DS 6003 Amazon Spark Exercise

Fang You (fy6vj)

01/30/2019

Motivation

- Data source: Red variants of the Portuguese "Vinho Verde" wine in 2009
- Regression task to analyse and predict wine quality based on predictors such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol

Code Snippet

- Use VectorAssembler to combine various predictors into a a single feature vector.

```
assembler = VectorAssembler(
    inputCols = ["fixed_acidity", "volatile_acidity", "citric_acid", "residual_sugar", "chlorides", "free_sulfur_dioxide", \
                 "total_sulfur_dioxide", "density", "pH", "sulphates", "alcohol"],
    outputCol = "features")

trainingDF = assembler.transform(trainingDF)
testDF = assembler.transform(testDF)
```

Visualization