Universidade Federal do Agreste de Pernambuco
Bacharelado em Ciências da Computação
Prof. Tiago Buarque A. de Carvalho

Aprendizagem de Máquina:
Pré-processamento de Dados

1. Nesta questão você deve utilizar a base "Student Performance", `archive.ics.uci.edu/ml/datasets/Student+Performance` (ver arquivo `student-mat.csv` no `student.zip`).

   (a) (5 pontos) Explique qual a forma mais adequada para converter todos os atributos da base para numéricos.

   (b) (10 pontos) Converta todos os atributos da base para numéricos (exceto a classe).

   (c) (10 pontos) Assuma a última coluna (G3, que representa a nota final de cada estudante) como classe. Converta esta coluna (atributo numérico) para uma variável categórica binária. Após esta conversão é possível realizar a tarefa a seguir.

   (d) (5 pontos) Calcule o intervalo de confiança da acurácia para o 100 repetições de holdout 50/50 utilizando o classificador 1-NN com distância Euclidiana.

2. Utilizando a base "Forest Fires". `archive.ics.uci.edu/ml/datasets/Forest+Fires`

   (a) (5 pontos) Indique a forma mais adequada de converter para numéricos cada um dos atributos da base.

   (b) (10 pontos) Realize a conversão da base conforme a resposta indicada.

3. Utilizando a base "Car Evaluation". `archive.ics.uci.edu/ml/datasets/Car+Evaluation`

   (a) (5 pontos) Indique a forma mais adequada de converter para numéricos cada um dos atributos da base.

   (b) (10 pontos) Realize a conversão da base conforme a resposta indicada.

   *talvez deixar esta questão mais completa (lista de AM) e remover/reduzir a questão anterior

4. A base "Heart Disease (hungarian)" possui alguns valores de atributos omissos. Realize o experimento descrito abaixo utilizando o classificador 1-NN. Divida a base em treino (90%) e teste (10%) de forma estratificada. Calcule o intervalo de confiança para a taxa de acerto do classificador utilizando 100 repetições deste experimento.

   `https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.hungarian.data`

   `https://archive.ics.uci.edu/ml/datasets/Heart+Disease`

   (a) (10 pontos) Preencha os valores omissos no conjunto de treino.

   (b) (10 pontos) Preencha os valores omissos no conjunto de teste utilizando o método e os valores definidos para o conjunto de treino.

5. Utilizando a base de dados Wine `https://archive.ics.uci.edu/ml/datasets/wine`, para cada um dos casos abaixo, realize 100 repetições de Holdout 50/50 e calcule o intervalo de confiança da acurácia utilizando o classificador 1-NN com distância Euclidiana. Realize testes de hipótese por sobreposição dos intervalos de confiança comparando os pré-processamentos de cada um dos casos abaixo com a base de dados original:

   (a) (10 pontos) Com todas as características ajustadas para o intervalo [0,1].

   (b) (10 pontos) Com todas as características ajustadas para ter média zero e desvio padrão igual a um.

   * fazer apenas uma transformação e comparar com os dados originais