INSTITUTO INFNET
MIT BIG DATA
Analytics com R
Prof.: Cassius Figueiredo

AVALIAÇÃO ALGORITMOS

INSTRUÇÕES

- Cada grupo de alunos deve apresentar um relatório com as respostas para os dois problemas propostos.
- Todas as informações disponíveis estão abaixo e, caso seja necessário, os grupos deverão complementar as informações por sua conta, deixando claramente indicadas as referências.
- A abordagem de solução do primeiro problema está definida na questão, já para o segundo problema a abordagem é livre, porém apenas o uso de linguagem R será permitido.
- O relatório final deverá ser acompanhado dos códigos fartamente comentados para documentação das soluções apresentadas.

1) Exercícios de Classificação (Clusterização - Agrupamentos)

BASE
tripadvisor_review.csv

FONTE
Renjith, Shini, A. Sreekumar, and M. Jathavedan. 2018

CENÁRIO
Uma agência de viagens quer começar a usar inteligência de dados para mapear as necessidades de seus clientes e fazer campanhas de marketing direcionadas. Para isso, contratou uma empresa que fez uma coleta de informações do site TripAdvisor e gerou a base de dados que vocês irão trabalhar. Esta base contém variáveis que representam IDs de usuários e as médias das avaliações para algumas localidades na Ásia para os seguintes itens:

DESCRIÇÃO DOS DADOS

- User ID : User ID único
- Category 1 : Feedback médio dos usuários para galerias de arte
- Category 2 : Feedback médio dos usuários para casas noturnas
- Category 3 : Feedback médio dos usuários para bares
- Category 4 : Feedback médio dos usuários para restaurantes
- Category 5 : Feedback médio dos usuários para museus
- Category 6 : Feedback médio dos usuários para resorts
- Category 7 : Feedback médio dos usuários para piqueniques
- Category 8 : Feedback médio dos usuários para praias
- Category 9 : Feedback médio dos usuários para cinemas
- Category 10 : Feedback médio dos usuários para espaços religiosos

As notas são dadas segundo o critério abaixo:

- Excelente (4)
- Muito bom (3)
- Médio (2)
- Ruim (1)
- Horrível (0)

EXERCÍCIO

Use um algoritmo de agrupamentos (clusterização) que apresente uma segmentação destes dados. Tente interpretar estas informações com base em seus conhecimentos, exemplos:

- Quais grupos de itens agradam mais?
- Poderíamos direcionar uma campanha promovendo o continente para grupos de clientes que se interessem por determinados tipos de atrações do local?
- Segundo estes clientes, quais são os maiores atrativos do continente?

Fundamente suas decisões, não se prenda ao algoritmo visto em sala, porém apenas algoritmos de clusterização serão aceitos. Documente seu processo de análise em um documento que será entregue à agência.

2) Análise de dados

BASE
online_shoppers_intention.csv

FONTE
Sakar, C.O., Polat, S.O., Katircioglu, M. et al. Neural Comput & Applic (2018)

CENÁRIO
Vocês devem modelar a previsão a intenção de compra de usuários online. O conjunto de dados oferece informações de sessões de compra, sem tendência.

DESCRIÇÃO DOS DADOS
O conjunto de dados consiste em 10 atributos numéricos e 8 categóricos. O atributo binário "Revenue" indica se a sessão terminou em uma compra efetiva ou não.

- "Administrative", "Administrative Duration", "Informational", "Informational Duration", "Product Related" and "Product Related Duration" represent the number of different types of pages visited by the visitor in that session and total time spent in each of these page categories. The values of these features are derived from the URL information of the pages visited by the user and updated in real time when a user takes an action, e.g. moving from one page to another.

- The "Bounce Rate", "Exit Rate" and "Page Value" features represent the metrics measured by "Google Analytics" for each page in the e-commerce site. The value of "Bounce Rate" feature for a web page refers to the percentage of visitors who enter the site from that page and then leave ("bounce") without triggering any other requests to the analytics server during that session. The value of "Exit Rate" feature for a specific web page is calculated as for all pageviews to the page, the percentage that were the last in the session. The "Page Value" feature represents the average value for a web page that a user visited before completing an e-commerce transaction.

- The "Special Day" feature indicates the closeness of the site visiting time to a specific special day (e.g. Mother’s Day, Valentine's Day) in which the sessions are more likely to be finalized with transaction. The value of this attribute is determined by considering the dynamics of e-commerce such as the duration between the order date and delivery date. For example, for Valentina’s day, this value takes a nonzero value between February 2 and February 12, zero before and after this date unless it is close to another special day, and its maximum value of 1 on February 8.

- The dataset also includes operating system, browser, region, traffic type, visitor type as returning or new visitor, a Boolean value indicating whether the date of the visit is weekend, and month of the year.

EXERCÍCIO

De posse destes dados crie um modelo que indique a propensão de compra de um novo cliente com este mesmo conjunto de atributos. Documente seu processo de análise em um relatório com as escolhas, justificativas, testes, gráficos explicativos e qualquer outra informação que ache pertinente.