Apache Spark Machine Learning Library

Spark MLlib

Spark

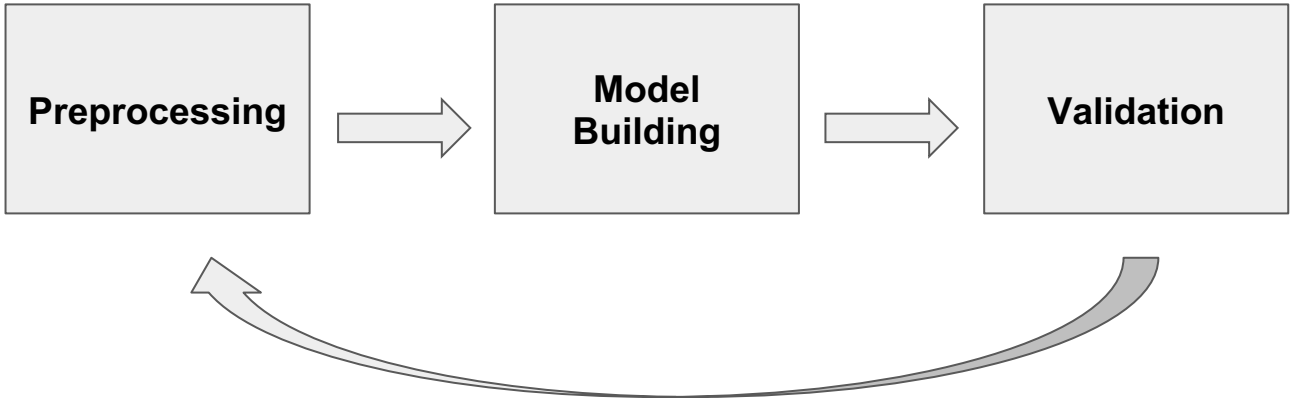
- A distributed,data processing platform for big data

- **Distributed:**
  - Runs in cluster of Servers
- **Processing:**
  - Performs computations, such as ETL and Modeling
- **Big Data:**
  - Terabyte and more volumes of data

Use Cases

- Real-time monitoring of financial data
- Text-analysis
- E-commerce pattern analysis
- Healthcare applications and genome analysis

Machine Learning Process



Machine Learning Process

1. **Preprocessing:**
    - **The first step is preprocessing which includes collecting, reformating, and transforming data.**
        - Extract, transform, and load data (it is similar to ETL in BI and data warehousing) to staging area
        - Review data to determine missing and invalid values
        - Normalizing or scaling numeric data
        - Standardize categorical values (e.g. 3-letter ISO code country names)

Machine Learning Process

1. **Model Building:**
    - **Applying machine learning algorithm to training data.**
        - Selecting algorithms (which works well with our data and use case)
        - Executing algorithms to fit data to the models
        - Tuning hyperparameters (some algorithms requires us to specify parameters, such as how many levels to have in decision tree.)

Machine Learning Process

1. **Validation:**
    - **Assess the quality of models**
        - **We can use:**
            - **Accuracy**
            - **Precision (positive predictive value)**
            - **Sensitivity (recall)**

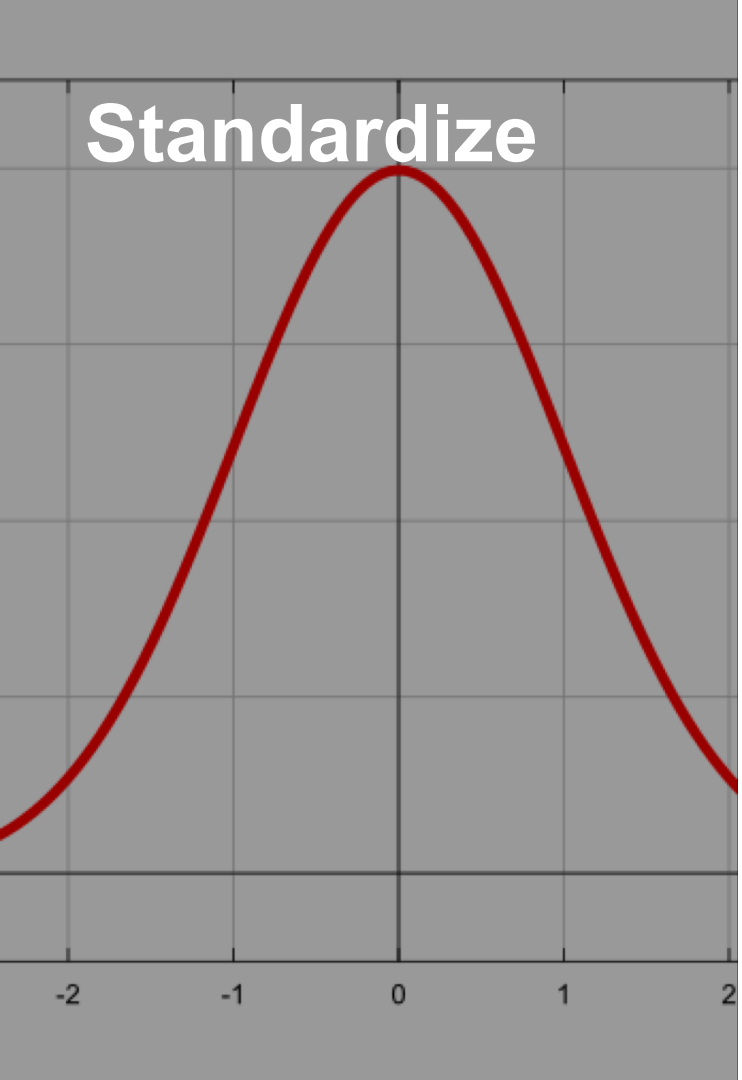
Normalizing

- Normalizing is the process of mapping numeric data from its original range into a range from zero to one.
- This is important, because you may have multiple attributes with different ranges.

Salary	Normalized Salary
60000	0
75000	0.33
90000	0.66
105000	1

Jump into coding

Normalization Notebook



- We may have a data that is pretty close to a bell-shaped curve or normally distributed.
- Standardization is the process of mapping data into a range of [-1,1]. (Mean is Zero)
- The main reason that we do this is some machine learning algorithms (i.e. SVM) work better when all of the features have unit variance and a zero mean.

Jump into coding

Standardize