BerkeleyX: CS110x Big Data Analysis with Apache Spark

Bookmarks

Week 1 - Big Data and Data Science

Lecture 1: Big Data and Data Science
Quizzes

Setting up the Course Software Environment
Setup

Lab 1: Power Plant Machine Learning Pipeline
Lab due Sep 13, 2016 at 04:30 IST

Lab 1 Quiz Questions
Quizzes

Week 1 - Big Data and Data Science > Lecture 1: Big Data and Data Science > Business Questions, Statistics, and Exploratory Data Analysis

Bookmark

Business Questions, Statistics, and Exploratory Data Analysis

BERCS1102016-V000200

Here is a good description of the difference between descriptive and inferential statistics.

SUPERVISED LEARNING

- kNN (k Nearest Neighbors)
- Naive Bayes
- Logistic Regression
- Support Vector Machines
- Random Forests

UNSUPERVISED LEARNING

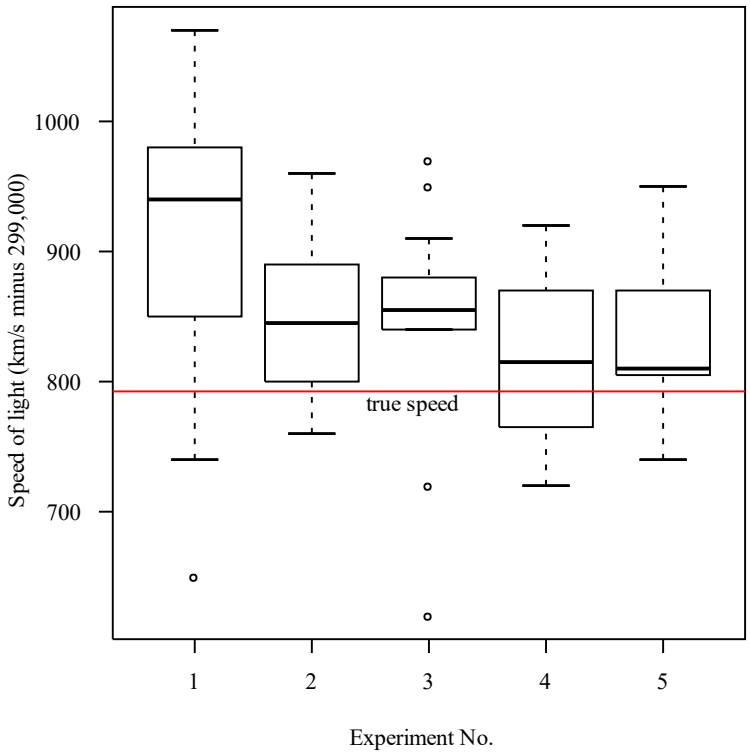
- Clustering
- Factor Analysis
- Latent Dirichlet Allocation

The US National Institute of Standards and Technology has an excellent primer on Exploratory Data Analysis.

The five-number summary is a descriptive statistic that provides information about a set of observations. It consists of the five most important sample percentiles:

1. The sample minimum (smallest observation)
2. The lower quartile or *first quartile*
3. The median (middle value)
4. The upper quartile or *third quartile*
5. The sample maximum (largest observation)

You can compare the five-number summaries of multiple observations using a box plot:



S Programming Language

(1/1 point)

Where was the S programming language invented?

- Google

- University of Auckland, New Zealand
- UC Berkeley
- Bell Labs ✔
- Dartmouth College

EXPLANATION

The S programming language was developed at Bell Labs for Exploratory Data Analysis.

Spark 1.4 introduced SparkR (R on Spark). SparkR provides a distributed data frame implementation that supports operations like selection, filtering, aggregation etc. (similar to R data frames) but on large datasets.

Exploratory Data Analysis

(1/1 point)
Which of the following is NOT a typical Exploratory Data Analysis activity?

- Visualizing data distributions
- Calculating summary statistics
- Examining data distributions
- Fitting a Support Vector Machine ✔

EXPLANATION

The activities of EDA include visualizing the data distributions, calculating summary statistics for the data, and examining the distributions of the data.


Download on the
App Store
ANDROID APP ON
Google play