PERGUNTAS E RESPOSTAS – MBA EM DATA SCIENCE E ANALYTICS

Disciplina: Unsupervised Machine Learning: Clustering I

Data: 08/06/2021

Carla Porto Veiga 20:34

No caso do trabalho das fraudes que você utilizou um modelo não-supervisionado, vc poderia ter usado o resultado extraído nesse modelo para usar num modelo supervisionado?

A partir de uma clusterização não supervisionada, você poderá usar as características dos grupos criados e utilizar outro método supervisionado como uma regressão múltipla ou logística.

Lucas Augusto Nogueira Henrique 19:37

o modelo RFV pode usado como um modelo baseline? para ser comparado com outros modelos posteriormente?

Caso você tenha esses dados do modelo do RFV, sim. Após uma clusterização você poderá utilizar os grupos formulados em modelos supervisionados, tipo regressão múltipla ou logística.

Gabriel Moisés 19:36

Em Marketing Digital, mais especificamente em Mídia Programática usamos muita clusterização baseado principalmente na navegação do Usuário. Com o fim dos cookies pelos browsers, há como criar cluster?

Acredito que utilizando técnicas de mineração de dados/web scrapping é possível, sim. Todavia, com a vigência da LGPD é necessário analisar essa possibilidade.

Gustavo Murad 19:42

o modelo de identificacao de imagem nao seria superviosionado? ou seja, ele nao aprende atraves de ver varias vezes foto de PATO como sendo o Y?

Depende. Caso você utilize parâmetros iniciais do que seria um PATO seria um modelo supervisionado. Caso não tivesse esses parâmetros seria um modelo não supervisionado.

Rafael Vieira Carelli 23:01

Tem algum exemplo para a padronização da variável?

Por exemplo, você pode padronizar as variáveis peso e altura para que ambas estejam compreendidas na mesma escala entre 0 e 1. A distribuição de frequência não seria alterada.

Carlos Henrique Ferreira de Souza 19:56

O software IRAMUTEQ cria clusters para análises textuais. Ele se baseia nos cálculos de frequências absolutas das palavras e no qui-quadrado. Este método ainda seria arbitrário?

É necessário estudar o algoritmo que o software usa para criação destes clusters.

Yuri Godoi Pereira Elias 19:36

Professora boa noite! Muitas vezes tenho que utilizar da criação de um Pseudo-Painel para análises supervisionadas, a minha dúvida é se eu perco eficiência no modelo ao utilizar cluster conjuntamente

Geralmente a utilização de mais de uma técnica de análise de dados vai te devolver insights. Mas sempre vai depender dos seus dados e de quais os seus problemas de pesquisa. É necessário ver se a utilização da técnica faz sentido sobre o que você procura.

Adonai Almeida 19:29

Referente a esse exemplo da SERASA, pode usar os dados dos clientes? Onde entra LGPD?

Conforme o artigo 7º da LGPD. O tratamento de dados pessoais poderá para proteção do crédito, inclusive quanto ao disposto na legislação pertinente. Desta forma, a utilização de dados

para fins de proteção ao crédito é autorizada pela lei. Todavia, é necessário observar toda legislação sobre o tema e os direitos do consumidor. Com certeza, a jurisprudência irá solver algumas dúvidas sobre o tema a partir da aplicação da referida lei.

Rodrigo Eduardo dos Santos Bernardes 19:42
Qual caracteristica fez com que os taxistas fossem clusterizados separadamente?
Para responder sua pergunta você poderia isolar o grupo e estudar os dados a partir de outras ferramentas de análise, tais como uma regressão múltipla e etc. Pode-se utilizar também uma análise simples de frequência das variáveis do grupo ou construção de gráfico bloxpot. Desta forma, seria possível saber quais as variáveis mais influenciaram na construção dos respectivos grupos.

Adriano Lombardi 22:59
padronizar escala é calcular Zscore por exemplo?
A utilização do Z-escore é a forma mais usual de padronização de variáveis.

Fabio Hemerson Araujo de Souza 19:35
Professora! Por que na maioria das vezes os índices de Dunn e Calinski-Harabasz divergem na decisão da quantidade de clusters?
O índice de Calinski-Harabasz (CH) indica que o número ótimo de cluster advém da média das somas dos quadrados dos clusters “between” e “within”. Enquanto isso, o índice de Dunn compara as distâncias intergrupos com tamanho do mais disperso. Enfim, são algoritmos que trabalham com fórmulas matemáticas diferentes. Cabe ao analista verificar qual deles responde melhor os seus questionamentos de pesquisa.

Patricia Hornink Mora 20:45
O Re x Zé não é um grupo? Tem o mesmo resultado de Li x Re
Oi Patrícia, conforme o cálculo de distância apresentado no modelo, não. Verifique que a distância entre Re x Zé é diferente do Li x Re.

Haraldo Cesar Saletti Filho 20:14
Ao analisar o dendograma, contamos com dados adicionais sobre grau de similaridade intra grupo e grau de diferença entre grupos?
Com certeza. O próprio dendrograma ajuda na visualização de como os grupos são homogêneos entre si. A depender do principalmente tamanho do dendograma do volume de dados e do software utilizado pode ser possível verificar no próprio gráfico as similaridades.

Kauê Bonato De Araujo 20:50
Como seriam feitos os cálculos de distâncias para quebra de grupos numa abordagem top-down do modelo hierárquico?
Os cálculos do dendrograma utilizam os resultados das distâncias fornecidas pelo algoritmo escolhido pelo analista. Desta forma, a quebra dos grupos depende de como é realizado o cálculo da distância e como o analista considera a melhor segmentação.

Teresa Arlinda De Souza Campos 20:34
Ainda com a saida dos taxistas, como extrair do cliente o publico alvo quando ele não sabe?
Após a exclusão dos taxistas, pode-se rodar uma nova clusterização para verificar as características dos clientes que permaneceram na base de dados.

Artur Pires De Jesus 21:56
Adriana, pode explicar melhor o conceito do centroide? Ele é um ponto de equilibrio preciso entre a média ou eixo? Queria entender mais o conceito por trás da equação

O centróide é o ponto médio central dos grupos criados e ajudarão a encontrar as semelhanças deste agrupamento. Como se fosse o centro de gravidade do grupo.

Quezia Abreu Salles Barbosa 21:35
Prof não seria 3,16 no Bru? Não entendi.....
Oi Quezia. Foi exatamente esse resultado apresentado no cálculo.

Quezia Abreu Salles Barbosa 22:00
Prof o BRU está como 3,46. Não seria 3,16?
Oi Quezia. Foi exatamente esse resultado apresentado no cálculo.

Francisco Flavio Ribeiro Viana 21:31
Professora, boa noite! Obrigado pela aula! Se eu tivesse que escrever o código, poderia considerar essa distância do ponto médio (LiRe) com Bru (euclidiana)?
Olá. Todas as distâncias calculadas na tabela foram extraídas a partir da distância euclidiana. Os procedimentos posteriores servem exatamente para calcular as distâncias entre os grupos e os indivíduos para depois agrupá-los. Mas, lembre-se que você poderá usar os outros métodos: vizinho mais próximo, vizinho mais longe e etc.

Flavia Cristiane Pinto Maciel 21:18
professora não consegui entender a análise que entre os dendogramas. Como avaliar a diferença para ver qual usar ?
Vai depender de quais agrupamentos fazem sentido na sua análise. O dendograma apresenta uma representação gráfica de como os grupos podem ser divididos. Assim, cabe ao analista estudar qual a melhor divisão dos grupos para responder seu problema de pesquisa.

Murilo Urssi Malek-Zadeh 21:31
Se eu calcular o centro geometrico entre Li e Re, e entao calcular a distancia desse centro ate Bru fica igual [interrogacao] - meu teclado nao esta indo pontuacao
O centróide é o ponto médio central dos grupos criados e ajudarão a encontrar as semelhanças deste agrupamento. Como se fosse o centro de gravidade do grupo. Mas acredito que sua afirmação é verdadeira, pois estamos calculando a distância dos grupos e dos indivíduos para agrupá-los.

Antonio Rodrigues Neto 21:28
Professora, essa fórmula da distância do centroide vale também quando o grupo tiver n pontos (mais do que 2)? Para mim faz sentido que no caso geral, tivesse algum somatório dos n pontos do grupo...
O centróide é o ponto médio central dos grupos criados e ajudarão a encontrar as semelhanças deste agrupamento. Como se fosse o centro de gravidade do grupo. Desta forma, a técnica de clusterização não utiliza vários centroides para cada grupo. Cada grupo só pode ter 1 centróide.

Isabella Montanhal de Araujo 21:21
Porque Dri pra Zé deu 5,38? não deveria ser 5,09?
Olá Isabella. O próprio software realizou o cálculo, portanto provavelmente não deve estar errado.

Mauricio Eneas De Oliveira 22:03
Adriana, pode dar um exemplo pratico do dendograma? com algum case de negocio

O dendograma é um gráfico utilizado na clusterização hierárquica para visualizar a melhor distribuição dos grupos. Qualquer modelo de segmentação de mercado ou de clientes pode usar esse método.

Marco Ruiz 21:28
entao o centroide calcula o centro do grupo em relacao ao novo q vamos testar ? ou seja diff da media, aquele se preocupa com quem esta fora ?
O centróide é o ponto médio central dos grupos criados e ajudarão a encontrar as semelhanças deste agrupamento, além de servir para medir a distância para os outros grupos ou indivíduos. Como se fosse o centro de gravidade dos grupos.

Andrea Garcia Ferraz 20:26
dúvida sobre o dendograma (exemplo)... o ponto 4 (quadradinho da JU) poderia ser considerado um outlier já que ele foi o último a se "transformar" em grupo, com maior distância ou variância?
Não, obrigatoriamente. O cálculo de um outlier obedece a parâmetros diferenciados. Para detectar um outlier você pode calcular a partir da sua base de dados da seguinte forma: 1ºquartil – 1,5x(3ºquartil - 1ºquartil). Também pode utilizar o boxplot. Mas a depender dos dados, o dendograma pode indicar a presença de outlier em função da distância do grupo no gráfico.

Fernando Gonçalves 21:48
Prof, se eu tivesse que explicar o que estou fazendo neste exercício, em poucas palavras o que eu diria?
Está agrupando os grupos, conforme suas distâncias. Assim você consegue determinar qual a melhor segmentação para cada grupo.

Roberto dos Santos Cordeiro Júnior 21:21
Professora, esse valor 5.1 foi arredondado, os outros valores não deviam ter a mesma regra??
Por se tratar de um exercício didático não há problema com arredondamento. Obviamente, em uma análise real deverá ser verificado se houve perda de eficácia na análise em caso de arredondamento.

MARCOS SGARBI 21:06
Se estamos no método do vizinho mais longo, porque agrupamos pelo com a menor distância?
Por que dentro do grupo procurados os participantes mais próximos, ou seja, aqueles que tem características mais homogêneas. A análise sempre se inicia pelos participantes mais próximos.

Franco Iranzi 21:28
Neste caso o Q foi Li e P foi Re, isso é definido arbitrariamente e ou como defino isso?
Sempre iniciar o cálculo pelos indivíduos que tem menor distância e que são agrupados primeiramente na matriz. Depois depende do modelo adotado: single linkage, complete linkage e etc.

Douglas Fernandes de Albuquerque 22:59
Além da padronização Zscore, posso usar a codificação linear, por exemplo, de mínimo valor = -1 a máximo valor = +1?
A padronização Z-score é a mais utilizada, mas não exclui outras formas de padronização. Sempre vai depender dos seus dados e das análises que você pretende realizar.

Carlos Vinicius Taborda Santos 22:57
No exemplo da categorização de usuários para criação de e-mails, quais seriam as variáveis utilizadas? Ex: Número de Acesso vs Número de Compras efetuadas

Depende dos seus objetivos de pesquisa. O que você pretende estudar com a clusterização. Neste ponto é necessário um pouco de experiência e imaginação do analista.

Eduardo Luís Hammes 21:49
Professora, em que momento o valor de N será maior que 1?
No decorrer dos agrupamentos de maneira sistemática, o número de indivíduos nos grupos vai aumentado.

Scott Carrara 22:20
Você comentou que se eleva ao quadrado pra deixar sem o sinal negativo. Utilizar a diferença em módulo também funcionaria pra isso! Por que elevar ao quadrado e não usar modulo?
Olá, Scott. O próprio conceito de variância indica que deve-se elevar ao quadrado. É uma regra matemática amplamente aplicada em cálculos estatísticos como uma notação cientifica usada por todos os analistas de dados e estatísticos.

Vagner Marques 23:06
Quando eu tenho um painel de dados e vou padronizar, eu devo fazê-la pelas médias e desvios das empresas ou ignorar a existência de empresas diferentes ao longo do tempo ?
Depende, pois existem diversos modos de padronização. O Z-score é o mais conhecido deles.

Denis Pereira 22:33
O R2 poderia ser uma metrica de entropia ?
Entropia significa basicamente um nível de desordem de um sistema, todavia a análise de cluster busca ordenar os indivíduos em grupos com característica semelhantes.

Guilherme Santos Vasconcelos 22:27
Professora, eu usei valores randômicos e minha variabilidade ficou oposta. SQG1+SQG2 ficou maior que SQT, e SQ entre grupos ficou negativo. É possível isso?
Acredito que sua análise deve ter algum cálculo equivocado, pois a SQT indica o total da variabilidade intragrupos e intergrupos.

Eliane Chinaglia 22:25
Neste exemplo do excel, não precisa levar e conta o número de elementos de cada grupo?
Não seria necessário, pois a variabilidade do grupo independe da quantidade de elementos de cada grupo. Pode existir um grupo pequeno com alta variabilidade oi vice-versa.

Rodrigo Ribeiro Pereira 22:29
O fato de ter muito menos variáveis não faz com que a soma total já tenda a uma diminuição? Como equolizar isso?
Quanto mais variáveis, melhor é a distinção dos grupos. Todavia, a variabilidade não depende, necessariamente, da quantidade de variáveis.

Renato Cortez Schreiner 22:27
Poderia dividir em grupos e se a variancia de um for baixa, posso formar um grupo definido e dividir o grupo de maior variancia?
Depende dos dados e do seu propósito de pesquisa. Pense, também, que o grupo de maior variância pode ter poucos indivíduos. Assim, deveria analisar para saber se faz sentido dividir os grupos.

Lincoln Amaral Sotto 22:30
O conceito de dividir a variável em grupos para reavaliar a variância é o mesmo que Entropia?/

Entropia é o nível de desorganização de um sistema. A variância mede a dispersão dos dados que podem estar afastados ou próximos.

Bruno Marcos Gonçalves 22:34
Podia mostrar no Excel o SSB e o SSR, ficou um pouco confuso.
Olá, Bruno. Você pode verificar novamente na aula de forma mais pausada. Tente assistir seguindo os mesmos passos no excel com os mesmos dados.

Lucas Fernandes De Oliveira 22:34
Entendi! Posso dizer que o maior R^2 será quando o número de grupos for igual ao número de indivíduos?
Quanto menor o número de grupos (maior variabilidade entre), menor será o R2.

Thiara Patricia Silva Gomes 22:37
No exemplo feito no Excel, não ficou claro quem é a soma de quadrado entre grupos (SQEG)
Olá Thiara, você pode assistir novamente a aula para tirar sai dúvida. Tente assistir seguindo os mesmos passos no excel com os mesmos dados.

Alexsandro Nicácio Siqueira 22:41
Pra eu usar o R^2, preciso primeiro montar o dendograma e analisar cada faixa de corte?
Não. Se você utilizar o método k-means não precisa fazer o dendograma. Basta calcular as variâncias dos grupos formados.

Rafael Viegas De Carvalho Carlos Gomes 22:31
Poderiam demonstrar como calcular o SSB a partir da formula demonstrada no slide?
Olá, Rafael. Você pode verificar novamente na aula de forma mais pausada. Tente assistir seguindo os mesmos passos no excel com os mesmos dados.

Debora Duarte Pinheiro 22:37
É necessário calcular o R^2 com o SSB de cada um dos meus grupos?
Somente se fizer sentido para o estudo e para sua pergunta de pesquisa. Entretanto, o R2 é uma boa medida para verificar como os grupos estão distribuídos.

Lucas Muchon Pavanelli 22:37
O metodo de cluster hierarquico (mais prox, menos prox, etc) tem relacao a resposta do R^2? ou somente a quantidade de grupos?
Não. Você pode calcular o R2 a partir do método k-means. Não precisa ser necessariamente pelo método hierárquico.

Alfredo Barbosa Salerno Junior 22:35
Adriana , ssr seria uma medida de entropia e o r quadrado uma medida de ganho de informação?
Entropia é o nível de desorganização de um sistema. A clusterização busca organizar os dados com características semelhantes.

Rodrigo Squaiella 22:40
professora, entre os metodos R2 e o elbow, se comparar um ao lado do outro, posso dizer que os resultados serão iguais? chegarei na mesma qtde de grupos?
Não necessariamente. Depende muito do conjunto de dados. O melhor sempre é realizar vários modelos para aprender aquele mais viável para responder sua pergunta de pesquisa.

Fabio Ismerim 22:51
Elbow e Silhueta são medidas complementares, ou devo utilizar somente uma delas para definição de numero de cluster?
Não são complementares. Mas sempre é bom utilizar vários métodos para comparar os resultados que melhor atendem sua pergunta de pesquisa.

Márcio de Lima Camargo22:52
Professora, o S deve ter um valor alto. Valor alto em relação a quê? Como sei se o valor está alto?
Sempre comparece seus resultados a partir de várias análises, com diferentes métodos. Assim, você poderá estabelecer parâmetros para verificar como os dados e os grupos se comportam.

Isadora Salvador Rocco 22:29
Essa formula da Soma dos quadrados totais s elevar em consideração o grau de liberdade vira calcula da ANOVA? O grau de liberdade permite a supervisão?
A Anova utiliza o conceito da soma dos quadrados totais e os graus de liberdade. A técnica Anova tem função diferente da análise de cluster.

Danilo José Pereira 22:34
Como esse R^2 se mostra na regressão... O conceito é o mesmo, mas como? Fiquei muito curioso
A regressão também utiliza o conceito de R2 para determinar o grau de explicação estatística da equação em relação ao fenômeno estudado, de forma resumida.

Adriana Melges Quintanilha Weingart 22:39
Dúvida: Tenho 1 base pqna, consigo montar a matriz e o dendograma - consigo definir os grupos, e então avaliar R2 e os outros métodos para escolha do Nr de clusters... Mas, e qdo tiver 1 base mto gde?
Vai depender da sua capacidade computacional para realizar análise com grandes base de dados.

Andre Kenji Yai 23:01
Silhueta é mais indicado para datasets menores?
Pode ser usado com datasets maiores, mas sempre depende do seu poder computacional.

Andre Kenji Yai 22:57
Esse metodo de silhueta e mais pessado computacionalmente. Então mais indicado para datasets pequenos. Certo?
Pode ser usado com datasets maiores, mas sempre depende do seu poder computacional.

Israel Filipe de Melo Tenorio 22:40
No gráfico de elbow, de 8 grupos para 9 aumentou o SSR ? ou é impressão minha ?
Manteve-se praticamente inalterado. Dependendo dos dados isso pode ocorrer. Mas geralmente são variações mínimas de aumento.

Matheus Garcia 22:41
Professora, no gráfico referente a escolha do número de clusters, do 8 para o 9, o eixo Y acaba aumentando mesmo com um cluster a mais, há algum exemplo disto?
Manteve-se praticamente inalterado. Dependendo dos dados isso pode ocorrer. Mas geralmente são variações mínimas de aumento.

ANTONIO THYRSO CORSINO PEREIRA de SOUZA 22:43
Derivar este gráfico ajudaria a encontrar o ponto de corte?
Olá, Antônio. Não conheço esta técnica.

Sergio Fonseca Da Silva 22:48
Prof,. o que ocorrer com os individuos, estão nos grupos apos o corte, são distribuidos no grupos selecionados ?
Isso mesmo. Após o corte, os indivíduos são distribuídos nos grupos selecionados.

João Batista Carvalho Nunes 22:50
Além dos métodos CCC, Elbow e silhueta, o que dizer do método VRC (também chamado Calinski-Harabasz pseudo-F) e do índice Duda-Hart para a determinação do número de clusters?
Existem vários métodos para clusterizar. Nenhum deles exclui o outro. Depende muito da sua base de dados e das suas respostas de pesquisa, bem como do seu poder computacional.

Douglas Fernandes de Albuquerque 22:48
Que função é essa E(R^2)? Função Erro complementar?
Serve para comparar o R2 calculado com aquele esperado por uma distribuição uniforme.

André Nicolas Petridis De Oliveira 22:57
Dri, a quantidade de membros em cada cluster podem atrapalhar o calculo de silhouette?
Não, necessariamente. Vai depender muito do seu poder computacional para calcular grandes base de dados.

Maurício Acconcia Dias 22:46
como calcula o elbow nao tenendi como calcula, como faz? poderia calcular um exemplo?
Olá, Rafael. Você pode verificar novamente na aula de forma mais pausada. Pelo tempo da aula não havia como fazer esse cálculo. Na aula prática acredito que você entenderá melhor.

Maria Cristina Miranda Ramos Chierigatti 22:51
o que é o max (a,b)?
"a" é a distância média entre o ponto e todos os demais pontos do cluster.
"b" é a distância média entre o ponto e todos os pontos do cluster vizinho mais próximo.

Gabriel Ferreira Primo 22:57
Boa noite, não entendi muito bem a parte da conta da Silhoutte. Se na Silhoutte tiver 3 grupos X,Y e Z, voce calcular o ponto X para Y e depois X para Z ou para Y para Z?
Olá, Gabriel. Você pode verificar novamente na aula de forma mais pausada. Acredito que assim você conseguirá tirar suas dúvidas.