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 06 - Exploração de dados - Parte IV

Caso os atributos que você escolheu nas práticas anteriores não for adequado à realização da prática, escolha outros. Por exemplo, você precisará verificar se há amostras com problemas, como valor ausente, se os atributos que você escolheu não possui valors ausentes, escolha outros atributos que tenha. Caso queira e tenha tempo você também pode para todos os atributos do conjunto de dados.

Para cada conjunto escolhido na prática anterior, realizar as seguintes atividades:

1- Reportar porcentagem de amostras com problema. Definir o que é problema no seu cenário. Problema pode ser valor ausente, inconsistente, ou outras formas que você acredita ser um problema. Especificar o problema por atributo, se você identificou problema nos atributos A e C, reportar a porcentagem de problema para cada um desses atributos.

2- Argumente se você acha que esses problemas têm origem sistemática ou aleatória.

3- Realize uma etapa de limpeza que você achar necessária. Comente as etapas realizadas. Por limpeza vide as atividades nos slides 13 em diante. Você pode achar apenas necessário fazer imputação dos dados ou então se forem poucas amostras com valores ausentes, remover essas amostras, mas sempre justifique o porque você fez isso.

4- Execute a função com os classificadores ou regressores, dependendo do seu problema, antes e depois da limpeza dos dados. Comente se o classificador conseguiu executar antes da limpeza ser realizada, e se houve melhora no desempenho.

Abaixo está implementada uma função que realiza a classificação/regressão para 4 modelos (reg_clas). Utilize essa função. A medida de desempenho da classificação é a acurácia (número de acertos pelo total de exemplos), e a medida da regressão é o erro quadrático médio (média da diferença ao quadrado do valor predito pelo valor real $\sum ni=1(yipred−yireal)2n)$

Se você tiver problema na execução da função tente executar apenas com colunas numéricas.

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

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

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

- isna()
- isnull()
- sum(axis=None, skipna=None, level=None, numeric_only=None, min_count=0)
- drop(labels=None, axis=0, index=None, columns=None, level=None, inplace=False, errors='raise')
- dropna(axis=0, how='any', thresh=None, subset=None, inplace=False)
- drop_duplicates(subset=None, keep='first', inplace=False)
- fillna(value=None, method=None, axis=None, inplace=False, limit=None, downcast=None)
- replace(to_replace=None, value=None, inplace=False, limit=None, regex=False, method='pad')