Trabalho Final (Em Dupla ou Trio)

INF-0613 – Aprendizado de Máquina Não Supervisionado

O objetivo deste trabalho é exercitar o uso de algoritmos de agrupamento. Neste trabalho, vamos analisar diferentes atributos de carros com o objetivo de verificar se seus atributos são suficientes para indicar um valor de risco de seguro. O conjunto de dados já apresenta o risco calculado no campo *symboling* indicado na Tabela 1. Quanto mais próximo de 3 maior o risco. O conjunto de dados que deve ser usado está disponível na página do Moodle com o nome `imports-85.data`. Este trabalho está dividido em três atividades: *Análise e Preparação dos Dados*, *Agrupamento com o K-means* e *Agrupamento com o DBscan*. Além do código implementado, neste trabalho também deverá ser entregue um relatório contendo as análises feitas para as duas atividades que são descritas com maiores detalhes na sequência.

Tabela 1: Nome para acesso do atributo no data-frame, nome do atributo, tipo de dado, valores possíveis para cada atributo.

Nome	Atributo	Tipo	Valores
V1	symboling	integer	-3, -2, -1, 0, 1, 2, 3
V2	normalized-losses	numeric	[65.00,256.00]
V3	make	string	22 marcas
V4	fuel-type	string	diesel, gas
V5	aspiration	string	std, turbo
V6	num-of-doors	string	four, two
V7	body-style	string	hardtop, wagon, sedan, hatchback, convertible
V8	drive-wheels	string	4wd, fwd, rwd
V9	engine-location	string	front, rear
V10	wheel-base	numeric	[86.60,120.90]
V11	length	numeric	[141.10,208.10]
V12	width	numeric	[60.3,72.3]
V13	height	numeric	[47.80,59.80]
V14	curb-weight	numeric	[1488.00,4066.00]
V15	engine-type	string	dohc, dohcv, l, ohc, ohcf, ohcv, rotor
V16	num-of-cylinders	string	eight, five, four, six, three, twelve, two
V17	engine-size	numeric	[61.00,326.00]
V18	fuel-system	string	1bbl, 2bbl, 4bbl, idi, mfi, mpfi, spdi, spfi
V19	bore	numeric	[2.54,3.94]
V20	stroke	numeric	[2.07,4.17]
V21	compression-ratio	numeric	[7.00,23.00]
V22	horsepower	numeric	[48.00,288.00]
V23	peak-rpm	numeric	[4150.00,6600.00]
V24	city-mpg	numeric	[13.00,49.00]
V25	highway-mpg	numeric	[16.00,54.00]
V26	price	numeric	[5118.00,45400.00]

Atividade 1: Análise e Preparação dos Dados

O conjunto de dados é composto por 205 amostras com 26 atributos cada descritos na Tabela 1. Os atributos são dos tipos `factor`, `integer` ou `numeric`. O objetivo desta etapa é a análise e preparação desses dados de forma a ser possível agrupá-los nas próximas tarefas. Nessa atividade, você deve:

1. *Selecionar os atributos:* queremos trabalhar com atributos numéricos (dos tipos `integer` ou `numeric`). Portanto, você deve analisar cada atributo (Tabela 1) e verificar se os atributos não números são descritivos para a realização dos agrupamentos. Caso um dos atributos não numéricos seja necessário, use a técnica do *one hot encoding* para transformá-lo em numérico. Não utilize os atributos *symboling* e *make* para os agrupamentos subsequentes;

2. *Tratar dados incompletos:* amostras incompletas deverão ser tratadas, e você deve escolher a forma que achar mais adequada. Considere como uma amostra incompleta uma linha que faltam dados em alguma das colunas selecionadas anteriormente.

Relatório: descreva o conjunto de dados e como foi realizado o tratamento. Além disso, descreva o conjunto de dados que resulta das operações de tratamento.

Atividade 2: Agrupamento com o *K-means*

Faça um agrupamento dos dados com o algoritmo *K-means*. Para a escolha do número de agrupamentos k, utilize duas métricas: a soma de distâncias intra-cluster e o coeficiente de silhueta. Os passos a serem seguidos são:

1. *Gráfico Elbow Curve:* construir um gráfico com a soma das distâncias intra-cluster para os valores de k variando de 2 a 30;

2. *Gráfico da Silhueta:* construir um gráfico com o valor da silhueta para os valores de k variando de 2 a 30;

3. *Escolha do k:* avalie os gráficos gerados (*Elbow Curve* e Silhueta) e escolha o melhor valor de k com base nas informações desses gráficos e na sua análise. Com o valor de k definido, utilize o rótulo obtido para cada amostra, indicando o *cluster* ao qual ela pertence, para gerar um gráfico de dispersão representando os *clusters* (atribuindo cores diferentes para cada cluster).

Relatório: apresente os gráficos *Elbow Curve* e *Silhueta*, descreva e analise o observado em cada um deles, apresente o valor escolhido para k explicando o motivo da escolha e, por fim, inclua e analise o gráfico de dispersão dos agrupamentos gerados para o valor de k escolhido.

Atividade 3: Agrupamento com o *DBscan*

Para a execução do algoritmo de agrupamentos *DBscan* analisaremos dois parâmetros principais do método, *eps* e *minPts*. Os passos a serem seguidos são:

1. *Análise do Raio da Vizinhança de Pontos:* nesta etapa, analisaremos o impacto do parâmetro *eps* para a formação de agrupamentos. Experimente com intervalos diferentes para o valor de *eps* e reporte apenas um deles.

2. *Determinando Ruídos:* nesta etapa, analisaremos o impacto do parâmetro *minPts* para a determinação de pontos de centro, de borda ou ruído. Experimente com intervalos diferentes para o valor de *minPts* e reporte apenas um deles.

Para cada conjunto de parâmetros, utilize o rótulo obtido para cada amostra, indicando o *cluster* ao qual ela pertence, para gerar um gráfico de dispersão representando os *clusters*.

Relatório: descreva e apresente os gráficos de dispersão para os resultados obtidos para cada conjunto de parâmetros utilizados indicando o melhor conjunto e explicando o motivo da escolha.

Conclusão dos Experimentos

Com base nas atividades anteriores, você deve incluir uma seção de conclusões no **relatório** contendo:

1. Qual dos métodos apresentou melhores resultados? Justifique.

2. Quantos agrupamentos foram obtidos?

3. Analisando o campo *symboling* e o grupo designado para cada amostra, os agrupamentos conseguiram separar os níveis de risco?

4. Analisando o campo *make* que contém as marcas dos carros, os agrupamentos conseguiram separar as marcas?

Considerações Finais

- O relatório corresponderá a 85% da sua nota.
- Você não deve remover qualquer linha já existente no arquivo (`.data`) da base 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 `imports-85.data`** na versão que foi disponibilizada no Trabalho 3.
- Comente o seu código, de modo que ele possa ser facilmente entendido e usado durante a correção.
- Use semente fixas para a aleatorização.
- Teste o seu código antes de submeter. Códigos que não executam serão penalizados.
- Apenas um membro da dupla (ou trio) deve enviar a solução. Os nomes dos membros devem constar tanto no relatório quanto no cabeçalho de cada arquivo "`.R`" a ser submetido.
- A submissão deve ser um arquivo "`.zip`" contendo: um arquivo com o código em *R* ("`.R`") e um relatório ("`.pdf`").
- O envio deve ser feito pelo sistema Moodle, clicando no link "`Trabalho 3`" 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: 10 de Maio de 2020 (Domingo), até às 23h55.

Forma de entrega: via sistema Moodle:

- `https://moodle.lab.ic.unicamp.br/moodle/course/view.php?id=393`

Pontuação: Este teste será pontuado de 0 a 10, e corresponderá a 40% da nota final.