Trabalho 2 (Individual ou em Dupla)

Módulo 2 – Aprendizado de Máquina Não Supervisionado

O objetivo deste trabalho é exercitar o conhecimento de técnicas de redução de dimensionalidade vistos em sala. Essas técnicas serão usadas tanto para extração de características quanto para visualização dos conjuntos de dados. Este trabalho está dividido em duas atividades: *Análise de Componentes Principais* e *Visualizando a partir da Redução*. Essas atividades são descritas com maiores detalhes na sequência. A base de dados a ser utilizada neste trabalho está disponível na página da disciplina no Moodle.

Atividade 1: Análise de Componentes Principais

Como foi visto na primeira aula da disciplina, é importante conhecer as características dos dados de uma base disponíveis para a mineração dos mesmos. Por isso, o objetivo desta atividade é reduzir a dimensionalidade dos dados apresentados na **base de dados `frogs.csv`** de forma que seja mantido o poder de representação dos dados. A seguir são listadas as análises que devem ser feitas.

1. *Obtenção dos autovetores e autovalores:* como o método de Análise de Componentes Principais (PCA) baseia-se na obtenção de autovetores e autovalores da matriz de covariância, espera-se que seja utilizado o método PCA para obtenção desses autovetores e autovalores. Para isso, escreva uma função com a seguinte especificação:

   Entrada: `data-frame` contendo o conjunto de atributos de cada amostra;

   Aplique o método para redução de dimensionalidade PCA;

   Saída: um objeto do tipo `princomp` ou `prcomp`.

2. *Escolha do número de dimensões para redução:* 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. Nesta atividade, espera-se que sejam mantidas K dimensões de forma que a soma dos quadrados do desvio padrão dos K primeiros autovetores totalizem pelo menos $x\%$ da variância total. Sua função deve seguir a seguinte especificação:

   Entrada: um objeto do tipo `princomp` ou `prcomp` e um valor x que indicará o mínimo de variância desejada;

   Selecione o menor K possível, tal que o valor mínimo da variância seja respeitado;

   Saída: `data-frame` contendo o K autovetores escolhidos.

   Além disso, indique no arquivo que você irá submeter:

   (a) O menor valor de K, tal que a soma dos quadrados dos desvios padrão totalize pelo menos 90% da variância;

   (b) O menor valor de K, tal que a soma dos quadrados dos desvios padrão totalize pelo menos 95% da variância;

   (c) O menor valor de K, tal que a soma dos quadrados dos desvios padrão totalize pelo menos 99% da variância.

Atividade 2: Visualizando a partir da Redução

O objetivo desta atividade é comparar as visualizações dos dados obtidos a partir da redução de dimensionalidade utilizando as técnicas PCA e T-sne. Para isso, escreva as funções com as seguintes especificações:

Função 1:

Entrada: `data-frame` contendo o conjunto de atributos de cada amostra;

Projete em um gráfico de dispersão os dados reduzidos para 2 dimensões utilizando o PCA;

Saída: – (esta função não precisa ter um retorno específico), mas um gráfico deve ser gerado.

Função 2:

Entrada: `data-frame` contendo o conjunto de atributos de cada amostra;

Projete em um gráfico de dispersão os dados reduzidos para 2 dimensões utilizando o T-sne;

Saída: – (esta função não precisa ter um retorno específico), mas um gráfico deve ser gerado.

Escolha dentre os dados categóricos da (*species, genus* e *family*) para colorir os gráficos. Não use dados categóricos para os cálculos de redução de dimensionalidade. Escreva, como comentário, qual técnica você acredita que apresentou a melhor projeção e por qual motivo. Lembre de fixar uma semente antes de usar o T-sne, e de especificar no arquivo qual a semente foi usada.

Considerações Finais

- Você não deve remover qualquer linha já existente nos arquivos das bases de dados.
- As funções que serão criadas neste trabalho não precisam ser genéricas, e devem assumir que a base de dados usada é a **base de dados `frogs.csv`**.
- Teste o seu código antes de submeter. Códigos com erros sintáticos ou que retornam tipos diferentes daqueles especificados na questão serão penalizados.
- Na página da disciplina no Moodle, fornecemos um arquivo `trabalho2.R` que contém um esqueleto da entrega da tarefa, seu uso é facultativo.
- Para os trabalhos feitos em dupla, apenas um membro da dupla deve enviar a solução. Os nomes dos membros devem constar no cabeçalho de cada arquivo "`.R`" a ser submetido.
- Salve os arquivos utilizando o mesmo nome, e os envie no sistema Moodle, clicando no link "`Trabalho 2`" da Seção "`Avaliações`". Clique em "`Adicionar tarefa`", anexe os arquivos e, por fim, clique em "`Salvar mudanças`". Você voltará para a tela da atividade e deverá constar o status "`Enviado para avaliação`". A qualquer momento, antes do prazo final de submissão, você pode alterar sua submissão clicando em "`Editar envio`".

Prazo de entrega: 08 de Abril de 2019 (Segunda-Feira), até às 23h55.

Forma de entrega: Deverá ser submetido um arquivo com `trabalho1.R` contendo o código especificado em cada atividade via Moodle:

`https://moodle.lab.ic.unicamp.br/moodle/course/view.php?id=315`

Pontuação: Este trabalho será pontuado de 0 a 10, e corresponderá a 30% da nota final.