4/12/2021

Model Planning

- What are the strengths and weaknesses of the tools that you are using?
- What kinds of variables or models are they designed to handle well?
- What packages are needed?

- Exploring the data to learn about the relationships between the variables
- Explore individual variables
- Determine which types of models and which variables will be most useful
- Separate the data into test and training sets (maybe a third – does the model require only one test or multiple test sets)

This phase of data exploration goes beyond the original data exploration phase – now we prepare the data for a specific type of analysis, check assumptions, etc.

Sometimes being removed from the domain of inquiry is beneficial, to be more objective – don't be too invested in a particular outcome. Avoid gut feelings, and pre-defined hunches. If these are offered, they must be tested against correlations with variables.

Types of models, options that are available:

- Map/Reduce
- Natural Language Processing (NLP)
- Clustering
- Classification
- Regression
- Graph Theory

Don't limit yourself to just one model – choose several to analyze and select the one that produces the best predictive value

Take care that our test/train split does not contaminate each other

- Normalize after doing the test/train split, not before
- Do dimensionality reduction after the test/train split, not before

Map/Reduce

It is a big data approach (Hadoop, Hive, Spark), takes data and maps it onto multiple processors, and then recombines the results into a single output

Spark works similarly to Hadoop (Python works with Spark using pyspark), Spark processes data in memory while Hadoop stores to disk – Spark is faster

Hadoop systems may be less expensive and are suitable for analyses that can run overnight. Spark is better for more real-time data processing.

Hadoop is better for linear processing. Spark is better for iterative processing, graphs, and joining data sets.

Natural Language Processing
Analysis of language in text form or in speech form

NLP is challenging because it's messy

Grew out of linguistics – semantics and syntax employed mathematical frameworks, formal methods, but many interesting problems remained unsolved

Modern NLP grew out of computational linguistics, developing a more statistical approach

NLTK is the Python package that does NLP

More recent analyses also use deep learning (neural networks)

Clustering Algorithms
Often used as a type of unsupervised learning – but can be modified to do as supervised or semi-supervised

Density-based methods – DBSCAN, OPTICS, etc.
Hierarchical Methods –
    Agglomerative (bottom-up)
    Divisive (top-down)
CURE, BIRCH, etc.

Partitioning methods – k-means, CLARANS, LDA, etc.
Grid-based methods – STING, CLIQUE, etc.

k-means is the most common of these methods