Statistical Modeling for Big-p, Big-n, Complex Data

Pradeep Ravikumar, UT Austin

University of Washington, Seattle 2013

Machine learning: Inferring conclusions from data

Statistical Machine Learning (SML): Doing so using statistical models that characterize the probabilistic relationship between each variable

Modern data

-across modern applications {images, signals, networks}

-m2m variables, observations

example:

fMRI

gene expression profiles

social networks

Characteristics of SML:

Big-p: a large number of variables p

Big-n: a large number of observations n

Complex: the data-type lies in some large unwieldy discrete space; permutations, strings, graphs

“off the shelf” SML techniques are frequently proving insufficient

“scruffy” over “neat”

How can SML help?:

GraphLab at UW, AMPLab at Cal

Optimization based approaches

-First-order methods very popular (iterave, with constant/linear time per step)

-but recently second-order, greedy, divide and conquer based approaches,…

**Statistical modeling that is tuned to big-p, big-n and complex data**

Big-p High dimensional

-sparse, group-sparse, low-rank, graphical models

Big-n Data

-non parapmetric data (n-dimetional space)

-improvements in cloud storage and capturing has greatly increased big data

Complex Data types

-trees, m2m

-permutations/rankings