Selecting Features in Gradient Boosting

Mike West
MACHINE LEARNING ENGINEER

Module Overview

Feature selection

Feature selection algorithms

Feature importance in gradient boosting

Feature construction

Feature selection demonstration
XGBoost

PassengerID	Survived	Name	Sex	Age
1	0	Braund, Owen	male	22
2	1	Cumings, John	female	38
3	1	Heikkinen, Laina	female	26
4	1	Futrelle, Jacques	female	35
5	0	Allen, Henry	male	35

Feature Selection Outcomes

Improving the predictive performance of your models

Increasing the models speed by combining or removing various features

The fewer number of attributes reduces model complexity

Feature Selection Algorithms

Filter Methods

Statistical measure to assign a scoring to each feature

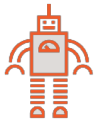
Wrapper Methods

The selection of a set of features as a search problem

Embedded Methods

Which features contribute to the accuracy of the model

Feature Importance



Gradient boosting models can automatically provide estimates of feature importance

The more an attribute is used to make key decisions with decision trees, the higher its relative importance

Importance is calculated for a single decision tree by the amount that each attribute splits

Gini impurity can be seen as a way to quantify how “good” a group is, so that you can choose the threshold wisely

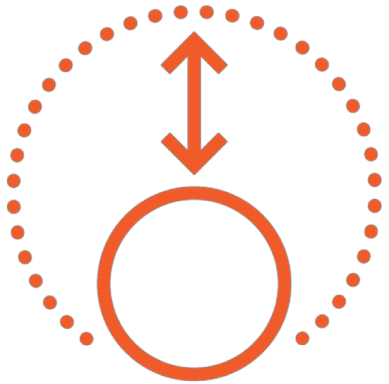
Feature Construction

Data Scientist

The results ultimately come from the practitioner, crafting the features

Art Form

Thinking about the underlying form of the problem



Limiting Features

Not all features are created equal. Move from many to a few

JavaScript Object Notation

JSON

New standard for data exchange among various platform

Minimalism

Transmit data between a server and web application

Keys and Values

Specific syntax with the key followed by a colon followed by the value

Demo

Import your libraries

Automatic feature calculation

Importance plot

Feature selection in SciKit-Learn

Feature selection demonstration

Summary

- Defined feature selection
- Feature selection modeling
- Gradient boosting feature importance
- Feature construction
- Feature selection in XGBoost