05

Organizando os dados

Transcrição

Faremos o pré processamento dos nosso dados, isto é, retirar o que for desnecessário e organizar as informações para diminuir o risco de possíveis problemas durante a clusterização.

Retiraremos o atributo de id, pois não nos é útil agrupar clientes por sua identificação individual. Retiraremos também o atributo "Tenure", que é o tempo que o contrato de crédito demora para ser renovado, isto é, 12 meses. Por isso esse atributo tem o mesmo valor para dos os elementos.

Utilizaremos a função `drop()` para remover colunas específicas, e então declararemos o nome dessas colunas.

```
import padas as pd

dataframe = pd.read_csv("cc GENERAL.csv")
detaframe.drop(columns["CUST_ID", "TENURE"], implace=True)
dataframe.head()
```

Feito isso, essas colunas serão removidas do dataframe, e temos 16 atributos dos 18 originais.

A próxima etapa é buscar pelos missing data, isto é, dados faltantes do dataframe. Para tanto, escreveremos:

```
missing = dataframe.isna().sum()
print(missing)
```

Dessa maneira, buscaremos por todos os dados nulos e em quais colunas se localizam. No atributo `minimum_payments` encontraremos 313 dados faltantes.

Substituiremos esse valor pela mediana dos valores deste atributo. Então escreveremos:

```
dataframe.fillna(dataframe.median(), inplace=True)
missing = dataframe.isna().sum()
print(missing)
```

Feito isso, não teremos nenhum dado faltante em nosso dataframe.

O próximo passo é normalizar nossos dados. Em alguns atributos relacionados à frequência, teremos um limite que varia de 0 a 1, sendo que 0 é 0% de frequência e 1, por sua vez, significa 100%. Para o atributo de balanço, não teremos a mesma organização de mínimo ou máximo, e isso pode fazer com que o algorítimo gere dados pouco satisfatórios para nós.

Para realizar a normalização dos dados utilizaremos o biblioteca Scikit learn. Escreveremos:

```
from sklearn.preprocessing import Normalizer
values = Normalizer().fit_transform(dataframe.value)
print.(values)
```

- o `Normalizar` colocará nossos valores entre 0 e 1. Dessa maneira temos nosso dados limpos e completos.