PERGUNTAS E RESPOSTAS – MBA EM DATA SCIENCE E ANALYTICS

Disciplina: Unsupervised Machine Learning: Clustering I

Data: 08/06/2021

Eduardo Silva Costa dos Santos 19:32
Professora, a Lei de Benford poderia ser usada para analise não supervisionada?
Eu, particularmente, nunca ouvi falar. Qual seria a ideia?

Fernando Gonçalves 23:01
professora, pegar variáveis categóricas (produtos por exemplo) distribuindo o valor faturado para cada uma é uma forma mais interessante de fugir da ponderação arbitrária?
Não ficou muito claro para mim. Se você está olhando o valor faturado, não tem categoria aí...

Murilo Urssi Malek-Zadeh 22:59
Se eu padronizar minhas variaveis, eu nao resolvo o problema do comeco da aula? O problema que nao pode usar variavel categorica (sendo 0 e 1)...
Infelizmente não. Se colocamos 0 e 1 e padronizamos... continuará sendo apenas dois números e continua sendo uma distância arbitrária.

Sabrina Portela Costa Ferreira 22:57
Eu não entendi o que o S poderia me ajudar na hora de plotar os números de grupos pra identificar qual seria a melhor quantidade, sendo que eu calculo um S por observação e a distância entre eles.
O S é exatamente isso, no entanto o método da Silhueta é a média dos S´s de todos os pontos. Então quanto maior for, melhor o agrupamento foi realizado.

Gabriela Werner Ceschini 22:57
Boa noite! Vou refazer a pergunta de outra forma: há estudos da função R^2, derivada, concavidade, que possam sugerir os pontos de corte do dendograma? O CCC é um método desse tipo? Obrigada!
Sim, o R2 pode me auxiliar para definir uma sugestão do número de grupos, assim como o CCC.

Lúcio Nunes Carmo de Oliveira 22:46
Tem algum pacote em R pra esse CCC?
Sim: https://stats.stackexchange.com/questions/212293/cubic-clustering-criterion-using-r-update

Nara De campos Costa 20:33
Professora ou Monitores, se eu tenho uma pesquisa sindicalizada que tem os dados extrapolados para o total da população (como o TGI da Kantar), consigo rodar uma análise de cluster?
Sim, tudo depende de como você definiu as linhas e as informações que utilizará para agrupar. E as limitações que isso pode causar na sua interpretação... conseguir você consegue, só precisa ver se faz sentido.

Kauê Bonato De Araujo 20:03
Professora, poderia explicar sobre a maldição da dimensionalidade? Existe um limite de número de variáveis a qual os pontos da minha amostra ficam "equidistântes"?
Vocês verão isso em outras disciplinas, como análise fatorial e componentes principais.

Rebeca Dieb Holanda Silva 19:56
Existe ponderação arbitrária na Teoria de Resposta ao Item?
Não sei, não estudo TRI, infelizmente.

Hernandes Matias Junior 19:55
No processo de aprendizagem de modelos de clusterização, nós vamos estudar autoencoders?

Não abordaremos nessa disciplina, mas eu acredito que será abordado no próximo módulo.

João Castro 23:06
poderia fazer encoding das categoricas (dummies) para determinar os cluster?
Não me fez muito sentido... eu creio que não entendi sua pergunta. Dummies utilizamos para variáveis categorias, no entanto fica sendo uma distância arbitrária... e, sendo assim, não é recomendado utilizar variáveis categóricas.

Cleyton Nunes de Oliveira 19:51
Duvida: Podemos analisar a "proximidade" das frequencias das variáveis categoricas usando os "bins" do histograma sem incorrer em ponderação arbitrária?
Eu creio que não entendi a pergunta. Para variáveis categóricas não temos histograma, temos gráfico de barras... e isso é uma sumarização da base de dados como todo. A análise de cluster tenta agrupar indivíduos a indivíduos... não consegui ver como as frequências no gráfico de barras poderiam ajudar nisso.