Instituto Federal de Educação, Ciência e Tecnologia – IFB
Campus Taguatinga
Curso Bacharelado em Ciências da Computação
Disciplina: Aprendizagem de Máquina
professor Lucas Moreira

Trabalho 04

O objetivo do segundo trabalho é familiarizar o(a) aluno(a) com o algoritmo de agrupamento não-supervisionado *K-means*.

Nesse trabalho, o(a) aluno(a) irá implementar um programa computacional que realiza o agrupamento de um conjunto de dados em classes com características similares.

Problema

Aparelhos eletrodomésticos no Brasil são classificados pelo INMETRO de acordo com seu consumo e eficiência energética, sendo aparelhos de classe A os mais eficientes e econômicos, e aparelhos da classe G os de maior consumo e menos eficientes.

Deseja-se fazer a mesma classificação para automóveis de passeio movidos a gasolina, identificando-os pelo seu consumo e pela emissão de carbono. Dessa forma, foram coletados dados de vários modelos de carros, de diferentes fabricantes.

O(a) aluno(a) deverá desenvolver um script em Matlab/Octave para agrupar esses modelos de carro segundo seus dados de consumo e emissão de carbono, para que o INMETRO possa rotulá-los posteriormente.

O número ótimo de classes (K) deverá ser determinado quantitativamente pelo estudante, justificando a escolha. Tal justificativa deve ser colocada, na forma de comentário, no script principal do trabalho.

O trabalho deve ser constituído por 3 (três) arquivos Matlab/Octave (com extensão .m) distintos. São eles:

- O script principal *trabalho4.m*, que deverá ser executado para realizar o agrupamento;
- Um script de função para inserir o rótulo das amostra em cada iteração;
- Um script de função para cálculo dos centroids em cada iteração.

O arquivo "*trabalho4.m*" pode ser usado para realizar o teste quantitativo do número ótimo de classes K.

Após o agrupamento com o número ótimo de classes, deverá ser disponibilizado o valor médio de consumo e emissão de carbono (centroids) de cada classe, para documentação junto ao INMETRO.

Dados

Os dados usados nesse trabalho são provenientes de um estudo sobre emissão de carbono de modelos de automóveis de passeio nos EUA.

Esses dados estão disponíveis no arquivo "data.mat" contendo 80 (oitenta) amostras de 3 (três) variáveis. A primeira delas, "model", contém o modelo e fabricante de um determinado automóvel, na forma de uma *string* de caracteres. Como o número de caracteres de cada elemento não é constante, não é possível armazenar tais dados de forma matricial, por isso os textos são armazenados por um vetor de células (*cell*). Consulte a documentação do Matlab/Octave para saber como manipular dados armazenados como células.

A segunda variável, "carbon", contém a emissão de carbono do modelo de automóvel medido em gramas de CO_2 emitido por litro de gasolina consumido. A terceira variável, "millage", contém o consumo do carro, medido em km/l.

Os três vetores contém o mesmo ordenamento, ou seja, o i-ésimo elemento dos vetores "carbon" e "millage" correspondem aos dados do i-ésimo elemento do vetor "model".

Entrega

Os 3 (três) arquivos citados acima, devem ser compactados em um único arquivo .zip e enviados para o endereço de e-mail lucas.moreira@ifb.edu.br até às 23h55min do dia 25 de novembro de 2018.

O trabalho é individual, e deve ser entregue um arquivo compactado, contendo os 3 (três) scripts solicitados, por aluno(a).

Cópias de trabalhos receberão nota 0 (zero).

Em caso de dúvidas, pede-se que entrem em contato pelo endereço de e-mail acima.