Part 2: Spark and ML

Applications to Finance

Miha Torkar
Jozef Stefan Institute
17th October 2019
miha.torkar@ijs.si
Overview

• Machine Learning (ML) pipeline
• Feature extraction
• ML methods
  • Logistic Regression
  • Random Forrest
  • SVM
• Implementation in Spark
Machine Learning pipeline in general
Pre-processing and Feature extraction

- Standardize numeric features
- Transform categorical features to numeric (one hot encoding)
- Remove outliers (if necessary)
- Domain knowledge feature engineering
- Feature extraction from raw data
ML Methods – quick overview

• Classification
  • Discrete target variable
  • Binary vs Multi class

• Regression
  • Continuous target variable
  • Time series of data; predicting the next step

• The better the input features, the better the model
Linear and Logistic Regression
Random Forrest

![Random Forest Diagram](image.png)
Support Vector Machine (SVM)
Implementation in Spark

• API documentation: https://spark.apache.org/docs/latest/api/python/pyspark.ml.html

• Programing guide: https://spark.apache.org/docs/latest/ml-guide.html
## Model scalability for classification

<table>
<thead>
<tr>
<th>Model</th>
<th>Features count</th>
<th>Training examples</th>
<th>Output classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic regression</td>
<td>1 to 10 million</td>
<td>No limit</td>
<td>Features x Classes &lt; 10 million</td>
</tr>
<tr>
<td>Decision trees</td>
<td>1,000s</td>
<td>No limit</td>
<td>Features x Classes &lt; 10,000</td>
</tr>
<tr>
<td>Random forest</td>
<td>10,000s</td>
<td>No limit</td>
<td>Features x Classes &lt; 100,000</td>
</tr>
<tr>
<td>Gradient-boosted trees</td>
<td>1,000s</td>
<td>No limit</td>
<td>Features x Classes &lt; 10,000s</td>
</tr>
</tbody>
</table>
# Model scalability for regression

<table>
<thead>
<tr>
<th>Model</th>
<th>Number features</th>
<th>Training examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear regression</td>
<td>1 to 10 million</td>
<td>No limit</td>
</tr>
<tr>
<td>Generalized linear regression</td>
<td>4,096</td>
<td>No limit</td>
</tr>
<tr>
<td>Isotonic regression</td>
<td>N/A</td>
<td>No limit</td>
</tr>
<tr>
<td>Decision trees</td>
<td>1,000s</td>
<td>No limit</td>
</tr>
<tr>
<td>Random forest</td>
<td>10,000s</td>
<td>No limit</td>
</tr>
<tr>
<td>Gradient-boosted trees</td>
<td>1,000s</td>
<td>No limit</td>
</tr>
<tr>
<td>Survival regression</td>
<td>1 to 10 million</td>
<td>No limit</td>
</tr>
</tbody>
</table>