Apache Spark

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

Spark Overview

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

What is Apache Spark?

- A fast and general engine for large-scale data processing
- A Open-source cluster computing framework
- End-to-End Analytics platform
- Developed to overcome limitations of Hadoop/Map Reduce
- Runs from a single desktop or a huge cluster
- Iterative, interactive or stream processing
- Supports multiple languages – Scala, Python, R, Java
- Major companies like Amazon, eBay, Yahoo use Spark.

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

When to use Spark?

- Data Integration and ETL
- Interactive Analytics
- High Performance Batch computation
- Machine Learning and Advanced Analytics
- Real time stream processing
- Example applications
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Advertisement Targeting

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Typical Spark workflow

- Load data from source
  - HDFS, NoSQL,S3, real time sources
- Transform Data
  - Filter, Clean, Join, Enhance
- Store processed data
  - Memory, HDFS, NoSQL
- Interactive Analytics
  - Shells, Spark SQL, third-party tools
- Machine Learning
- Action

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Online Reference

- http://spark.apache.org

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

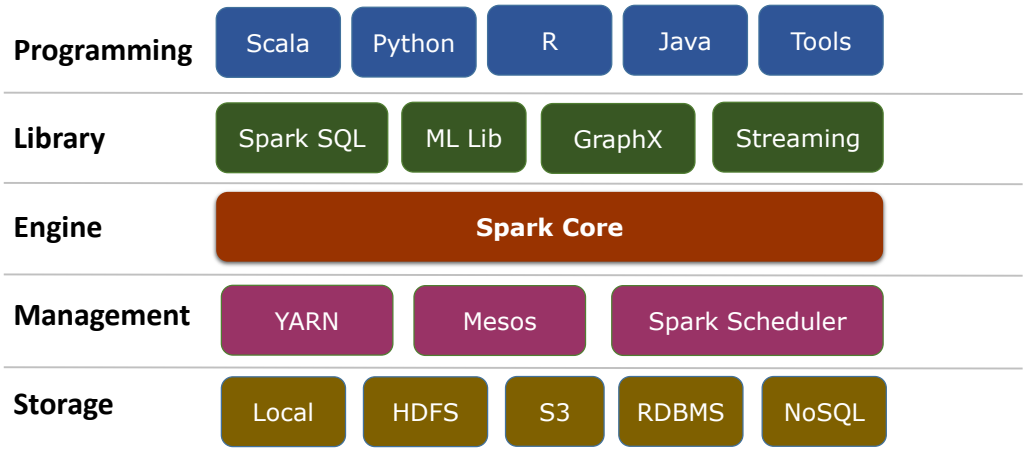
Spark Eco-system

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Spark Framework



Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

RDD

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Resilient Distributed Datasets (RDD)

- Spark is built around RDDs. You create, transform, analyze and store RDDs in a Spark program.
- The Dataset contains a collection of elements of any type.
  - Strings, Lines, rows, objects, collections
- The Dataset can be partitioned and distributed across multiple nodes
- RDDs are immutable. They cant be changed.
- They can be cached and persisted
- Transformations act on RDDs to create a new RDD
- Actions analyze RDDs to provide a result

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

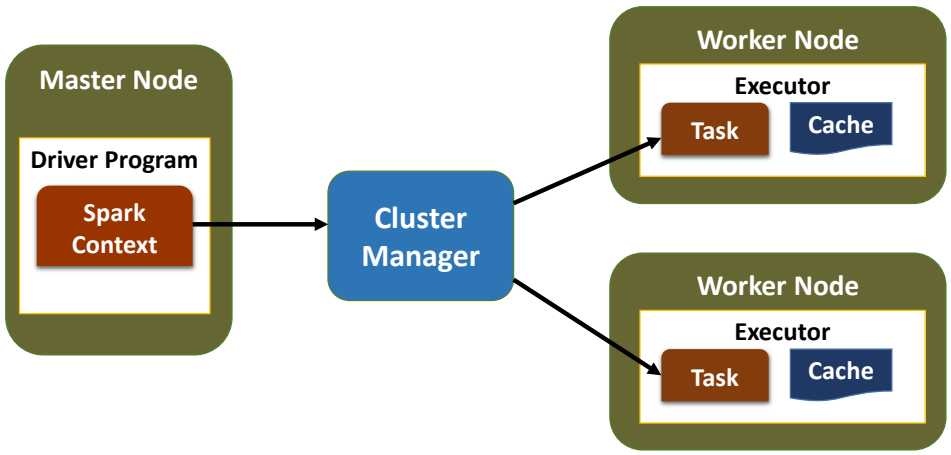
Spark Architecture

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Spark Architecture



Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Driver Program

- The main executable program from where Spark operations are performed
- Runs in the master node of a cluster
- Controls and co-ordinates all operations
- The Driver program is the “main” class.
- Executes parallel operations on a cluster
- Defines RDDs
- Each driver program execution is a “Job”

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

SparkContext

- Driver accesses Spark functionality through a SparkContext object.
- Represents a connection to the computing cluster
- Used to build RDDs.
- Partitions RDDs and distributes them on the cluster
- Works with the cluster manager
- Manages executors running on Worker nodes
- Splits jobs as parallel “tasks” and executes them on worker nodes
- Collects results and presents them to the Driver Program

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Spark modes

- Batch mode
  - A program is scheduled for execution through the scheduler
  - Runs fully at periodic intervals and processes data
- Interactive mode
  - An interactive shell is used by the user to execute Spark commands one-by-one.
  - Shell acts as the Driver program and provides SparkContext
  - Can run tasks on a cluster
- Streaming mode
  - An always running program continuously processes data as it arrives

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Spark scalability

- Single JVM
  - Runs on a single box (Linux or Windows)
  - All components (Driver, executors) run within the same JVM
- Managed Cluster
  - Can scale from 2 to thousands of nodes
  - Can use any cluster manager for managing nodes
  - Data is distributed and processed on all nodes

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Loading and Storing Data

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Creating RDDs

- RDDs can be created from a number of sources
  - Text Files
  - JSON
  - Parallelize() on local collections
    - Java Collections
    - Python Lists
    - R Data Frames
  - Sequence files
- RDBMS – load into location collections first and create RDD
- Very large data – create HDFS files outside of Spark and then create RDDs from them

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Storing RDDs

- Spark provides simple functions to persist RDDs to a variety of data sinks
  - Text Files
  - JSON
  - Sequence Files
  - Collections
- For optimization use language specific libraries for persistence than using Spark utilities.
- saveAsTextFile()
- RDBMS – move to local collections and then store.

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Lazy evaluation

- Lazy evaluation means Spark will not load or transform data unless an action is performed
  - Load file into RDD
  - Filter the RDD
  - Count no. of elements (only now loading and filtering happens)
- Helps internally optimize operations and resource usage
- Watch out during troubleshooting – errors found while executing actions might be related to earlier transformations

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Transformations

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Overview

- Perform operation on one RDD and create a new RDD
- Operate on one element at a time
- Lazy evaluation
- Can be distributed across multiple nodes based on the partitions they act upon

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Map

```
newRdd=rdd.map(function)
```

- Works similar to the Map Reduce "Map"
- Act upon each element and perform some operation
  - Element level computation or transformation
- Result RDD may have the same number of elements as original RDD
- Result can be of different type
- Can pass functions to this operation to perform complex tasks
- Use Cases
  - Data Standardization – First Name, Last Name
  - Data type conversion
  - Element level computations – compute tax
  - Add new attributes – Grades based on test scores

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

flatMap

```
newRdd=rdd.flatMap(function)
```

- Works the same way as map
- Can return more elements than the original map
- Use to break up elements in the original map and create a new map
  - Split strings in the original map
  - Extract child elements from a nested json string

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Filter

```
newRdd=rdd.filter(function)
```

- Filter a RDD to select elements that match a condition
- Result RDD smaller than the original RDD
- A function can be passed as a condition to perform complex filtering
  - Returns a true/false

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Set Operations

- Set operations are performed on two RDDs
- Union – Return a new dataset that contains the union of the elements in the source dataset and the argument.
  - `unionRDD=firstRDD.union(secondRDD)`
- Intersection - Return a new RDD that contains the intersection of elements in the source dataset and the argument.
  - `intersectionRDD=firstRDD.intersect(secondRDD)`

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Actions

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Introduction to actions

- Act on a RDD and product a result (not a RDD)
- Lazy evaluation – Spark does not act until it sees an action
- Simple actions
  - `collect` – return all elements in the RDD as an array. Use to trigger execution or print values
  - `count` – count the number of elements in the RDD
  - `first` – returns the first element in the RDD
  - `take(n)` – returns the first n elements

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

reduce

- Perform an operation across all elements of an RDD
  - sum, count etc.
- The operation is a function that takes as input two values.
- The function is called for every element in the RDD

```
inputRDD = [ a, b, c, d, e ]
and the function is  func(x,y)
func( func( func( func(a,b), c), d), e)
```

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Reduce Example

```
vals = [3,5,2,4,1]
sum(x,y) { return x + y }
sum( sum( sum( sum(3,5), 2), 4), 1) = 15
```

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Pair RDD

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Pair RDDs

- Pair RDDs are a special type of RDDs that can store key value pairs.
- Can be created through regular map operations
- All transformations for regular RDDs available for Pair RDDs
- Spark supports a set of special functions to handle Pair RDD operations
  - mapValues : transform each value without changing the key
  - flatMapValues : generate multiple values with the same key

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Pair RDD Actions

- countByKey – produces a count by each key in the RDD
- groupByKey – perform aggregation like sum, average by key
- reduceByKey – perform reduce, but by key
- aggregateByKey – perform aggregate by key
- Join - join multiple RDDs with the same key

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Advanced Spark

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Local Variables in Spark

- Spark makes copies of your code (one per partition) and executes them
- Any variable you create in the base programming language are local to a cluster
- Duplicate copies of local variables for each cluster
- Each variable acted upon independently

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Broadcast variables

- A read-only variable that is shared by all nodes
- Used for lookup tables or similar functions
- Spark optimizes distribution and storage for better performance.

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Accumulators

- A shared variable across nodes that can be updated by each node
- Helps compute items not done through reduce operations
- Spark optimizes distribution and takes care of race conditions

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Partitioning

- By default all RDDs are partitioned
  - spark.default.parallelism parameter
  - Default is the total no. of cores available across the entire cluster
- Should configure for large clusters
- Can be specified during RDD creation explicitly
- Derived RDD take the same number as the source.

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Persistence

- By default, Spark loads an RDD whenever it required. It drops it once the action is over
  - It will load and re-compute the RDD chain, each time a different operation is performed
- Persistence allows the intermediate RDD to be persisted so it need not have to be recomputed.
- persist() can persist the RDD in memory, disk, shared or in other third party sinks
- cache() provides the default persist() – in memory

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Spark SQL

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Overview

- A library built on Spark Core that supports SQL like data and operations
- Make it easy for traditional RDBMS developers to transition to big data
- Works with “structured” data that has a schema
- Seamlessly mix SQL queries with Spark programs.
- Supports JDBC
- Helps “mix” n “match” different RDBMS and NoSQL Data sources

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Spark Session

- All functionality for Spark SQL accessed through a Spark Session
- Data Frames are created through Spark Session
- Provides a standard interface to work across different data sources
- Can register Data Frames as temp table and then run SQL queries on them

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

DataFrame

- A distributed collection of data organized as rows and columns
- Has a schema – column names, data types
- Built upon RDD, Spark optimizes better since it knows the schema
- Can be created from and persisted to a variety of sources
  - CSV
  - Database tables
  - Hive / NoSQL tables
  - JSON
  - RDD

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Operations supported by Data Frames

- `filter` – filter data based on a condition
- `join` – join two Data Frames based on common column
- `groupBy` – group data frames by specific column values
- `agg` – compute aggregates like sum, average.
- Operations can be nested.

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Temp Tables

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Temp tables / Views

- Provide SQL table like operations
- A “wrapper” around a data frame
- Execute ANSI SQL Queries against temp tables.
- Very simple and yet very powerful
- A query on a Temp table generates another data frame
- `createOrReplaceTempView` – register the Data Frame as a table within SQL Session
- SparkSession provides the ability to execute SQL

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Spark Streaming

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Why Spark Streaming?

- Typically, Analytics is performed on data at rest
  - Databases, flat files etc.
- Some use cases require real time analytics
  - Fraud detection, click stream processing
- Spark Streaming is built for this purpose
- Spark Streaming helps you to
  - Look at data as they are created/arrive from source
  - Transform, summarize, analyze
  - Perform machine learning
  - Predict in real time

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

When to use Spark Streaming

- Credit Card Fraud Detection
- Spam Filtering
- Network Intrusion Detection
- Real time social media analytics
- Click Stream analytics and recommendations
- Ad recommendations
- Stock Market analytics

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

What sources does Spark Streaming supports?

- Flat files (as they are created)
- TCP/IP
- Apache Flume
- Apache Kafka
- Amazon Kinesis
- Twitter, Facebook and other social media

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

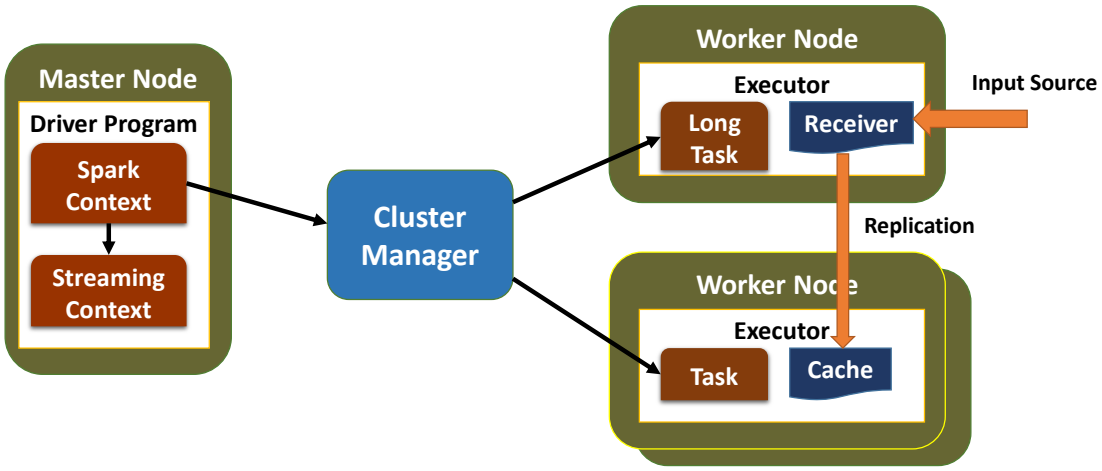
Streaming Architecture

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Spark Streaming Architecture



Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

DStreams

- A Streaming context is created from the Spark context to enable streaming
- Streaming creates a DStream (Discretized Stream) on which processing occurs
- A micro-batch window is setup for the DStream
- Data is received, accumulated as a micro-batch and processed as a micro-batch
- Each micro-batch is an RDD
- Regular RDD operations can be applied on the DStream RDD

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

DStream processing

- Spark collects incoming data for each interval (micro-batch)
- Data is collected as an RDD for that interval.
- It then calls all transformations and operations that applies for that DStream or derived DStreams
- Global variables can be used to track data across DStreams
- Windowing functions are available for computing across multiple DStreams.
  - Window size multiple of interval
  - Sliding interval multiple of interval

Copyright @2016 V2 Maestros, All rights reserved.

Types of Analytics

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

Machine Learning with Spark

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Overview

- Make ML practical and easy.
- Contains algorithms and utilities
- Packages
  - spark.mllib – original APIs built on RDDs
  - spark.ml – new higher level API built on DataFrames (Spark SQL) and pipelines
- http://spark.apache.org/docs/latest/ml-guide.html
- Special data types
  - Local vector
  - Labeled Point

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Local Vector

- A vector of double values
- Dense Vector
  - (1.0, 3.0, 4.5)
- Sparse Vector
  - Original : (1.0,0.0,0.0,2.0,0.0)
  - Representation: (5, (0,3), (1.0,2.0)

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Labeled Point

- Represents a Data point in ML
- Contains a “label” (the target variable) and a list of “features” (the predictors)
- LabeledPoint(1.0, Vectors.dense(1.0,0.0,3.0))

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Pipelines

- A pipeline consist of a series of transformations and actions that need to be performed to create a model
  - DataFrame (the source)
  - Transformers (data transformations)
  - Estimators (model building)
  - Parameters (common parameters across algorithms)
- Internally optimized for better parallelism and resource utilization

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Spark Project Work flow

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Copyright @2016 V2 Maestros, All rights reserved.

V2 Maestros
The Data Science Experts

Project Work flow

- Problem statement and Questions to Answer
- Acquire Data
- Process Data
  - Cleanse, Filter, Augment
- Exploratory Analysis
- Machine Learning
  - Correlation Analysis
  - Training and Test split
  - Model building
  - Prediction
- Action
- Feedback

Copyright @2016 V2 Maestros, All rights reserved.

Analytics and Predictions

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

Types of Analytics

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

V2 Maestros
The Data Science Experts

Types of Analytics

Type of Analytics	Description
Descriptive	Understand what happened
Exploratory	Find out why something is happening
Inferential	Understand a population from a sample
Predictive	Forecast what is going to happen
Causal	What happens to one variable when you change another
Deep	Use of advanced techniques to understand large and multi-source datasets

V2 Maestros
The Data Science Experts

Exploratory Data Analysis

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

V2 Maestros
The Data Science Experts

Goals of EDA

- Understand the predictors and targets in the data set
  - Spreads
  - Correlations
- Uncover the patterns and trends
- Find key variables and eliminate unwanted variables
- Detect outliers
- Validate previous data ingestion processes for possible mistakes
- Test assumptions and hypothesis

V2 Maestros
The Data Science Experts

Tools used for EDA

- Correlation matrices
- Boxplots
- Scatterplots
- Principal component Analysis
- Histograms

V2 Maestros
The Data Science Experts

Machine Learning

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

V2 Maestros
The Data Science Experts

Overview

- Data contains attributes
- Attributes show relationships (correlation) between entities
- Learning – understanding relationships between entities
- Machine Learning – a computer analyzing the data and learning about relationships
- Machine Learning results in a model built using the data
- Models can be used for grouping and prediction

V2 Maestros
The Data Science Experts

Data for machine learning

- Machines only understand numbers
- Text Data need to be converted to equivalent numerical representations for ML algorithms to work.
- Number representation
  - (Excellent, Good, Bad can be converted to 1,2,3)
- Boolean variables
  - 3 new Indicator variables called Rating-Excellent, Rating-Good, Rating-Bad with values 0/1
- Document Term matrix

V2 Maestros
The Data Science Experts

Unsupervised Learning

- Finding hidden structure / similarity / grouping in data
- Observations grouped based on similarity exhibited by entities
- Similarity between entities could be by
  - Distance between values
  - Presence / Absence
- Types
  - Clustering
  - Association Rules Mining
  - Collaborative Filtering

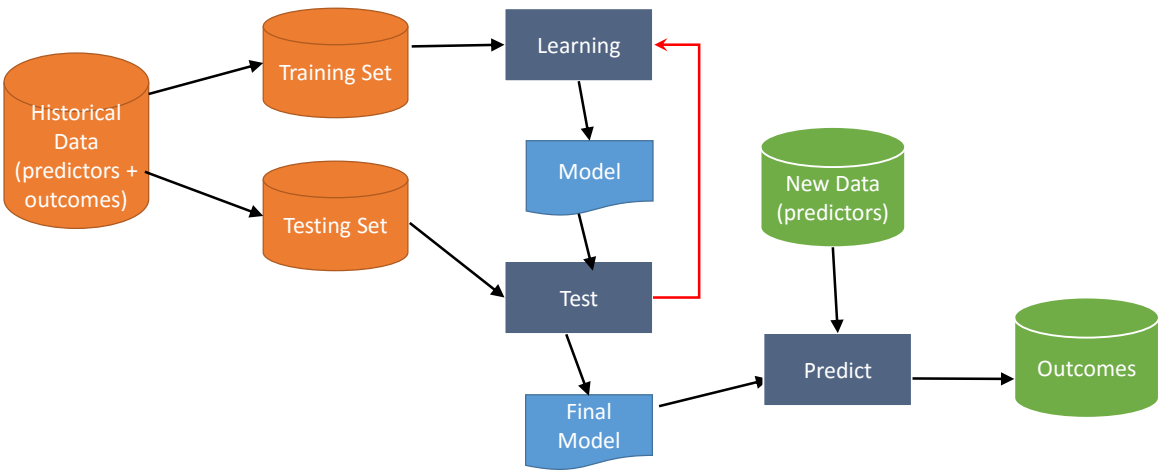
V2 Maestros
The Data Science Experts

Supervised Learning

- Trying to predict unknown data attributes (outcomes) based on known attributes (predictors) for an entity
- Model built based on training data (past data) where outcomes and predictors are known
- Model used to predict future outcomes
- Types
  - Regression (continuous outcome values)
  - Classification (outcome classes)

V2 Maestros
The Data Science Experts

Supervised Learning Process



V2 Maestros
The Data Science Experts

Training and Testing Data

- Historical Data contains both predictors and outcomes
- Split as training and testing data
- Training data is used to build the model
- Testing data is used to test the model
  - Apply model on testing data
  - Predict the outcome
  - Compare the outcome with the actual value
  - Measure accuracy
- Training and Test fit best practices
  - 70-30 split
  - Random selection of records. Should maintain data spread in both datasets

V2 Maestros
The Data Science Experts

Comparing Results

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

V2 Maestros
The Data Science Experts

Confusion Matrix

- Plots the predictions against the actuals for the test data
- Helps understand the accuracy of the predictions
- Predictions can be Boolean or classes

		Actual		
		TRUE	**FALSE**	**Total**
Prediction	**TRUE**	44	6	50
	FALSE	9	41	50
	Total	53	47	100

V2 Maestros
The Data Science Experts

Prediction Types

- The importance of prediction types vary by the domain
- True Positive (TP) and True Negative (TN) are the correct predictions
- False Negative (FN) can be critical in medical field
- False Positive (FP) can be critical in judicial field

		Actual	
		TRUE	**FALSE**
Prediction	**TRUE**	True Positive	False Positive
	FALSE	False Negative	True Negative

V2 Maestros
The Data Science Experts

Confusion Matrix metrics

- Accuracy
  - Measures the accuracy of the prediction
  - Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Sensitivity
  - Hit rate or recall
  - Sensitivity = TP / (TP + FN)
- Specificity
  - True negative rate
  - Specificity = TN / (TN + FP)
- Precision
  - Precision = TP / (TP + FP)

		Actual	
		TRUE	**FALSE**
Prediction	**TRUE**	True Positive	False Positive
	FALSE	False Negative	True Negative

V2 Maestros
The Data Science Experts

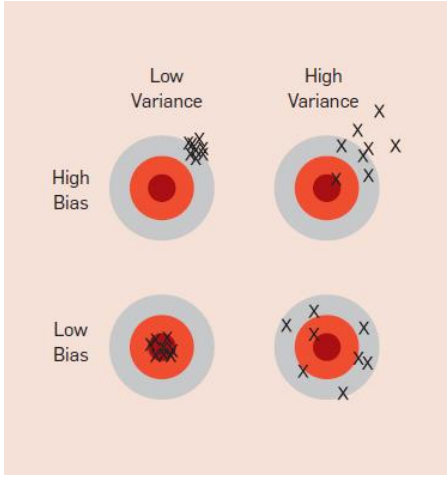
Prediction Errors

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

V2 Maestros
The Data Science Experts

Bias and Variance

- Bias happens when the model "skews" itself to certain aspects of the predictors, while ignoring others. It is the error between prediction and actuals.
- Variance refers to the stability of a model – Keep predicting consistently for new data sets. It is the variance between predictions for different data sets.



V2 Maestros
The Data Science Experts

Types of Errors

- In-Sample error is the prediction error when the model is used to predict on the training data set it is built upon.
- Out-of-sample error is the prediction error when the model is used to predict on a new data set.
- Over fitting refers to the situation where the model has very low in-sample error, but very high out-of-sample error. The model has "over fit" itself to the training data.

V2 Maestros
The Data Science Experts

Linear Regression

Linear Relationships

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

V2 Maestros
The Data Science Experts

Regression Analysis

- Method of investigating functional relationship between variables
- Estimate the value of dependent variables from the values of independent variables using a relationship equation
- Used when the dependent and independent variables are continuous and have some correlation.
- Goodness of Fit analysis is important.

V2 Maestros
The Data Science Experts

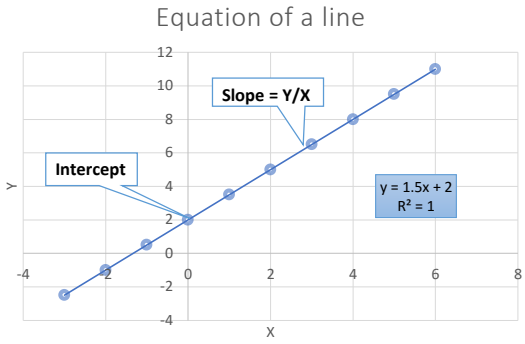
Linear Equation

- X is the independent variable
- Y is the dependent variable
- Compute Y from X using

$$Y = \alpha X + \beta$$

Coefficients:

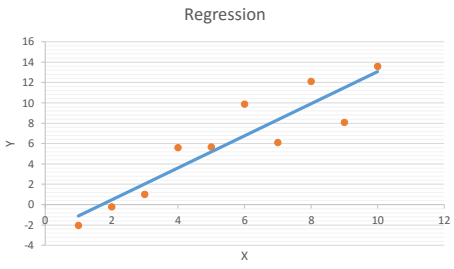
- α = Slope = Y/X
- β = Intercept = value of Y when X=0

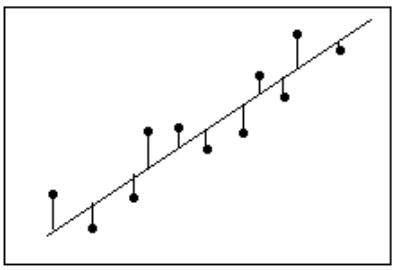


V2 Maestros
The Data Science Experts

Fitting a line

- Given a scatter plot of Y vs X, fit a straight line through the points so that the sum of square of vertical distances between the points and the line (called residuals) is minimized
- Best line = least residuals
- A line can always be fitted for any set of points
- The equation of the line becomes the predictor for Y

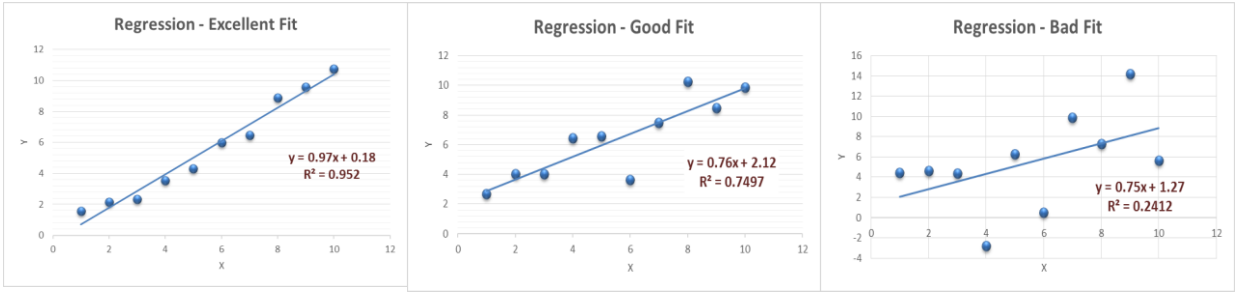




V2 Maestros
The Data Science Experts

Goodness of Fit

- R-squared measures how close the data is to the fitted line
- R-squared varies from 0 to 1. The higher the value, the better the fit
- You can always fit a line. Use R-squared to see how good the fit is
- Higher correlation usually leads to better fit



V2 Maestros
The Data Science Experts

Multiple regression

- When there are more than one independent variable that is used to predict the dependent variable.
- The equation $Y = \beta + \alpha_1 * X_1 + \alpha_2 * X_2 + \ldots + \alpha_p * X_p$
- Same process used for prediction as a single independent variable
- Different predictors have different levels of impact on the dependent variable

V2 Maestros
The Data Science Experts

Using Linear Regression for ML

- ML Technique to predict continuous data – supervised learning
- Predictors and outcomes provided as input
- Data analyzed (training) to come up with a linear equation
  - Coefficients
  - Intercept
  - R-squared
- Linear equation represents to model.
- Model used for prediction
- Typically fast for model building and prediction

V2 Maestros
The Data Science Experts

Summary – Linear Regression

Advantages

- Fast
- Low cost
- Excellent for linear relationships
- Relatively accurate Continuous variables

Shortcomings

- Only numeric/ continuous variables
- Cannot model non-linear / fuzzy relationships
- Sensitive to outliers

Used in

- Oldest predictive model used in a wide variety of applications to predict continuous values

V2 Maestros
The Data Science Experts

Decision Trees

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

V2 Maestros
The Data Science Experts

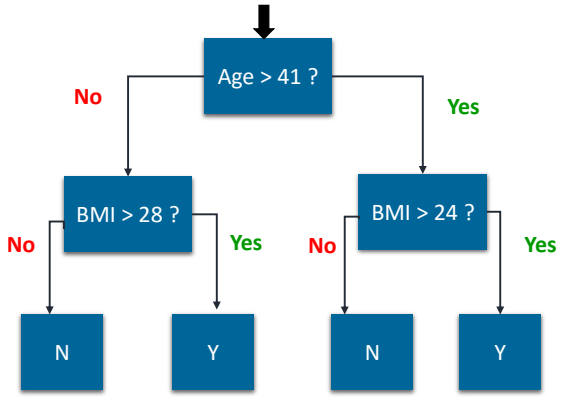
Overview

- The simplest, easy to understand and easy to explain ML technique.
- Predictor variables are used to build a tree that would progressively predict the target variable
  - Trees start with a root node that start the decision making process
  - Branch nodes refine the decision process
  - Leaf nodes provide the decisions
- Training data is used to build a decision tree to predict the target
- The tree becomes the model that is used to predict on new data

V2 Maestros
The Data Science Experts

Example

Age	BMI	Is Diabetic
24	22	N
33	28	N
41	36	Y
48	24	N
58	31	Y
61	35	Y



Predicting "Is Diabetic ?"

V2 Maestros
The Data Science Experts

Choosing the right Predictors

- The depth of trees are highly influenced by the sequence in which the predictors are chosen for decisions
- Using predictors with high selectivity gives faster results
- ML implementations automatically make decisions on the sequence /preference of predictors

V2 Maestros
The Data Science Experts

Summary – Decision Trees

Advantages

- Easy to interpret and explain
- Works with missing data
- Sensitive to local variations
- Fast

Shortcomings

- Limited Accuracy
- Bias builds up pretty quickly
- Not good with large predictors

Used in

- Credit approvals
- Situations with legal needs to explain decisions
- Preliminary categorization

V2 Maestros
The Data Science Experts

Naïve Bayes

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

V2 Maestros
The Data Science Experts

Bayes' theorem (too) simplified

- Probability of an event A = P(A) is between 0 and 1
- Bayes' theorem gives the conditional probability of an event A given event B has already occurred.

  P(A/B) = P(A intersect B) * P (A) /P(B)
- Example
  - There are 100 patients
  - Probability of a patient having diabetes is P(A) = .2
  - Probability of patient having diabetes (A) given that the patient's age is > 50 (B) is P(A/B) = .4

V2 Maestros
The Data Science Experts

Naïve Bayes Classification

- Application of Bayes' theorem to ML
- The target variable becomes event A
- The predictors become events B1 – Bn
- We try to find P(A / B1-Bn)

Age	BMI	Is Diabetic	
24	22	N	Probability of Is Diabetic = Y given that Age = 24 and BMI = 22
41	36	Y	Probability of Is Diabetic – Y given that Age = 41 and BMI = 36

V2 Maestros
The Data Science Experts

Model building and prediction

- The model generated stores the conditional probability of the target for every possible value of the predictor.

	Overall	Age						Gender	
Salary		1 to 20	20 to 30	30 to 40	40 to 50	50 to 60	60 to 100	Female	Male
< 50K	.75	0.1	0.3	0.25	0.17	0.1	0.08	0.39	0.61
> 50K	.25	0.03	0.08	0.3	0.32	0.2	0.07	0.15	0.85
Overall		.08	.24	.26	.21	.12	.08	.33	.67

- When a new prediction needs to be done, the conditional probabilities are applied using Bayes' formula to find the probability
  - To predict for Age = 25
  - P(Salary < 50K / Age=25) = 0.3 * 0.75 / 0.24 = ~ 0.92
  - P(Salary > 50K / Age=25) = 0.08 * 0.25 / 0.24 = ~ 0.08

V2 Maestros
The Data Science Experts

Summary – Naïve Bayes

Advantages

- Simple and fast
- Works well with noisy and missing data
- Provides probabilities of the result
- Very good with categorical data

Shortcomings

- Limited Accuracy
- Expects predictors to be independent
- Not good with large numeric features

Used in

- Medical diagnosis
- Spam filtering
- Document classification
- Sports predictions

V2 Maestros
The Data Science Experts

Random Forests

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

V2 Maestros
The Data Science Experts

Overview

- Random Forest is one of the most popular and accurate algorithms
- It is an Ensemble method based on decision trees
  - Builds multiple models – each model a decision tree
  - For prediction – each tree is used to predict an individual result
  - A vote is taken on all the results to find the best answer

V2 Maestros
The Data Science Experts

How it works

- Lets say the dataset contains m samples (rows) and n predictors (columns)
- x trees are built, each with a subset of data
- For each tree, a subset of m rows and n columns are chosen randomly.
- For example, if the data has 1000 rows and 5 columns, each tree is built using 700 rows and 3 columns
- The data subset is used to build a tree
- For prediction, new data is passed to each of the x trees and x possible results obtained
- For example, if we are predicting buy=Y/N and there are 500 trees, we might get 350 Y and 150 N results
- The most found result is the aggregate prediction.

V2 Maestros
The Data Science Experts

Summary – Random Forest

Advantages

- Highly accurate
- Efficient on large number of predictors
- Fully parallelizable
- Very good with missing data

Shortcomings

- Time and Resource consuming
- For categorical variables, bias might exist if levels are disproportionate

Used in

- Scientific Research
- Competitions
- Medical Diagnosis

V2 Maestros
The Data Science Experts

K-means Clustering

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

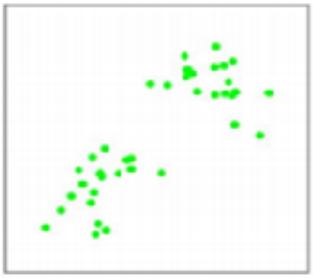
V2 Maestros
The Data Science Experts

Overview

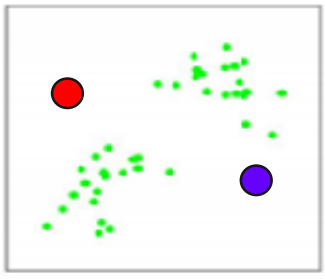
- Unsupervised Learning technique
- Popular method for grouping data into subsets based on the similarity
- Partitions n observations with m variables into k clusters where by each observation belongs to only one cluster
- How it works
  - An m dimensional space is created
  - Each observation is plotted based on this space based on the variable values
  - Clustering is done by measuring the distance between points and grouping them
- Multiple types of distance measures available like Euclidian distance and Manhattan distance

V2 Maestros
The Data Science Experts

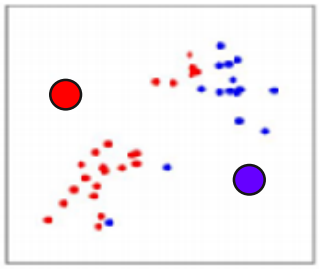
Clustering - Stages



- Dataset contains only m=2 variables. We will create k=2 clusters
- Plot observations on a two dimensional plot



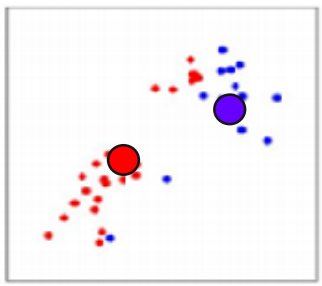
- Choose k=2 centroids at random
- Measure the distance between each observation to each centroid



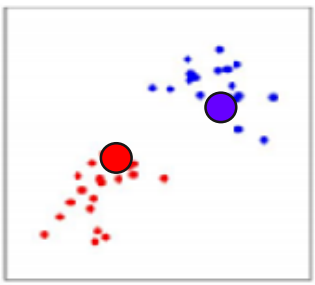
- Assign each observation to the nearest centroid
- This forms the clusters for round 1

V2 Maestros
The Data Science Experts

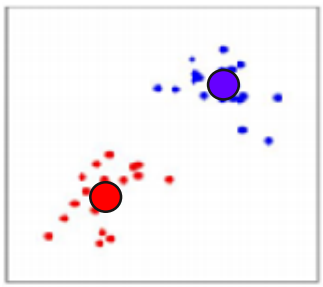
Clustering - Stages



- Find the centroid of each of the cluster
- Centroid is the point where the sum of distances between the centroid and each point is minimum



- Repeat the process of finding the distance between each observation to each centroid (the new one) and reassign each point to the nearest one



- Find the centroid for the new clusters
- Repeat the process until the centroids don't move

V2 Maestros
The Data Science Experts

Summary – K-means clustering

Advantages

- Fast
- Efficient with large number of variables
- Explainable

Shortcomings

- K needs to be known
- The initial centroid position has influence on clusters formed

Used in

- Preliminary grouping of data before other classification
- General grouping
- Geographical clustering

V2 Maestros
The Data Science Experts

Collaborative Filtering

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

V2 Maestros
The Data Science Experts

Dimensionality Reduction

Principal Component Analysis

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

V2 Maestros
The Data Science Experts

Issues with too many predictors

- Memory requirements
- CPU requirements / time taken for machine learning algorithms
- Correlation between predictors
- Over fitting
- Some ML algorithms don't work fine with too many predictors

V2 Maestros
The Data Science Experts

Manual selection

- Using domain knowledge
  - Purely based on hypothesis
  - Risky – there could be unknown correlations
- Using Correlation co-efficients
  - Variables with good correlation can only be picked up.
- Using Decision Trees
  - Decision trees are fast and choose variables based on correlation
  - Variables used in the decision trees can be picked for further processing

V2 Maestros
The Data Science Experts

Principal Component Analysis

- Used to reduce the number of predictors
- Based on Eigen Vectors and Eigen Values.
- Given a set of M predictors, PCA transforms this to a set of N predictors such that N < M
- The new predictors are derived predictors called PC1, PC2, PC3
- The new predictors retain similar levels of correlation and predictability like the original predictors

V2 Maestros
The Data Science Experts

Recommendation Engines

Unauthorized copying, distribution and exhibition of this presentation is punishable under law

V2 Maestros
The Data Science Experts

What is a Recommendation Engine?

- Also called Collaborative filtering
- Analyze past data to understand user / entity behavior
- Identify "similar" items /users / entities
- Recommend based on similarity of behavior
- Example
  - Tom and Chris both like similar items. In the past 1 year, Tom has brought 42 items and Chris has brought 35 items. 28 of these are same.
  - Tom buys a new item which Chris has not bought. Recommend that to Chris.
- Used by
  - Netflix for movie recommendations
  - Amazon for product recommendations
  - YouTube for video recommendations

V2 Maestros
The Data Science Experts

Recommendation Types

- User based Recommendations
  - Identify similar users and form User neighborhoods.
  - If one user buys a new product, recommend that to all users in the neighborhood.
  - "Similar customers brought…"
- Item based Recommendations
  - Identify Items that are frequently brought together.
  - If one item is brought by a user, recommend the related items to the user.
  - "Users who brought this item also brought…"

V2 Maestros
The Data Science Experts

Input for Recommenders

- Recommender algorithms take as input as specific format
  - User ID
  - Item ID
  - Score
- Scores indicate relative preference of the user for the item
  - Boolean values
  - Rating scores
  - Measure of sales volume

V2 Maestros
The Data Science Experts

Building a User based recommender

- Find affinity scores between users based on similarity of behavior.
  - Uses similarity measures like cosine similarity, Pearson correlation etc.
- For user neighborhoods with each neighborhood containing users with high inter-user scores.
- If one user shows a new behavior (buys an item), recommend that to other users in the neighborhood.
- Continuous processing of past data and building of neighborhoods.

V2 Maestros
The Data Science Experts

Building a Item based recommender

- Find affinity scores between items based on usage (used together).
  - Uses similarity measures like cosine similarity, Pearson correlation etc.
- If one item is brought by a user, recommend items with high similarity scores with that item.
- Continuous processing of past data and building of neighborhoods.
- Item based recommenders are superior to user based, since
  - No. of items are limited
  - More usage data available since the list of items are relatively static.