Reconhecimento de Padrões: Atividade 05 – Pré-processamento de Dados

Carlos Emmanuel Pereira Alves Curso de Bacharelado em Ciência da Computação Universidade Federal do Agreste de Pernambuco (UFAPE) Garanhuns, Brasil carlos.emmanuel.236@gmail.com

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

Primeiro vamos separar os dados que já são numéricos e não precisam ser convertidos, que são: age, Medu, Fedu, traveltime, studytime, failures, famrel, freetime, goout, Dalc, Walc, health, absences, G1, G2, G3.

Agora com os atributos restantes vamos fazer a conversão para os que são binários e não-binários.

Binários (nominais e ordinais):

school, sex, address, famsize, Pstatus, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic.

Não binário (nominais):

- Mjob, Fjob, reason, guardian.
- b) Converta todos os atributos da base para numéricos (exceto a classe).
- c) Assuma a última coluna (G3, que representa a nota final de cada estudante) como classe. Converta esta coluna (atributo numérico) para

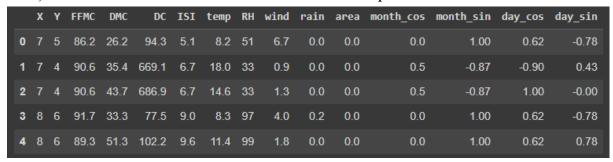
uma variável categórica binária. Após esta conversão é possível realizar a tarefa a seguir.

Converti utilizando o critério de que 14 é a nota mínima exigida para ser aprovado então se G3 < 14; G3 = Failed; e se G3 > 13; G3 = Approved.

- d) Calcule o intervalo de confiança da acurácia para o 100 repetições de holdout 50/50 utilizando o classificador 1-NN com distância Euclidiana. Intervalo de confiança da taxa de acerto: [0.86419 0.92945]
- 2) Utilizando a base Forest Fires. archive.ics.uci.edu/ml/datasets/Forest+Fires
 - a) Indique a forma mais adequada de converter para numéricos cada um dos atributos da base.

Os únicos atributos que precisam ser convertidos são: month e day. A forma mais adequada vai ser a utilização de dados cíclicos, por se tratar de dias da semana e do mês.

b) Realize a conversão da base conforme a resposta indicada.



- 3) Utilizando a base Car Evaluation. archive.ics.uci.edu/ml/datasets/Car+Evaluation
 - a) Indique a forma mais adequada de converter para numéricos cada um dos atributos da base.

Para a conversão a melhor forma será utilizar o ordinal não binário porque os valores indicam uma ordem que vai do mais baixo ao mais alto. Teremos de converter toda a base pois alguns atributos podem assumir valores numéricos e categóricos.

b) Realize a conversão da base conforme a resposta indicada.

	buying	maint	doors	persons	lug_boot	safety	class
0	3.0	3.0	0.0	0.0	0.0	0.0	0.0
1	3.0	3.0	0.0	0.0	0.0	1.0	0.0
2	3.0	3.0	0.0	0.0	0.0	2.0	0.0
3	3.0	3.0	0.0	0.0	1.0	0.0	0.0
4	3.0	3.0	0.0	0.0	1.0	1.0	0.0

- 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.
 - a) Preencha os valores omissos no conjunto de treino.
 - b) Preencha os valores omissos no conjunto de teste utilizando o método e os valores definidos para o conjunto de treino.

Intervalo de confiança da taxa de acerto: [0.34126 0.88523]

- 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) Com todas as características ajustadas para o intervalo [0,1].
 - b) Com todas as características ajustadas para ter média zero e desvio padrão igual a um.

Intervalo de confiança: [0.62758 0.80186]
Intervalo de confiança ajustado para [0, 1]: [0.90954 0.99293]
Intervalo de confiança da taxa de acerto padronizado: [0.90983 0.98994]

No original e no ajustado para [0, 1] não existe sobreposição de intervalos, então temos uma diferença significativa entre eles.

No original e no padronizado também não temos sobreposição de intervalos, e também temos uma diferença significativa entre eles.

No ajustado para [0, 1] e no padronizado temos sobreposição de intervalos, portanto não existe uma diferença significativa entre eles.