ME732 / ME921 / MI407: Análise Multivariada e Aprendizado Não-Supervisionado de Máquinas, atividade prática I

Guilherme Ludwig
gvludwig@ime.unicamp.br

13 de Abril de 2021

- A atividade é individual. A consulta de material eletrônico e livros será permitida, mas as respostas devem ser desenvolvidas inteiramente na solução entregue, e devem estar completas. Referencie o trabalho citado, mas desenvolva as soluções na atividade. Explique resultados com suas próprias palavras.
- A solução deve ser desenvolvida em detalhe. Todas as contas devem estar explicadas e desenvolvidas. Não é preciso fazer operações elementares manualmente mas as contas devem estar claramente indicadas.
- Cada aluno deverá ser responsável por garantir que não há cópia do seu trabalho. Indícios de plágio podem levar à anulação da atividade de todos os envolvidos. Use seu bom senso para guiar as conversas com seus colegas.
- A atividade deve ser entregue através do Moodle, em um único documento em formato `PDF`.
  - As páginas da sua solução devem estar numeradas, e as linhas também devem estar numeradas. Use o pacote no LaTeX `\usepackage{lineno}` e `\linenumbers`. Isso é importante pois me ajudará a dar *feedback* à sua solução.
  - Se você usar o `Word` ou `LibreOffice`, numere linhas e páginas e imprima o documento em formato `PDF` antes de enviá-lo.
- Use a linguagem que você preferir, de preferência com o pacote `listings` do LaTeX para exibir código: `R`, `C`, `python`, `julia`, ou outra qualquer. Todos os códigos devem estar listados no final da atividade, com instruções de execução (à parte das linguages listadas).
- Se você quiser, pode usar `Rmarkdown`, `Jupyter notebooks` ou o que preferir para mesclar código e texto.
- Organize os resultados mas não deixe de entregar soluções parciais, se você não conseguiu finalizar o exercício.
- **Prazo de entrega:** 20/04/2021, às 23:59, via Moodle. Vou tentar configurar o Moodle para permitir múltiplos envios, a última versão enviada é a final.

O conjunto de dados "`Atividade1.csv`" foi extraído na data de hoje do site `https://www.whiskybase.com/whiskies/brands` e corresponde a notas, dadas por usuários, a diferentes marcas de whisky. As marcas estão organizadas por país de origem, e classificação de especialistas (*WB Ranking*, A até G), mas vamos usar para clustering somete as notas dadas pelos usuários, isto é, a coluna `Rating`.

O conjunto de dados tem um problema de balanço de classes, e balanço de avaliações. Por exemplo, o whisky da marca "A Drop of the Irish" tem 24 whiskies na sua avaliação (possivelmente correspondento a 24 anos de engarrafamento). A nota média por ano pode variar, bem como o número de notas por ano.

Por outro lado, whiskies famosos (e.g. "Jack Daniel's") possuem mais de 400 votos (e 679 tipos de whisky). A menos que exista alguma distorção exagerada, podemos considerar o whisky "médio"de cada marca como sendo o representante da nota dada pelo site.

Minhas sugestões:

- Considere usar apenas um país para fazer sua análise. Você pode escolher o país, mas dê preferência a países com um número grande de whiskies (por exemplo, New Zealand tem 22, um ponto de corte razoável). Quanto maior o tamanho do corte que você fizer, melhor; mas a análise de um conjunto de dados muito grande pode ser problemática.

- Considere descartar whiskies com um número pequeno de votos, por exemplo considerando apenas entradas com pelo menos 5 votos (faça isso para países com pelo menos uma centena de whiskies listados, a saber Suiça, França, Japão, Canadá, Irlanda, Alemanha, Estados Unidos e Escócia).

- Considere alguma correção nas notas para incorporar o número de votações. Por exemplo, a IMDB (o site de filmes) corrige *rankings* usando como nota suavizada

$$\hat{u}_j^{(s)} = \frac{n_j \hat{u}_j + 1}{n_j + 2} \times 100\%,$$

  em que n_j é o número de avaliadores, e u_j a nota média entre 0 e 1.

Atividade:

- Utilize o conjunto de dados e a descrição do problema para preparar um pequeno tutorial de clustering, usando clustering tipo *PAM* (partition around medoids). Você pode usar o pacote `cluster` no `R`, ou ferramenta equivalente na sua linguagem de preferência.

- Introduza o método e faça a análise do número de clusteres, inclusive examinando as silhuetas dos clusteres.

- Produza um relatório de no máximo 5 páginas, em formato de artigo (não é preciso capa, índice etc.). O relatório também será avaliado por organização e apresentação.

- Como os seus clusteres se comportam com relação ao `WB Ranking` informado pela página?

```
> # Se tiver algum erro de leitura dos dados, o encoding que eu usei foi utf8 do Linux
> # whiskies <- read.csv("atividade1.csv", fileEncoding = "utf8")
> whiskies <- read.csv("atividade1.csv")
```