Stata for Big Data and Data Science

Overview and Prospects

Debora Bilard

Economist, Associate Consultant at Timberlake Consultants

www.linkedin.com/in/deborabilard

d.e.kusmerskibilard@gmail.com

Stata Users Group Meeting, FEA - USP, Sao Paulo, December 2, 2016

Overview

The conceptual framework - Statistical analysis

A comparison of approaches

Big Data e Data Science – an overview

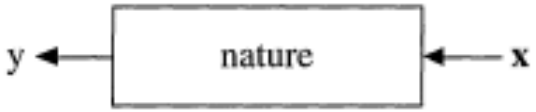
Selected topics of Data Science and Stata tools

Prospects

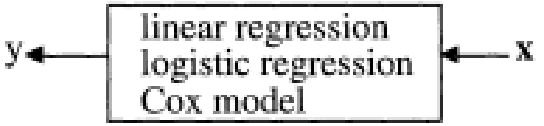
The conceptual framework
Statistical analysis

Breiman (2001), “Statistical Modeling: The Two Cultures”, link

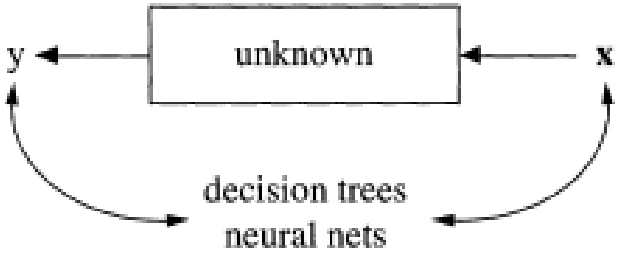
Goals: Inference and Prediction



Approaches: 1. Model

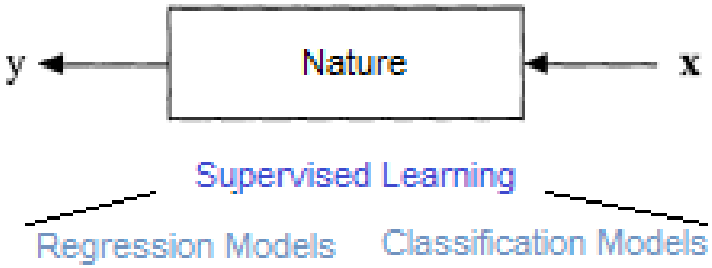


2. Algorithm



The overfitting problem -> training and test samples

A comparison of approaches



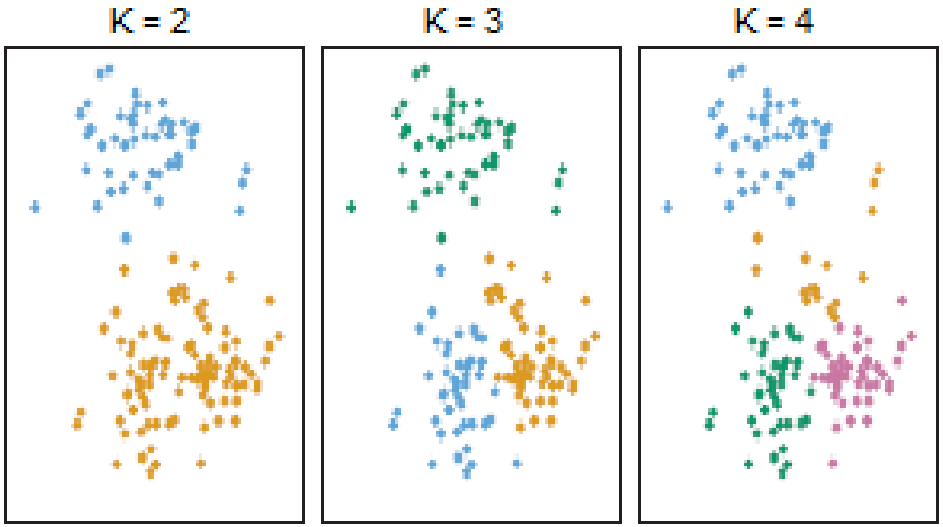
Supervised Learning

Regression Models Classification Models

Ex.: Fraud problem

Deep Learning

Ensembled models



Ex.: Clustering

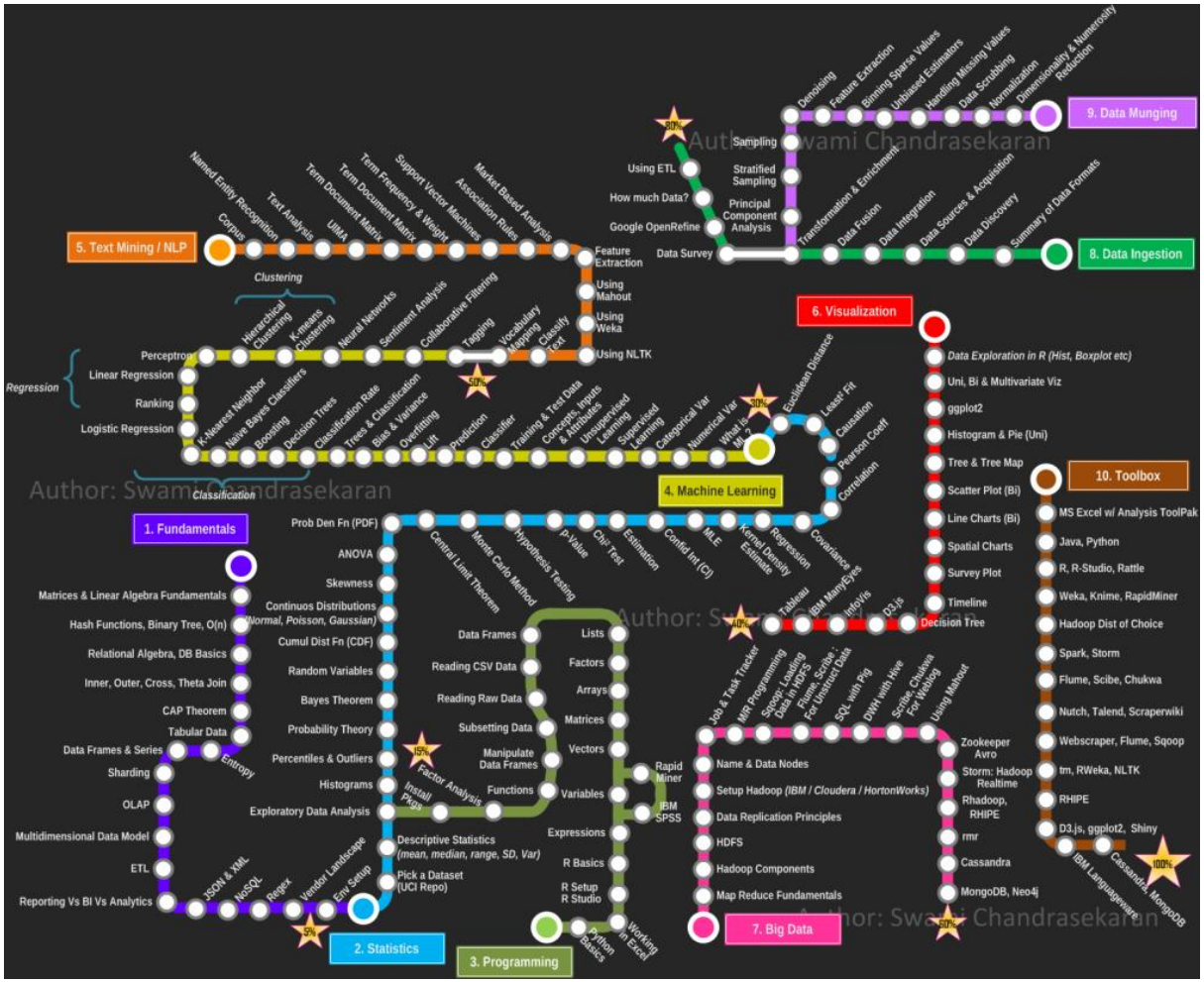
Unsupervised Learning

James et al (2013), "An Introduction to Statistical Learning - with Applications in R", link

Varian (2014), "Machine Learning and Econometrics", link1,
"Big Data: New Tricks for Econometrics", link2

Athey and Imbens (2015), "NBER Lectures on Machine Learning", link1, link2

Big Data e Data Science an overview



link

Selected topics of Data Science and Stata tools

- ✓ Statistics
- Machine Learning (partial tools)
- ✓ Data Visualization
- Big Data (no tools)

Prospects

What are Stata plans for Big Data and Data Science?

We, Debora Bilard and Timberlake, are planning to add Machine Learning algorithms to Stata and show applications via Timberlake website.

We would like to collaborate with and receive support from Stata, in what concerns libraries and other technical issues.

Thank you!