



FIAP



MACHINE LEARNING & MODELLING

Esta disciplina aborda os principais conceitos sobre aprendizado de máquina e as técnicas clássicas de modelagem



Na última aula...

- Conceitos avançados sobre dados
- Árvores de decisão

Agenda

- Parametrização dos modelos baseados em árvores
- Como evitar overfitting
- Como observar a composição da árvore construída
- Trabalhando com outra métrica de avaliação (AUC)
- Pipeline para comparação de múltiplos modelos
- Criação do modelo final (e seu uso!)

Árvores de decisão

Vantagens

- Pouco esforço de preparação dos dados
- São capazes de lidar com múltiplos labels (classes)
- Alta explicabilidade

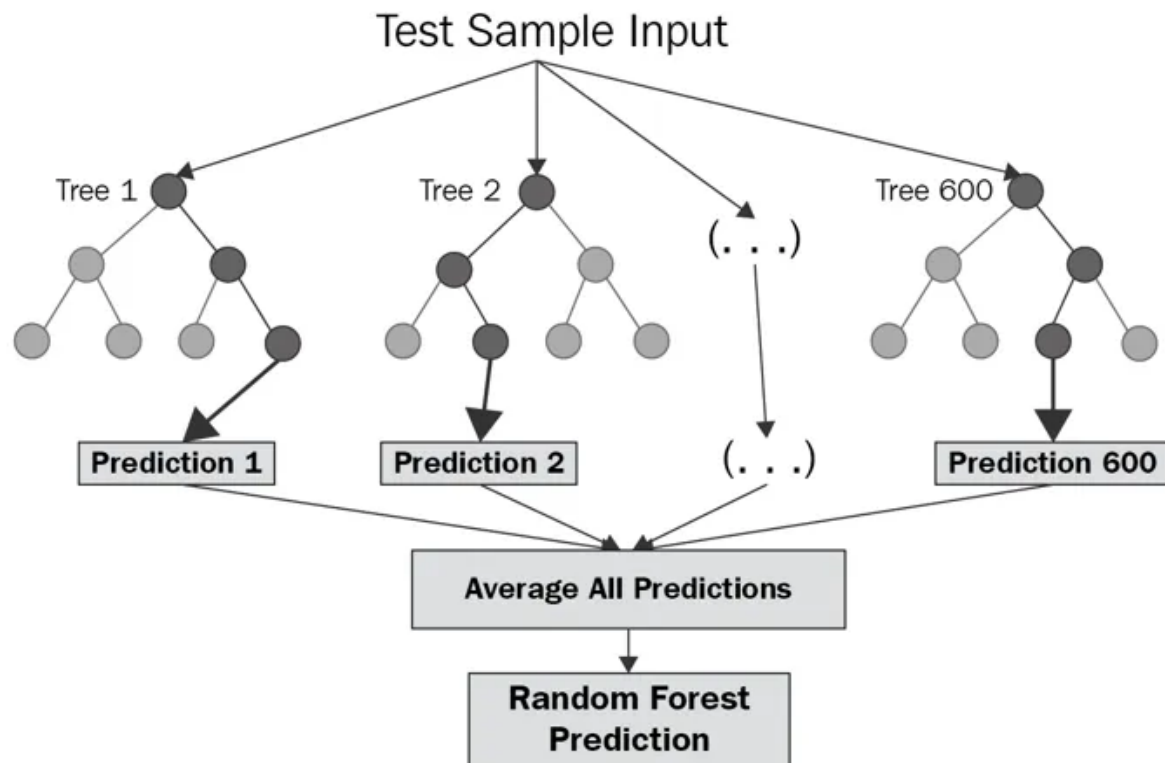
Desvantagens

- Podem ser desbalanceadas (*overfitting*)
- São instáveis (alta variância)
- Podem não ser ótimas

Como contornar os 2 últimos problemas?



Florestas Aleatórias



Parametrizando os modelos

```
# Arvore  
tree = DecisionTreeClassifier()
```

```
floresta = RandomForestClassifier(n_estimators=5)
```

sklearn.tree.DecisionTreeClassifier

```
class sklearn.tree.DecisionTreeClassifier(*, criterion='gini', splitter='best', max_depth=None, min_samples_split=2,  
min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features=None, random_state=None, max_leaf_nodes=None,  
min_impurity_decrease=0.0, min_impurity_split=None, class_weight=None, ccp_alpha=0.0)
```

[\[source\]](#)

sklearn.ensemble.RandomForestClassifier

```
class sklearn.ensemble.RandomForestClassifier(n_estimators=100, *, criterion='gini', max_depth=None, min_samples_split=2,  
min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features='auto', max_leaf_nodes=None, min_impurity_decrease=0.0,  
min_impurity_split=None, bootstrap=True, oob_score=False, n_jobs=None, random_state=None, verbose=0, warm_start=False,  
class_weight=None, ccp_alpha=0.0, max_samples=None)
```

[\[source\]](#)

Parametrizando a Árvore

Parameters:

criterion : {"gini", "entropy"}, default="gini"

The function to measure the quality of a split. Supported criteria are "gini" for the Gini impurity and "entropy" for the information gain.

splitter : {"best", "random"}, default="best"

The strategy used to choose the split at each node. Supported strategies are "best" to choose the best split and "random" to choose the best random split.

max_depth : int, default=None

The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

min_samples_split : int or float, default=2

The minimum number of samples required to split an internal node:

- If int, then consider min_samples_split as the minimum number.
- If float, then min_samples_split is a fraction and $\text{ceil}(\text{min_samples_split} * n_samples)$ are the minimum number of samples for each split.

Changed in version 0.18: Added float values for fractions.

min_samples_leaf : int or float, default=1

The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of the left and right branches. This

Parametrizando os modelos

- Veja mais:
 - Árvores: <https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html>
 - Florestas: <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>

Modelos baseados em árvores: evitando overfitting

Algumas estratégias para ajudar a evitar o overfitting:

- Especificar o **número mínimo de amostras para divisão do nó**. Valores altos previnem a criação de modelos complexos ou podem causar underfitting.
- Especificar o **número mínimo de amostras para o nível folha**. Nos permite controlar o crescimento da árvore e deve ser considerado em cenários com dados desbalanceados.
- **Profundidade máxima da árvore**.
- **Número máximo de features para considerar durante a divisão**.



E na prática?

Vamos ver como refinar o modelo criado na última aula

OBRIGADO!



Prof. Michel Fornaciali

<https://www.linkedin.com/in/michelfornaciali/>
profmichel.fornaciali@fiap.com.br

Copyright © 2021 | Professor Michel Fornaciali

Todos os direitos reservados. Reprodução ou divulgação total ou parcial deste documento, é expressamente proibido sem consentimento formal, por escrito, do professor/autor.