Aprendizado de Máquinas – Avaliação 1

Caio Lins

26 de março de 2022

1 Apresentação do Dataset

Como banco de dados para ser utilizado no projeto final da disciplina, escolhemos o *"Census Income Data Set"*[1]. O dataset foi extraído da base de dados do censo populacional americano de 1994. Cada instância representa um indivíduo, que possui algumas caracteríscas sociais e econômicas disponíveis, correspondentes às colunas da tabela. Os dados foram extraídos já com um problema de classificação binária em mente: a coluna relativa à renda anual do indivíduo só informa se ele ganha mais ou menos que $50 000 por ano.

Para ler e manipular os dados, utilizamos a versão 1.2.4. da biblioteca *Pandas* [2], da linguagem de programção *Python* [3], versão 3.8.10. No repositório da UCI, os dados se encontram previamente divididos em um conjunto para treino e outro para teste, na proporção 2:1. Nós juntamos as duas tabelas para obter um data frame com todas as instâncias disponíveis. No total, são 48 840 entradas com 14 atributos distintos, listados a seguir:

- `Age`. Variável numérica que assume valores inteiros. Corresponde à idade do indivíduo.
- `workclass`. Variável categórica que assume os seguintes valores: `Private`, `Self-emp-not-inc`, `Self-em-inc`, `Local-gov`, `State-gov`, `Federal-gov`, `Without-pay`, `Never-worked`.
- `fnlwgt`. Variável numérica que assume valores inteiros. É um peso calculado pelo *United States Census Bureau* que indica quantas pessoas aquele indivíduo representa na população . Ele é necessário devido às estratégias de amostragem utilizadas para decidir quem será entrevistado no censo.
- `education`. Variável categórica que assume os seguintes valores: `Preschool`, `1st-4th`, `5th-6th`, `7th-8th`, `9th`, `10th`, `11th`, `12th`, `HS-grad`, `Some-college`, `Assoc-voc`, `Assoc-acdm`, `Bachelors`, `Masters`, `Prof-school`, `Doctorate`. Corresponde ao grau máximo de educação obtido pelo indivíduo.
- `education_num`. Variável numérica categórica que assume valores inteiros entre 1 e 16. Corresponde a uma codificação numérica da educação do indivíduo, na ordem apresentada anteriormente.

- `marital_status`. Variável categórica que assume os seguintes valores: `Married-civ-spouse`, `Divorced`, `Never-married`, `Separated`, `Widowed`, `Married-spouse-absent`, `Married-AF-spouse`. Corresponde à situação conjugal do indivíduo.

Referências

[1] Dheeru Dua e Casey Graff. *UCI Machine Learning Repository – Census Income Data Set*. 2017. URL: `https://archive.ics.uci.edu/ml/datasets/Census+Income`.

[2] Wes McKinney. «Data Structures for Statistical Computing in Python». Em: *Proceedings of the 9th Python in Science Conference*. Ed. por Stéfan van der Walt e Jarrod Millman. 2010, pp. 51–56.

[3] Guido Van Rossum e Fred L. Drake. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.