PERGUNTAS E RESPOSTAS – MBA EM DATA SCIENCE E ANALYTICS

Disciplina: Unsupervised Machine Learning: Análise de Correspondência Simples e Múltipla I

Data: 13/07/2021

Carla

professor, então não podemos ter missing values no dataset para fazermos a ANACOR? nesse caso, teríamos que tirar a observação com o valor faltante para rodar o modelo?

Dependo do número de missing values pode ser que não rode o modelo. Assim, você deverá voltar a campo para aumentar sua amostra.

Willians Dutra

é correto afirmar que a quantidade de eigenvalues será sempre o menor grau de liberdade ? (min(I-1,J-1))

Oi William, pesquisando sobre a sua dúvida achei um paper interessante que pode te ajudar. "O número máximo de dimensões (eixos nos gráficos) que pode ser estimado é um a menos do que o menor número entre a quantia de linhas ou de colunas. Por exemplo, em uma tabela de contingência com três colunas e cinco linhas, o número máximo de dimensões será dois [mín(linha,coluna) -1]. Após a determinação da dimensionalidade, os resultados podem ser examinados numa representação gráfica, chamada de mapa perceptual" (FÁVERO, L. P. L.; BELFIORE, P. P.; FIGUEIRA JUNIOR, M. F. Utilização da Anacor para a identificação de meios de pagamento em populações de média e baixa renda. SEMINÁRIOS EM ADMINISTRAÇÃO, v. 11, 2006.

Reinaldo

Professor, a Inercia total eu posso calcular usando o Qui2 (31..764) dividindo por 100 ou foi apenas coincidência dar o mesmo resultado?

Oi Reinaldo. A inercia total tem conceito diferente do teste qui-quadrado. Siga a técnica apresentada na sala que não tem erro. Não posso garantir que sempre dividindo por 100 será satisfatório.

Daniel Gabriel Padilha

erro linha 129 > #Matriz A > A - diag(diag(Dl) ^ (-1/2)) %*% (P - lc) %*% diag(diag(Dc) ^ (-1/2)) Error in diag(Dl) : object 'Dl' not found > A Error: object 'A' not found

Olá Daniel. Tente rodar todo script novamente, pois não verificamos qualquer erro. Caso ainda persista a dúvida, envie um email para: monitordsa@pecege.com

Alex Jornada

A Matriz W então seria equivalente a matriz de correlação, só que para variáveis qualitativas??? É isso? No caso, seria não a matriz de correlações, mas a matriz de associações!

O teste qui quadrado busca verificar se existem associações relevantes entre as variáveis qualitativas, enquanto na matriz de correlação busca verificar correlações das variáveis quantitativas. Mas cuidado, correlação é diferente de associação.

Priscila Schall

os eigenvalues estão dando algum erro aqui. a função está retornando o seguinte resultado: [1] 4.829233e-01 2.905629e-01 3.890265e-17

Olá Priscilla. Tente rodar todo script novamente, pois não verificamos qualquer erro. Caso ainda persista a dúvida, envie um email para: monitordsa@pecege.com

Ricardo José Pfitscher
O que é a matriz A?
Olá Ricardo, Poderia ser mais específico. Caso ainda persista a dúvida, envie um email para: monitordsa@pecege.com

ARTUR SOUZA
Professor, no caso do teste qui2 a hipotese escolhida não seria a nula H0. pois as variaveis se associam de forma aleatoria? não é isso que estamos buscando? Obrigado
Quando qui2 calculado é maior que o qui2 críticos, podemos rejeitar a hipótese nula de que as duas variáveis categóricas se associam de forma aleatória, ou seja, existe associação estatisticamente significante, ao nível de significância de 5%, entre o perfil do investidor e o tipo de aplicação financeira.

Thiago Colette Vegi
Como ajustar o gráfico quando temos uma quantidade muito grande de categorias?
Neste caso, acho melhor dividir o dataset, pois pode ficar incompreensível a depender do tamanho da sua tela.

Monica Falcon
Existe algum tipo de análise não supervisionada onde eu possa utilizar variáveis categóricas e métricas na mesma análise?
O ideal é combinar as técnicas, pois sempre terão um viés quanti ou quali.

Fernando
Professor, não usamos critério de Kaiser?
O critério de kaiser é uma técnica utilizada na Análise fatorial.

Adrisson Consoni Floriano
Vamos ver aplicações com dados binários?
Um dos exemplos utiliza a categoria casados x solteiros

Douglas Fernandes de Albuquerque
Se por definição o qui² é unilateral à direita, o teste F também seria?
Suas distribuições do plano cartesiano são bem próximas, portanto, sim.

Rafael Cavalcante De Oliveira
Para a tabela de valores padronizados ajustados, comparamos com o valor de 1,96. Neste caso, sempre com os valores positivos ou em módulo?
Você quis falar da tabela de resíduos padronizados ajustados que deve ter seus valores comparados com 1,96 de modo absoluto e real.

Isis
Professor, o caso que estamos vendo no momento temos uma matriz quadrada, mas ela pode não ser quadrada, certo? Quando ela não for quadrada, não será possível calcular autovetor/autovalor.
O dataset não precisa ser balanceado (tipo: 2 variáveis com 7 categorias ou uma variável com 7 categorias e outra com 10), bem como matriz de contingência não precisa ser quadrada necessariamente.

Sanches
na Anacor o número de categorias tem que ser a mesma para cada variável??
O dataset não precisa ser balanceado (tipo: 2 variáveis com 7 categorias ou uma variável com 7 categorias e outra com 10), bem como matriz de contingência não precisa ser quadrada necessariamente.

Guilherme Yukio Seki
Quanto maior o qui-quadro o observado está mais próximo ao esperado. Então porque se rejeita a hipótese alternativa quando a soma dos qui quadrados é menor que o o qui crítico?
Quando qui2 calculado é maior que o qui2 críticos, podemos rejeitar a hipótese nula de que as duas variáveis categóricas se associam de forma aleatória, ou seja, existe associação estatisticamente significante, ao nível de significância de 5%, entre o perfil do investidor e o tipo de aplicação financeira.

Adrisson Consoni Floriano
Vamos explorar alguma aplicação com dados binários? Parece ser possível utilizar uma abordagem parecido, e na vida real, é muito comum termos dados booleanos
Na ACM um dos exemplos irá utilizar uma variável categórica binária: solteiro x casado.

Barrella
Se tiver alguma variável NA ou NULL na base de dados a soma pelo R considera esse caso? (Se não considerar, não é primordial ter uma validação da soma total vs o número de observação inicial?
Dependendo da quantidade de NA ou NULL pode ser o algoritmo não consiga rodar o modelo, razão pela qual é necessário voltar a campo e coletar mais dados.

Ederson Aguiar De Lima
Prof. Não entendi para que serve o R padronizados. Pods explicar novamente?
Dado que a associação entre as duas variáveis não se dá de forma aleatória, podemos,por meio da análise dos resíduos padronizados ajustados, estudar a relação de dependência entre cada par de categorias.

Guilherme Pereira
Nao existe no R um metodo especifico para os residuos em si, sem ser padronizados ?
Dado que a associação entre as duas variáveis não se dá de forma aleatória, podemos, por meio da análise dos resíduos padronizados ajustados, estudar a relação de dependência entre cada par de categorias.

Gabriel
No meu sjt.xtab ainda aparecem outros número abaixo de cada associação. O que representam?
Olá Gabriel, Poderia ser mais específico. Caso ainda persista a dúvida, envie um email para: monitordsa@pecege.com

Rafael
Porque o lower tail é FALSE?
Olá Rafael, Poderia ser mais específico. Caso ainda persista a dúvida, envie um email para: monitordsa@pecege.com

Ricardo José Pfitscher
Professor vai mostrar um exemplo de interpretação dos dados?
A partir dos exemplos apresentados em sala você pode fazer as suas interpretações.

Gaby

prof sobre analise fatorial, qual a diferença entre rotação de fatores e espelhamento? Poderia falar um pouco mais sobre espelhamento? Teria alguma regra na elaboração de rankings ?

Após a determinação dos fatores e do cálculo das cargas fatoriais, é possível ainda que algumas variáveis apresen- tem correlações de Pearson (cargas fatoriais) intermediárias (nem tão altas, nem tão baixas) com todos os fatores ex- traídos, embora sua comunalidade não seja relativamente tão baixa. Nesse caso, embora a solução da análise fatorial já tenha sido obtida de forma adequada e considerada finalizada, o pesquisador pode, para os casos em que a tabela de cargas fatoriais apresentar valores intermediários para uma ou mais variáveis em todos os fatores, elaborar uma rotação desses fatores, a fim de que sejam aumentadas as correlações de Pearson entre as variáveis originais e novos fatores gerados. Na aula de fatorial esse conceito foi abordado. O espelhamento é uma continuidade da técnica.

Anna Rita

não funcionou instalar e carregar os pacotes Error in table(perfil_investidor$perfil, perfil_investidor$aplicacao) : todos os argumentos devem ter o mesmo comprimento In addition: Warning message:

Oi Anna. Espero que tenha consegui instalar. Mas se não foi possível, realize os seguintes passo: 1 - Tente atualizar os seus pacotes com comando: update.packages(repos='http://cran.rstudio.com/', ask=FALSE, checkBuilt=TRUE); 2 - Se não funcionar e continuar com o problema, tente os passos a seguir; 3 - Tente desinstalar o pacote (eu sei que já foi realizado mas vamos tentar de outra forma): uninstall.packages("magrittr"); 4 - Reinstalar dessa forma: install.packages("magrittr", dependencies = TRUE). Não esqueça de chamar o library (magrittr); 5 - Ao final, verifique se essa biblioteca está marcada na aba "packages" no lado inferior e direito.

Andressa

Estou com erro para instalar alguns pacotes e aparece a seguinte msg: ERROR to lock directory (diretorio) for modifying. Try to removing (diretorio)

Oi Andress. Espero que tenha consegui instalar. Mas se não foi possível, realize os seguintes passo: 1 - Tente atualizar os seus pacotes com comando: update.packages(repos='http://cran.rstudio.com/', ask=FALSE, checkBuilt=TRUE); 2 - Se não funcionar e continuar com o problema, tente os passos a seguir; 3 - Tente desinstalar os pacotes: uninstall.packages("....."); 4 - Reinstalar dessa forma: install.packages("....", dependencies = TRUE). Não esqueça de chamar o library (....); 5 - Ao final, verifique se essa biblioteca está marcada na aba "packages" no lado inferior e direito.

Catharina

A tabela de contingência pode ser considerada um diagrama de venn?

Os conceitos são próximos, mas os objetivos são diferentes.

Patrícia

A quantidade de categorias das variáveis analisadas influenciam o percentual da percepção?

Com certeza, considerando que vão fazer parte do cálculo realizado pelo algoritmo de forma concomitante.

Catharina

Se a soma das dimensões = 100% no mapa perceptual explica todo o comportamento, quando o mapa trouxer o total de 52% nas duas dimensões, como é o caso do exemplo mostrado agora, como pode ser lido?

Por tratar-se de uma ACM o gráfico não consegue capturar todas as dimensões.

Diego
Porque o 2o mapa perceptual explica sem dúvidas o tipo de investidor vs a aplicação e o primeiro não?
Depende das variáveis que foram utilizadas para realizar a análise. Quanto mais variáveis, melhor você conseguirá estabelecer suas interrelações.

Douglas Pasquali Pedroso
Boa noite, no gráfico, o ponto representa a posição da variável, o retângulo com escrito é apenas um label , certo ou não? Obrigado :)
Na verdade é a relação de interdependência das variáveis.

Marcos Henrique Da Silva
Professor, falar mais do exemplo sobre NPS com ANACOR?
NPS é quanti. Mas quando você estabelece faixas de clientes promotores ou detratores passa a ser quali, logo você poderá utilizar ANACOR combinando com outras variáveis.

Vera
Variáveis do tipo NPS então são consideradas qualitativas?
NPS é quanti. Mas quando você estabelece faixas de clientes promotores ou detratores passa a ser quali.

Francisco Felipe Gomes De Sousa
Se ao utilizar na base "perfil_investidor" eu tivesse um investidor com mais um tipo de aplicação, por exemplo, Gabriela tem Poupança e CDB isso afetaria a minha ANACOR de que forma?
Dependeria de como você arrumou essa observação que poderia se repetir com nomenclatura diferente (Tipo: Gabriela1, Gabriela2...).

Israel
No caso de um estudante com investimentos em 2 tipos de investimento, seria uma ACM em vez de Anacor?
Neste caso continuaria sendo uma ANACOR. A observação poderia se repetir com nomenclatura diferente (Tipo: Estudante1, Estudante2...). Entretanto, depende muito do que o pesquisador busca com a análise.

Kauê
Professor, é uma abordagem válida utilizar um modelo de clusterização para gerar uma variável catégirca a partir das variáveis métricas e levá-la em conta numa análise de correspondência?
Se existir um embasamento teórico forte neste sentido pode ser utilizado. Mas numa primeira análise o resultado da clusterização é classificar as observações e não criar variáveis.

Guilherme Sampaio Bacelo
Em caso de um formulário, poderia substituir para a pessoa preencher com números que representassem a categoria? Para assim evitar algum erro de português, e danificar a base
Acho melhor a pessoa marcar qual a categoria Ex: ( ) poupança ou ( ) CDB.

Mariana Borges Maié
Podemos supor que na ACM o percentual que as variáveis consegue explicar é menor?
ACM não tem objetivo de explicar fenômenos, mas apenas classificar as observações.

Eduardo
Qual o limite de dimensões que podemos ter na ANACOR E ACM? A quantidade de variáveis tambem?

Com base na determinação das coordenadas de cada categoria, pode ser construído um mapa perceptual com m dimensões. Embora essa possibilidade seja matematicamente possível, apenas as duas primeiras dimensões (m = 2) são geralmente utilizadas para a elaboração da análise gráfica, o que gera um mapa perceptual conhecido por biplot.

Eduardo

No mapa perceptual, possivelmente temos uma 3ª dimensão não plotada?

Com base na determinação das coordenadas de cada categoria, pode ser construído um mapa perceptual com m dimensões. Embora essa possibilidade seja matematicamente possível, apenas as duas primeiras dimensões (m = 2) são geralmente utilizadas para a elaboração da análise gráfica, o que gera um mapa perceptual conhecido por biplot.

Paula

Professor, o p-valor me confundiu. Ele é um teste extra, além da comparação dos qui-quadrados crítico e calculado, ou é uma validação complementar a ele?

Ele verifica se a associação entre as variáveis é estatisticamente significante. Assim, ele é uma das fases do qui-quadrado.

Alexandre

Pergunta relacionada com a da Samya: faz sentido clusterizar e depois aplicar ACM se há variáveis quanti e quali?

Nesse caso acredito ser melhor dividir o banco de dados e rodar as técnicas correspondentes.

Vitor Hugo Miro

Qual é a medida nos eixos do mapa perceptual?

São as duas inércias principais parciais para construção do mapa perceptual com duas dimensões (biplot), é importante enfatizar que 100% da inércia principal total estão representados no mapa bidimensional.

Lucas Nogueira

Prof. as técnica de ML não supervisionado podem ser interpretadas como técnicas que visam realizar a transformação do espaço de dados de forma a melhor explicá-lo?

É uma forma de interpretação, considerando que nas técnicas não supervisionadas não se busca estabelecer relações de predição entre as variáveis.