Decision Tree - Prática

Ciclo	Ciclo 06: Algoritmos baseado em árvores
# Aula	43
Created	@March 9, 2023 10:11 AM
Done	☑
Ready	☑

Objetivo da Aula:

- [ ] Decision Tree Classifier
- [ ] Hiperparametros de controle
- [ ] Resumo
- [ ] Próxima aula

Conteúdo:

▼ 1. Decision Tree Classifier

```
# Load datasets
import cv2
import pydotplus
import matplotlib.pyplot as plt

from six             import StringIO
from sklearn         import datasets as dt
from sklearn         import tree     as tr
from IPython.display import Image

# Import dataset
iris = dt.load_iris()
X = iris.data[:, 2:]
y = iris.target
```

```
# Model fit
tree_clf = tr.DecisionTreeClassifier( max_depth=2 )
tree_clf.fit( X, y )

# Export draw
dot_data = StringIO()
tr.export_graphviz(
    tree_clf,
    out_file='tree.dot',
    feature_names=iris.feature_names[2:],
    class_names=iris.target_names,
    rounded=True,
    filled=True
)

# Convert .dot to .png
!dot -Tpng tree.dot -o tree.png

# Load image on jupyter notebook
img = cv2.imread('tree.png')
plt.figure(figsize = (20, 20))
plt.imshow(img)

# Predict
tree_clf.predict ( [[5, 1.5]] )
```

▼ 2. Hiperparametros de controle

A algoritmo Decision Tree é um modelo não-paramétrico. Apesar de possuir parâmetros, seu modelo não depende de uma fórmula pré-estabelecida como as Regressões por exemplo.

A falta de um modelo matemático prévio, fornece um alto grau de liberdade que aumentam as chances de overfitting.

Para reduzir as chances de overfitting é necessário regular alguns parâmetros do algoritmo que controlam o crescimento da árvore ou limitam o número de recortes espaciais feito pelo algoritmo, no conjunto de dados.

Os parâmetros que regulam a Decision Tree são:

- *max_depth:* controle o tamanho máximo das quebras, ou seja, o tamanho máximo do crescimento da árvore ou ainda o número de recortes espaciais.

- *min_samples_leaf:* O número mínimo de amostras do nó deve ter, antes de fazer uma nova separação e gerar nós filhos.
- *min_weight_fraction_leaf:* A mesma definição do parâmetro min_samples_leaf, mas definida como a fração do número total dos pesos das evidências.
- *max_features:* O número máximo de atributos que são avaliados para a divisão de cada nó.

Em regras gerais, aumentando os parâmetros que começam com *min_* ou reduzindo os parâmetros que começam com *max_* vão regular o algoritmo.

▼ 3. Resumo

1. O aprendizado não-supervisionado tem o objetivo de agrupar indivíduos com características ou comportamentos semelhantes, para encontrar padrões.

▼ 4. Próxima aula

Métricas de avaliação I: Curva ROC