Tutorial: Formatação de Datasets Para Algoritmos de OPF

Este tutorial visa fornecer orientações sobre a formatação adequada de datasets para o processamento eficiente de algoritmos de Optimum-Path Forest (OPF).

Passos:

1. Formatar arquivo TXT
2. Converter para binário DAT

1. Converter CSV para TXT OPF

O armazenamento de dados é frequentemente realizado em formato tabular, sendo o formato CSV (Comma-Separated Values) uma escolha comum. O CSV utiliza vírgulas para separar os valores, onde a primeira linha geralmente atua como cabeçalho das colunas, e cada linha subsequente representa as amostras de dados. Outros delimitadores, como ponto e vírgula (;), tabulação (\t) ou espaços, também podem ser utilizados.

Exemplo 1: Formato CSV

```
CAMPO 1, CAMPO 2, CAMPO 3, CAMPO 4
Dado, Dado, Dado, Dado
Dado, Dado, Dado, Dado
```

Tabela 1: Tabela formato CSV

CAMPO 1	CAMPO 2	CAMPO 3	CAMPO 4
Dado	Dado	Dado	Dado
Dado	Dado	Dado	Dado

A maioria dos algoritmos de Machine Learning é adaptada para processar dados nesse formato. No entanto, para os algoritmos OPF, é necessário um formato específico que não viabiliza o uso do CSV. Nesse caso, o cabeçalho é representado por três valores: o total de

amostras (tuplas do dataset), o total de classes presentes no dataset e o total de features (variáveis do dataset, excluindo a classe). Além disso, cada amostra requer um rótulo, começando do 0 (zero) até o (total de amostras - 1), por exemplo:

Exemplo 2: Dataset no formato CSV

```
CLASSE, FEATURE 1, FEATURE 2,  FEATURE 3
C1, Dado, Dado, Dado, Dado
C1, Dado, Dado, Dado, Dado
C2, Dado, Dado, Dado, Dado
C2, Dado, Dado, Dado, Dado
C3, Dado, Dado, Dado, Dado
```

Tabela 2: Tabela no formato CSV

CLASSE	FEATURE 1	FEATURE 2	FEATURE 3
C1	Dado	Dado	Dado
C1	Dado	Dado	Dado
C2	Dado	Dado	Dado
C2	Dado	Dado	Dado
C3	Dado	Dado	Dado

Nesse exemplo 2 do formato CSV observamos 3 classes (C1, C2 e C3), 5 amostras e 3 features. Para tanto, para esse mesmo dataset ser processado em um algoritmo OPF as modificações necessárias seriam:

Exemplo 3: Dataset no formato TXT

```
5 3 3
0 C1 Dado Dado Dado Dado
1 C1 Dado Dado Dado Dado
2 C2 Dado Dado Dado Dado
3 C2 Dado Dado Dado Dado
5 C3 Dado Dado Dado Dado
```

No exemplo 3, as vírgulas foram substituídas por espaços, e o cabeçalho original com os campos foi removido e substituído pelo cabeçalho específico do formato OPF. Esse formato inclui o total de amostras, o total de classes e o total de features, seguido pelos dados de cada amostra com seu rótulo, classe e características correspondentes.

Você pode encontrar algoritmos de conversão de CSV para DAT ou vice versa nesse notebook.

2. Converter TXT OPF para binário DAT

Após o processo anterior, a partir do algoritmo encontrado em LibOPF é necessária a conversão do arquivo TXT já formatado para o processamento do dataset no algoritmo OPF, dessa forma é só seguir os passos a seguir:

2.1. Após fazer o clone do repositório LibOPF, encontre a pasta `/tools`, nela pode-se encontrar algumas ferramentas que ajudaram no processo de conversão de tipos de arquivos, como `txt2dat` (TXT para DAT), `dat2txt` (DAT para TXT), entre outros. Após encontrar o onde está seu arquivo TXT formatado, pelo terminal entre em `/tools`, em seguida por exemplo digite `txt2dat dataset_formatado.txt dataset.dat`, lembre-se de colocar o nome e caminho correto do dataset formatado TXT.

2.2. Após executar esse comando você terá como resultado um novo arquivo DAT, no caso o dataset.dat, a partir daqui encontre a pasta `/bin` dentro do arquivo OPF, nessa pasta contém os algoritmos executáveis para o processamento do dataset, mais detalhes em LibOPF/Examples.