UNIVERSIDADE DE SÃO PAULO - ICMC
Departamento de Ciências de Computação
SCC-0275 – Introdução à ciência de dados – Graduação - 2° Sem /2018
PROFa: Roseli Aparecida Francelin Romero
Estagiário PAE: Daniel Moreira Cestari

Aula 07 - Exploração de dados - Parte V

Escolha apenas UM dataset de CLASSIFACAÇÃO para realizar os exercícios abaixo. Caso você não tenha utilizado um dataset desse tipo anteriormente escolha um novo para essa prática.

1- O desbalanceamento, em geral, causa problemas para a classificação. Trate o desbalanceamento utilizando as abordagens de subamostragem, superamostragem. Mostre a distribuição das classes após o balanceamento.

Para realizar a subamostragem e a superamostragem chame a função *sample* do DataFrame. Abaixo tem um exemplo de como utilizar essa função.

Utilize a função classificação definida abaixo ajustando o parâmetro *weights* para um valor que você ache adequado. Esse parâmetro faz com que o algoritmo de classificação leve em conta o peso da classe no processo de treinamento. Utilize 2 valores de *weights* diferentes e comente o resultado (o que você acha que mudou com esses valores diferentes de *weights*).

2- Faça a conversão entre tipos para dois atributos. Escolha as conversões que achar mais adequada, quantitativo para qualitativo ou qualitativo para quantitativo. Explique o que você fez e porque achou essa a melhor forma de fazer. Para referência consulte os slides 9-21.

Esse tipo de conversão é importante pois se você for utilizar uma rede neural artificial. Quando temos uma classe numérica como no wine, 0,1,2, é melhor transformar a classe em 3 colunas e a primeira coluna é 1 quando a classe for 0, a segunda coluna tem valor 1 quando a classe for 1, e a terceira coluna é 1 quando a classe for 2. Isso resulta numa rede neual com 3 neuronios na saída onde cada neurônio sinaliza uma das classes. Esse processo também ajuda na convergencia da rede neural artificial.

3- Realize a normalização de duas colunas. Para normalizar utilize a abordagem 0-1 (slide 27) e z-score (slide 29). Apresente métricas que comprove que a normalização foi realizada (boxplot, histograma, ...).

OBS: A etapa de pré-processamento deve ser feita pensando nos passos seguintes. Por exemplo, se penso em utilizar posteriormente um algoritmo de classificação que necessita atributos numéricos no intervalo 0-1 sei que preciso remover os atributos não numéricos, ou transformá-los em numéricos, e fazer a normalização 0-1. Outro exemplo é o das funções simples mostradas no slide 24, o log e a raíz quadrada aproximam os dados de uma normal, logo se sua abordagem supoe normalidade dos dados pode ser interessante aplicar alguma dessas funções.

UNIVERSIDADE DE SÃO PAULO - ICMC
Departamento de Ciências de Computação
SCC-0275 – Introdução à ciência de dados – Graduação - 2° Sem /2018
PROFa: Roseli Aparecida Francelin Romero
Estagiário PAE: Daniel Moreira Cestari

Entrega

A entrega da prática deve ser feita em formato de notebook do jupyter. Crie um notebook seu, nas primeiras linhas identifique os membros da dupla, salve o notebook com o nome pratica_2_nUSP1_nUSP2. Submeta no tidia apenas o arquivo do notebook .ipynb, não crie pastas para separar as práticas, apenas faça o upload do arquivo .ipynb Não precisa fazer o upload dos datasets utilizados. A resolução desta prática deve seguir a mesma maneira da anterior. No corpo da prática primeiro identifique qual o conjunto de dados escolhido e começe a responder as perguntas no corpo do notebook. Procure utilizar as caixas de texto para a discussão (o notebook tem caixas de texto e de código).

Funções utilizadas na prática

DataFrame

- sample(n=None, frac=None, replace=False, weights=None, random_state=None, axis=None)

scikitlearn

- Procurar pelo módulo preprocessing

SciPy.stats

- zscore(a, axis=0, ddof=0)