Universidade do Minho
Departamento de Informática
Mestrado [integrado] em Engenharia Informática

Perfil de Machine Learning: Fundamentos e Aplicações
Classificadores e Sistemas Conexionistas
4º Ano, 2º Semestre
Ano letivo 2019/2020

Enunciado Prático nº 1
13 de fevereiro de 2020

Tema

Manipulação de dados categóricos

Enunciado

Dados categóricos, não numéricos, são, no geral, mais complexos de lidar do que dados numéricos. De facto, são vários os algoritmos de *Machine Learning* que exigem que o seu input seja numérico, sendo necessária a aplicação de técnicas que transformem esses mesmos atributos. Duas das técnicas mais utilizadas são o *one-hot encoding* e o *label encoding*. A primeira, utilizada essencialmente quando não existe uma ordenação entre as categorias, consiste na criação de novos atributos binários para cada chave única de um dado atributo não numérico (por exemplo, novos atributos binários para cada tipo de cozinha ou para cada género de filmes). A segunda, *label encoding*, consiste na criação de um *label* para cada chave única de um dado atributo não numérico ([Braga, Porto, Guimarães, Braga] para [0, 1, 2, 0]).

Tarefas

Esta ficha encontra-se dívida em duas partes distintas.

1. Na primeira parte desta ficha prática pretende-se que sejam instalados e criados ambientes virtuais para suportar os trabalhos práticos que serão realizados durante o semestre letivo.

2. Para a segunda parte da ficha prática, devem descarregar o *dataset* disponível em https://goo.gl/mRgSAq que contém um conjunto de informação sobre vários voos, incluindo o ano, a hora de chegada, o atraso sofrido, a origem e o destino do voo, entre outros. Pretende-se assim que sejam realizadas as seguintes tarefas:

   - Analisar o *dataset* em relação ao número de registos e a cada um dos seus atributos, os seus valores e respetivos tipos;
   - Desenvolver o algoritmo correspondente à técnica de *label encoding*;
   - Utilizar a biblioteca *pandas*, em particular o método *get_dummies*, para aplicar a técnica de *one-hot encoding*;
   - Como extra - utilizar uma plataforma na *cloud*, por exemplo *Google Colab*, para executar o código produzido.

Como base para o desenvolvimento da ficha deixa-se o seguinte excerto de código:

```
import numpy as np
import pandas as pd
```

```
#read dataset
df = pd.read_csv('flights_dataset.csv')

#dataset info
print(df.info())
print('--------')

#drop unwanted columns
df.drop(['hour','minute','tailnum'], 1, inplace=True)

#infer objects type
df.infer_objects()

#check and replace missing with -99 (masking)
print(df.isnull().sum())
print('--------')
df.fillna(-99, inplace=True)

#frequency distribution of categories within a feature
print(df['dest'].unique())
print('Unique count: %d' %df['dest'].value_counts().count())
print('--------')
print(df['dest'].value_counts())
print('--------')

'''
Function to encode all non-(int/float) features in a dataframe.
For each column, if its dtype is neither int or float, get the list of unique values,
store the relation between the label and the integer that encodes it and apply it.
Return a labelled dataframe and a dictionary label to be able to restore the original value.
'''
def label_encoding(df):
    pass

'''
Function to decode what was previously encoded - get the original value!
'''
def label_decoding(df_labelled, label_dictionary):
    pass

df_labelled, label_dictionary = label_encoding(df)
print(df_labelled['dest'].unique())
print('Unique count after Label Encoding: %d' %df_labelled['dest'].value_counts().count())

df_labelled_decoded = label_decoding(df_labelled, label_dictionary)
print(df_labelled_decoded['dest'].unique())
print('Unique count after dec.: %d' %df_labelled_decoded['dest'].value_counts().count())
print('--------')

'''
Use a pandas' function to apply one-hot encoding to the origin column
'''
print(df.columns.values)
df_pandas_ohe = TODO
print(df_pandas_ohe.columns.values)
```