF caso existisse sendo o identificador da linha?

Acredito que seria mais fácil mudar o nome da cidade. Tipo cidade1/Estado A e cidade2/Estado B.

Rodrigo Vitali Kramper

Linha 40: comando notas_alunos_pad - scale(notas_alunos), gera erro: Error in colMeans(x, na.rm = TRUE) : 'x' must be numeric. O que aconteceu?

Oi Rodrigo. Tente baixar todos os pacotes e depois chamá-los com “library” e repetir os códigos, uma vez que todos os código foram testados e funcionam bem.

Rodrigo Pravalão

um EDA com boxplot pegando entre o 1º e 2º quartil entre as metricas nos produtos seria uma boa iniciativa pra separar os dados pra clusterizar?

Se fazer sentido para sua análise, sim. Não existe uma forma única de analisar. Vai muito do contexto e do seu objetivo. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Lucas Rodrigues Gagliardi

Qual o código para mudar o ~ do projeto?

Oi Lucas. Não sei se entendi bem a sua pergunta. Mas você pode renomear o projeto diretamente no Windows Explorer.

Arthur Adabo De Camargo

Temos como substituir o nome dos eixos no gráfico da função fviz_cluster pelo nome dos componentes principais? Vi que tem o argumento "choose.vars" mas seria bom se houvesse uma forma automática

Olá, Arthur. Não conheço se existe tal função. Mas depois de exportar a figura para o seu relatório você pode manipular os dados para melhor compreensão, sempre mantendo o conteúdo original.

Rafael William Fabricio Campanholo

Estou com o seguinte erro ---> > rownames(municipios) - municipios[,1] Error in `.rowNamesDF-`(x, value = value) : 'row.names' com comprimento inválido Além disso: Warning message: Setting ro

Oi Rafael. Tente baixar todos os pacotes e depois chamá-los com “library” e repetir os códigos, uma vez que todos os código foram testados e funcionam bem.

Gustavo Murad

seria bom ter uma ideia do modelo/tabela que foi usada no exemplo de case apresentado...ainda claro que com dados não reais do caso do cliente.

Oi Gustavo. Com os modelos apresentados na aula você consegue ter visão do assunto. Não acho que professora tenha autorização para usar os dados do cliente.

Lucas Alves Dias Cardoso

O que a Profa fez foi uma análise em 5 dimensões, em vez de bidimensional? (ou seja, uma dimensão para cada produto?)

Na verdade, foram criados 6 clusters que agrupou 5 produtos. Não existe uma forma única de analisar. Vai muito do contexto e do seu objetivo. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Rafael Loureiro Smiderle de Moraes

Professora, pode explicar pq renomeamos as linhas?

Na verdade, a primeira linha passar a ser o objeto que vai ser distribuído nos clusters.

Paulo Henrique Real Leite

Eu não entendi o motivo do algoritmo pegar 84% da amostra, qual o motivo disso?

Oi Paulo, o algoritmo faz de maneira automática para melhor dispor os dados.

Gabriela Ribeiro

Estou fazendo a analise descritiva dos municípios, mas não sei o qual tabela colocar no group_by()

Escolha a tabela que os municípios estão inseridos. Esta função agrupa o dataframe por várias colunas com média, soma e outras funções como contagem, máximo e mínimo.

EMANUEL RODRIGUES DE VARGAS

Mas podemos utilizar variáveis multinomiais, para o problema distância com variáveis dummies? Como grau de cheiro ácido, entre 1 e 5?

Olá. Não verifico ganho de eficiência em dummizar este tipo de variável, considerando que o grau de cheio ácido pode variar numa escala numérica entre 1 e 5.

Adriano de Camargo Bisogni

Eu costumo arbitrar valores com valores diferentes, em ordem de grandeza diferente, das demais colunas e isso parece afetar menos o agrupamento. Faz sentido?

Depende muito dos seus dados. Só não vale usar variáveis categóricas em análise de cluster, considerando que existem outros modelos que aceitam esse tipo dado, tipo análise de correspondência.

Pedro Henrique Esteves Trindade

Adriana, é possível incluir algum fator na análise de cluster (séries temporais)? Ou deveria fazer clusters separados?

É possível fazer um cluster de cada unidade de tempo e depois comparar a evolução dos clusters.

Marcel Alexandre Fenerich

Boa noite. Uma dúvida. Falando em variáveis categóricas de CORES. E se convertermos pra RGB ou HEXA, seria válido a conversão de categórica pra numérica?

Acredito não ser possível, considerando que o formato não permitem calcular uma distância, como no exemplo Slateblue: HEXA (#6A5ACD) e RGB (106,90,205). Não conheço nenhuma função no R, mas no Phyton existem algumas ferramentas neste sentido: https://acertbr.com.br/extracao-de-paleta-de-cores-com-k-means-clustering/

Savio Costa

mas qual o conceito para definir o EPS?

Depende também muito dos seus dados. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.

Eduardo Augusto Da Costa Cordeiro

Ajuda professora! estou trabalhando com dbscan e ainda não consegui definir uma questão, qual a melhor forma de definir o eps e o Min de pontos

Depende também muito dos seus dados. O melhor é utilizar várias técnicas com vários parâmetros, até você conseguir responder suas questões de pesquisa.