INF0613 – Aprendizado de Máquina Não Supervisionado

Trabalho 2 - Redução de Dimensionalidade

O objetivo deste trabalho é exercitar o conhecimento de técnicas de redução de dimensionalidade. Essas técnicas serão usadas tanto para obtenção de características quanto para visualização dos conjuntos de dados. Usaremos a base de dados `speech.csv`, que está disponível na página da disciplina no Moodle. A base contém amostras da pronúncia em inglês das letras do alfabeto.

Informações Importantes

- **Prazo de entrega:** 12 de setembro de 2021 (Domingo), até às 23h55.
- **Forma de entrega:** Deverá ser submetido um arquivo comprimido no formato **.zip** via Moodle contendo:
  - Arquivo `inf0613-trabalho2.Rmd` com as respostas das atividades, e
  - Arquivo no formato `PDF` gerado pelo `Knit`, gerado a partir do arquivo `inf0613-trabalho2.Rmd` respondido.
- **Pontuação:** Este trabalho será pontuado de 0 a 10, e corresponderá a 30% da nota final.
- Na página da disciplina no Moodle, fornecemos um arquivo `inf0613-trabalho2.Rmd` que contém o template do trabalho, seu uso é obrigatório.
- O arquivo `base.zip` contém a base de dados `speech.csv`. Baixe o arquivo e o descomprima antes de começar o trabalho. O arquivo `.csv` não deve ser alterado diretamente, ou seja, todas as alterações devem ser feitas nos objetos lidos no script.
- **Atenção:** O atributo da coluna 618 deve ser transformado em `factor` ou `string`. Esse atributo contém a classe de cada amostra e não deve ser usado nos cálculos de redução de dimensionalidade.
- *Dica para esta tarefa:* revise os slides das aulas, todos os comandos em R necessários estão exemplificados neles.
- Teste o seu código antes de submeter. Códigos com erros sintáticos serão penalizados.
- Funções que não atendem às especificações serão penalizadas.
- Submissões com formatos diferentes dos especificados não serão corrigidas.
- Para os trabalhos feitos em grupo, apenas um membro do grupo deve enviar a solução. Os **nomes completos** dos integrantes devem constar no cabeçalho de cada arquivo a ser submetido no local indicado.

Atividade 1 – Análise de Componentes Principais (*3,5 pts*)

Durante a redução de dimensionalidade, espera-se que o poder de representação do conjunto de dados seja mantido, para isso é preciso realizar uma análise da variância mantida em cada componente principal obtido. Use função `prcomp`, que foi vista em aula, para criar os autovetores e autovalores da base de dados. Não use a normalização dos atributos, isto é, defina `scale.=FALSE`. Em seguida, use o comando `summary`, analise o resultado e os itens a seguir:

a) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos `80%` do total?

b) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos `90%` do total?

c) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos `95%` do total?

d) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos `99%` do total?

e) Faça um breve resumo dos resultados dos itens *a)-d)* destacando o impacto da redução de dimensionalidade.

Atividade 2 – Análise de Componentes Principais e Normalização (*3,5 pts*)

A normalização de dados em alguns casos, pode trazer benefícios. Nesta questão, iremos analisar o impacto dessa prática na redução da dimensionalidade da base de dados `speech.csv`. Use função `prcomp` para criar os autovetores e autovalores da base de dados usando a normalização dos atributos, isto é, defina `scale.=TRUE`. Em seguida, use o comando `summary`, analise o resultado e os itens a seguir:

a) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos `80%` do total?

b) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos `90%` do total?

c) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos `95%` do total?

d) Qual o menor número de componentes, tal que a variância acumulada seja pelo menos `99%` do total?

e) Quais as principais diferenças entre a aplicação do PCA nesse conjunto dados com e sem normalização?

f) Qual opção parece ser mais adequada para esse conjunto de dados? Justifique sua resposta.

Atividade 3 – Visualização a partir da Redução (*3,0 pts*)

Nesta atividade, vamos aplicar diferentes métodos de redução de dimensionalidade e comparar as visualizações dos dados obtidos considerando apenas duas dimensões. Lembre de fixar uma semente antes de executar o T-SNE.

a) Aplique a redução de dimensionalidade com a técnica PCA e gere um gráfico de dispersão dos dados. Use a coluna `618` para classificar as amostras e definir uma coloração.

b) Aplique a redução de dimensionalidade com a técnica UMAP e gere um gráfico de dispersão dos dados. Use a coluna `618` para classificar as amostras e definir uma coloração.

c) Aplique a redução de dimensionalidade com a técnica T-SNE e gere um gráfico de dispersão dos dados. Use a coluna `618` para classificar as amostras e definir uma coloração.

d) Qual técnica você acredita que apresentou a melhor projeção? Justifique.