IT4931
Tích hợp và xử lý dữ liệu lớn

IT4931

12/2021
Thanh-Chung Dao Ph.D.

Spark ML

- Spark ML is a machine-learning library
  - Classification: logistic regression, naive Bayes
  - Regression: generalized linear regression, survival regression
  - Decision trees, random forests, and gradient-boosted trees
  - Recommendation: alternating least squares (ALS)
  - Clustering: K-means, Gaussian mixtures (GMMs)
  - Topic modeling: latent Dirichlet allocation (LDA)
  - Frequent item sets, association rules, and sequential pattern mining

Classification vs Prediction

- Classification models predict categorical class labels [2]
  - Binary classification

- Prediction models predict continuous valued functions
  - Regression analysis is a statistical methodology that is most often used for numeric prediction

Predicting the arrival delay of commercial flights [1]

- Problem
  - We want to be able to predict, based on historical data
    - The arrival delay of a flight using only information available before the flight takes off

- Dataset
  - http://stat-computing.org/dataexpo/2009/the-data.html
  - The data used was published by the US Department of Transportation
  - It compromises almost 23 years worth of data

- Approach
  - Using a regression algorithm

Dataset Information

	Name	Description
1	Year	1987-2008
2	Month	1-12
3	DayofMonth	1-31
4	DayOfWeek	1 (Monday) - 7 (Sunday)
5	DepTime	actual departure time (local, hhmm)
6	CRSDepTime	scheduled departure time (local, hhmm)
7	ArrTime	actual arrival time (local, hhmm)
8	CRSArrTime	scheduled arrival time (local, hhmm)
9	UniqueCarrier	unique carrier code
10	FlightNum	flight number
11	TailNum	plane tail number
12	ActualElapsedTime	in minutes
13	CRSElapsedTime	in minutes
14	AirTime	in minutes
15	ArrDelay	arrival delay, in minutes
16	DepDelay	departure delay, in minutes
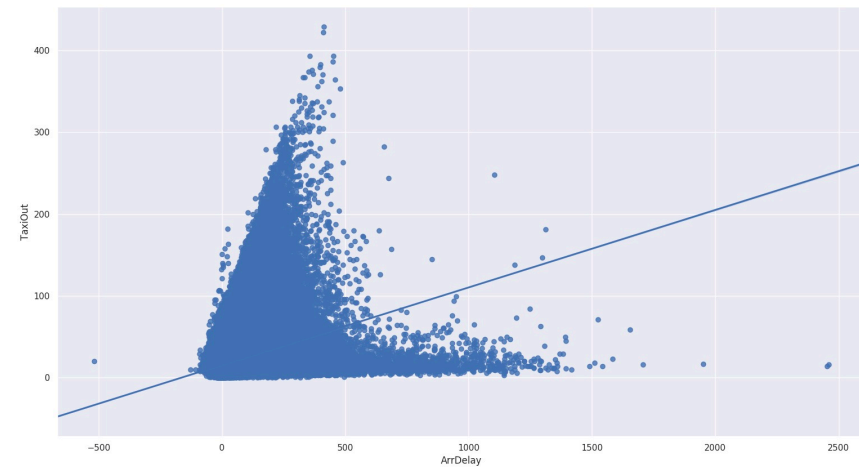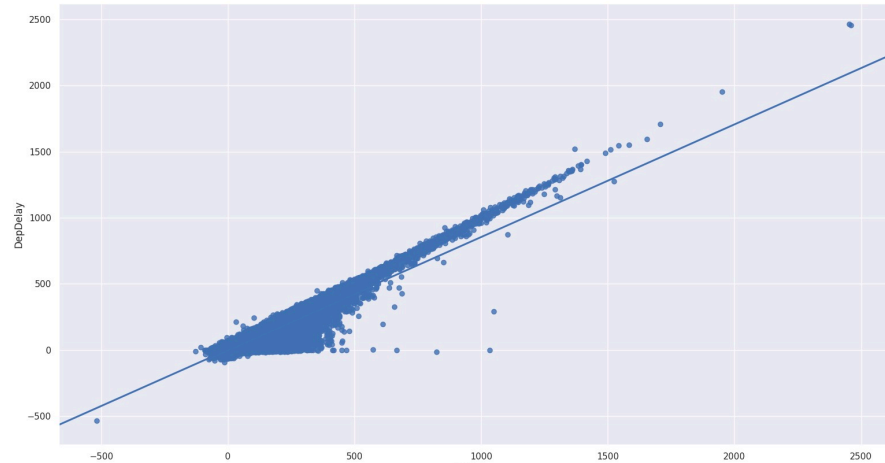
Analyzing Flow

- Load data
- Preprocess Data
- Train the data and obtain a model
- Evaluate the resulting model
- Make Predictions

Lab: Running Prediction of Flight Delay on Zeppelin

- Write code using PySpark
  - Get the prepared notebook

- Run and try to understand algorithms

- The original source code (in Scala)
  - https://github.com/pedroduartecosta/Spark-PredictFlightDelay

References

- [1] https://medium.com/@pedrodc/building-a-big-data-machine-learning-spark-application-for-flight-delay-prediction-4f9507cdb010

- [2] https://www.tutorialspoint.com/data_mining/dm_classification_prediction.htm