Part 2: Spark and ML

Applications to Finance

Miha Torkar

Jozef Stefan Institute

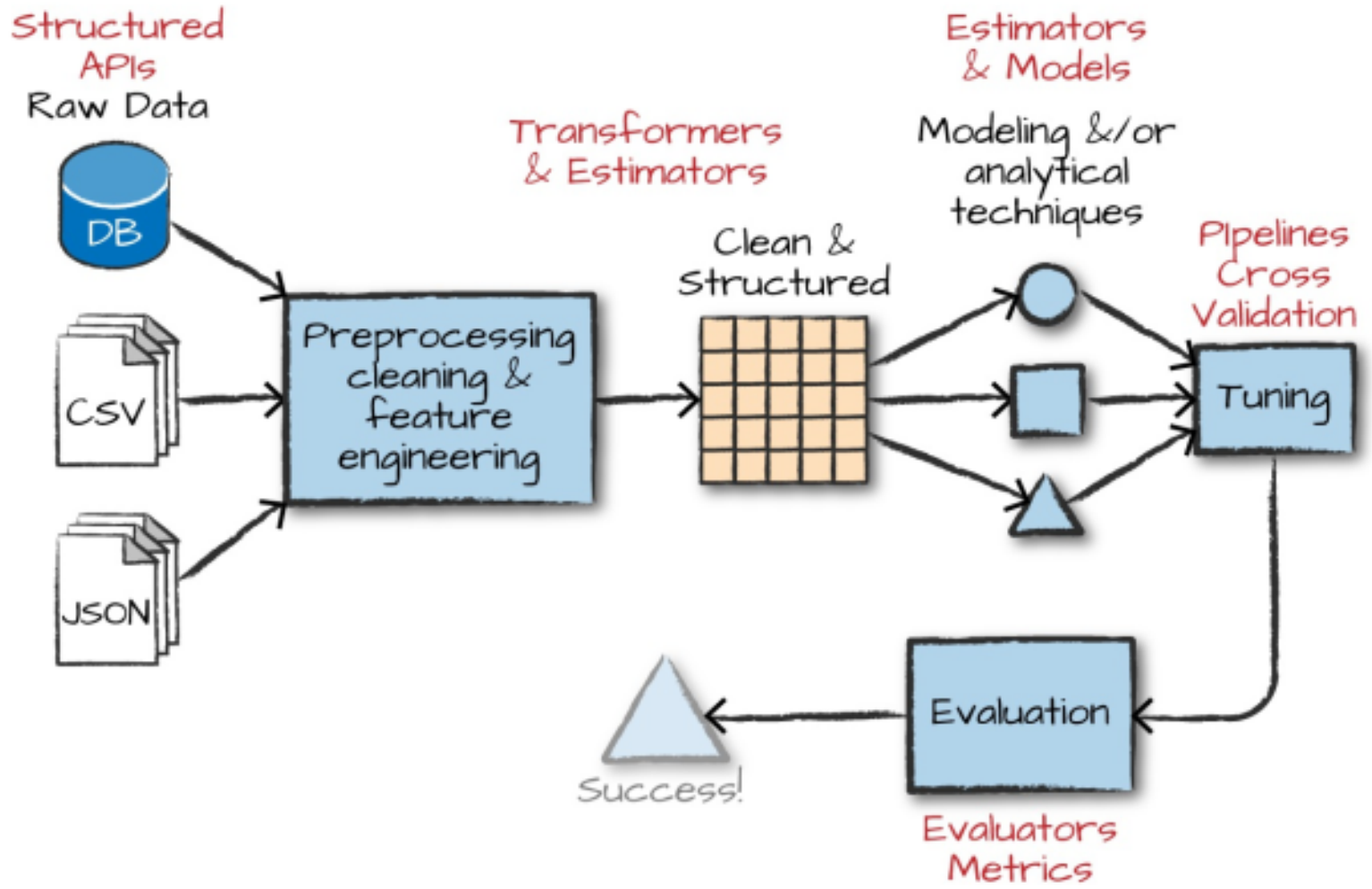
17th October 2019

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

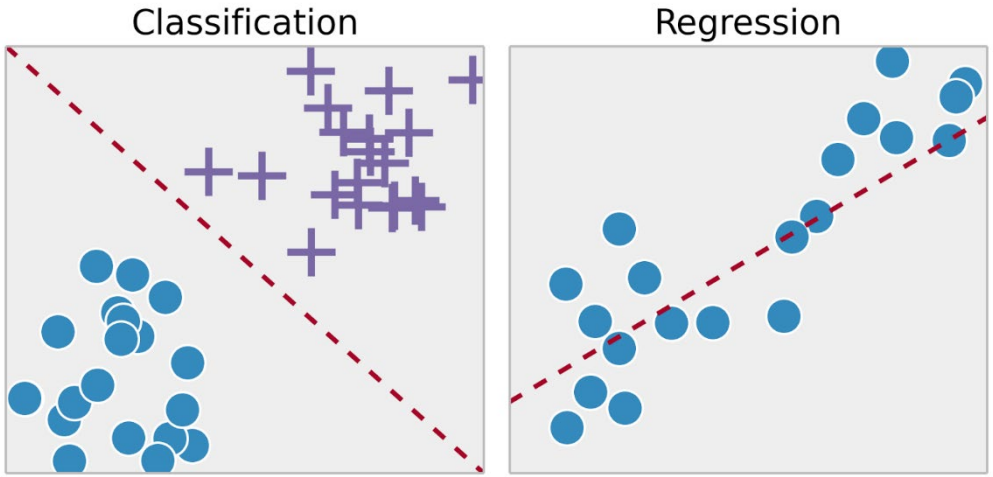


Pre-processing and Feature extraction

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

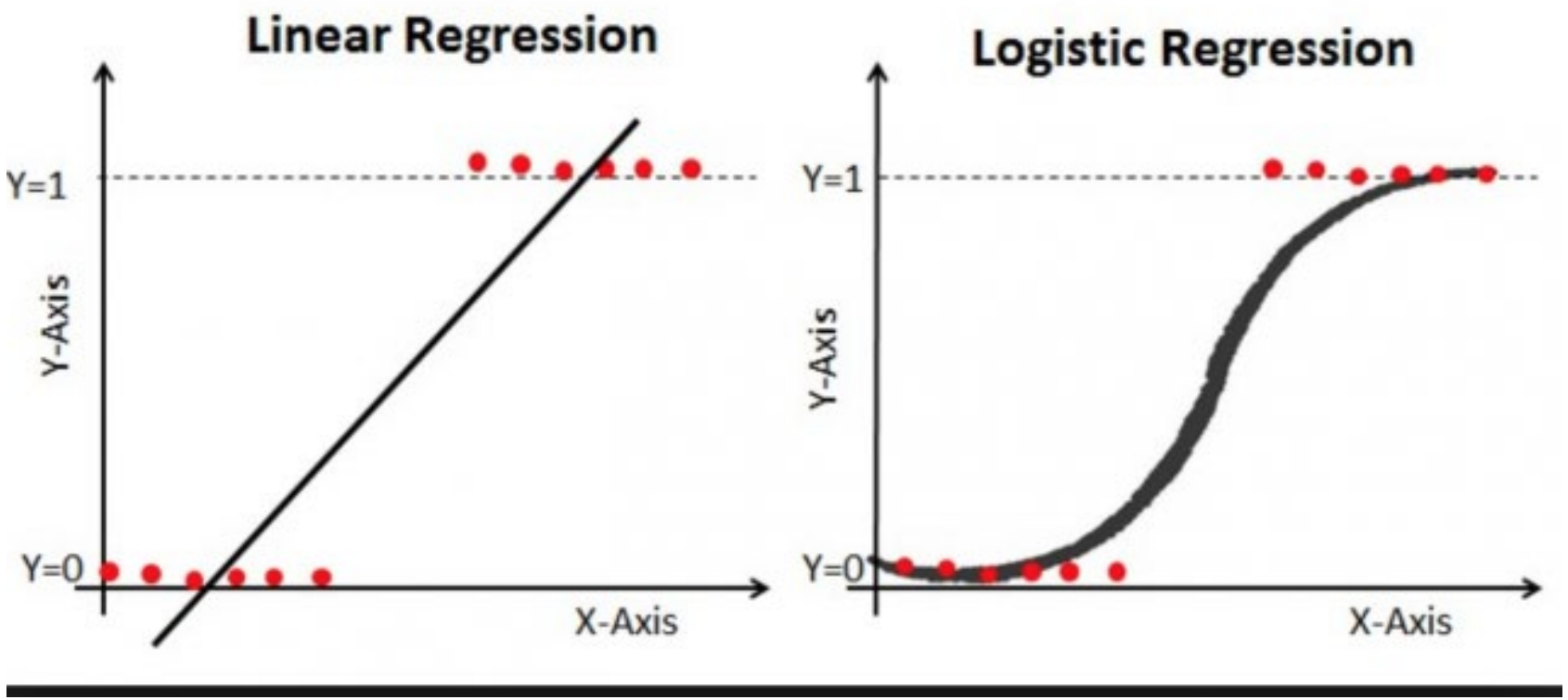
ML Methods – quick overview

- Classification
  - Discrete target variable
  - Binary vs Multi class
- Regression
  - Continuous target variable
  - Time series of data; predicting the next step

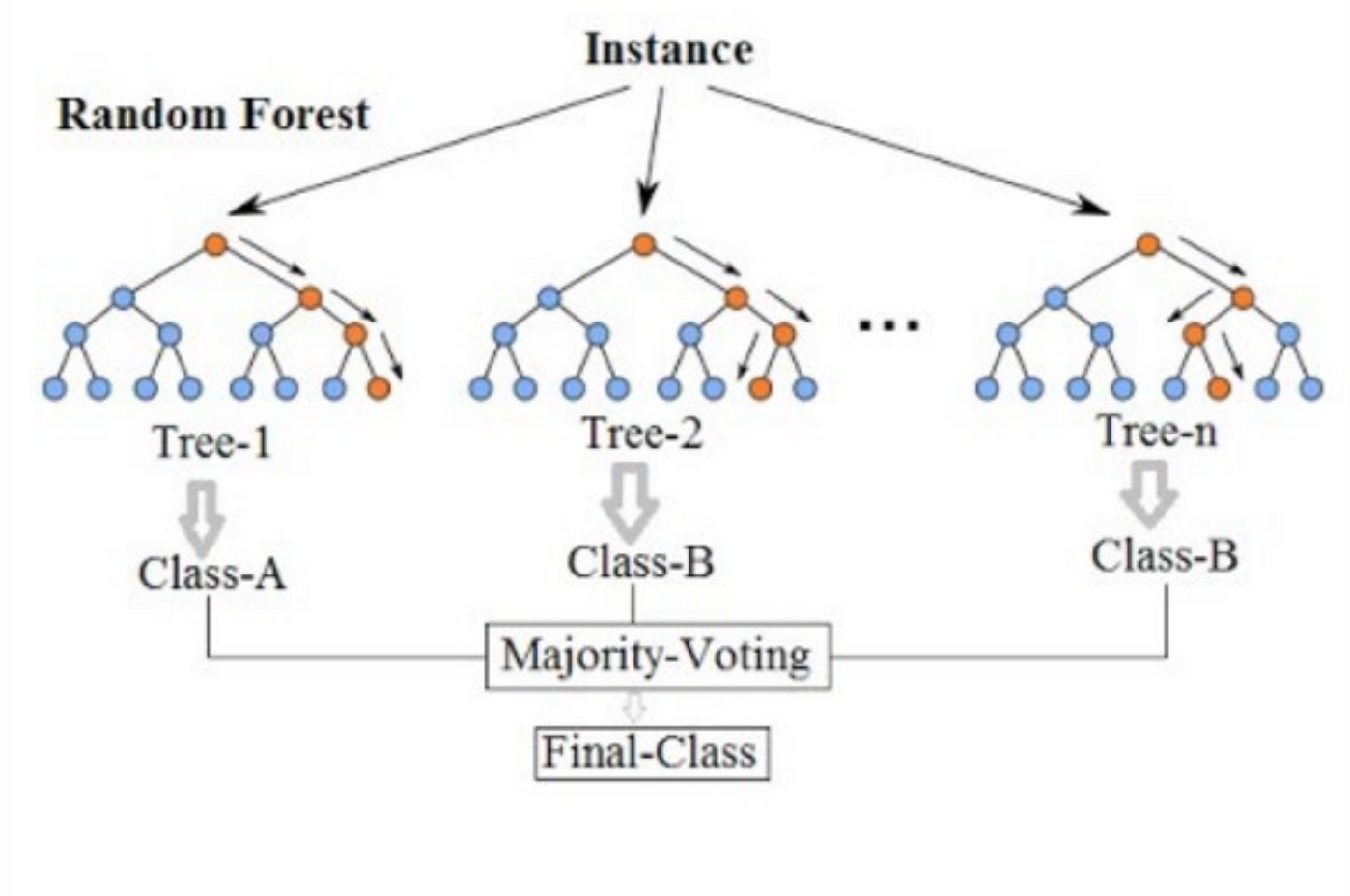


- The better the input features, the better the model

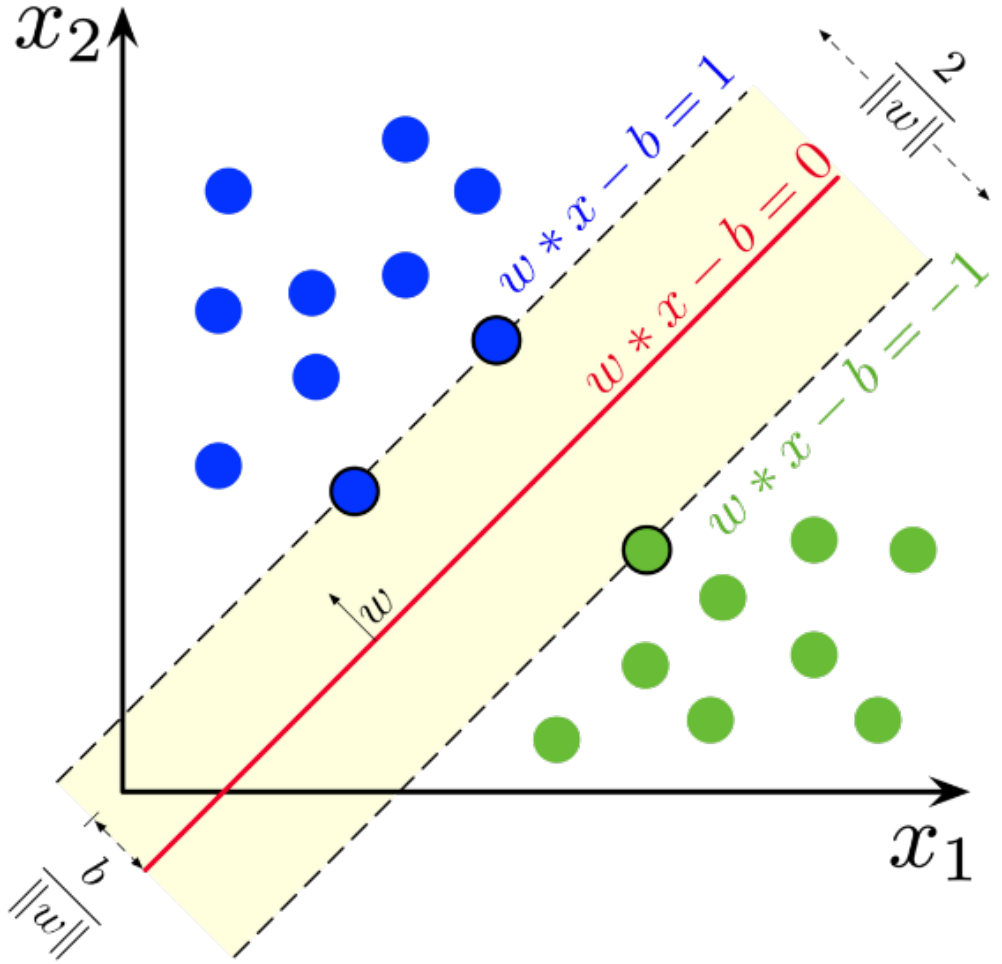
Linear and Logistic Regression



Random Forrest



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

Model	Features count	Training examples	Output classes
Logistic regression	1 to 10 million	No limit	Features x Classes < 10 million
Decision trees	1,000s	No limit	Features x Classes < 10,000
Random forest	10,000s	No limit	Features x Classes < 100,000s
Gradient-boosted trees	1,000s	No limit	Features x Classes < 10,000s

Model scalability for regression

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