Machine Learning

PyfA L8

Data science is mostly

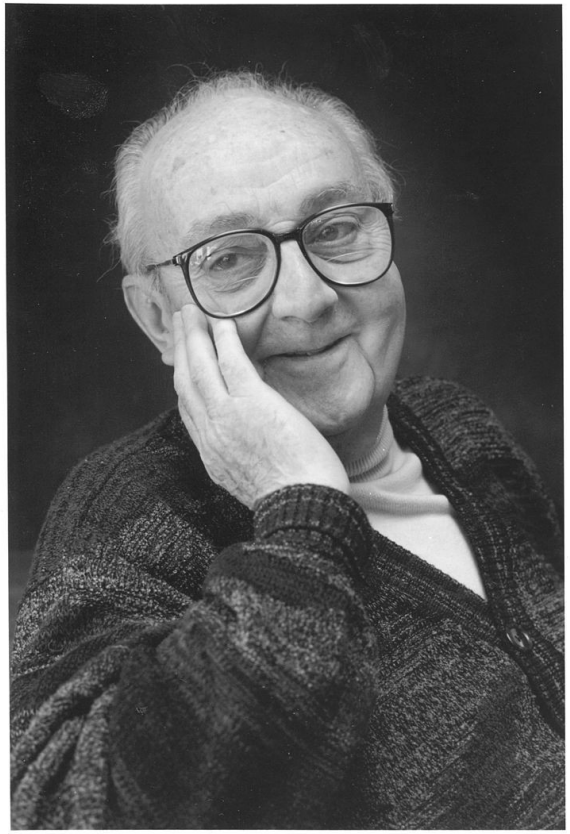
(1) turning business problems into data problems and

(2) collecting data and understanding data and

(3) cleaning data and formatting data

(4) then comes the machine learning part..

Machine Learning involves building mathematical models to help understand data. "Learning" enters the fray when we give these models *tunable parameters* that can be adapted to observed data. Once these models have been fit to previously seen data, they can be used to predict and understand aspects of newly observed data.



"All models are wrong; some models are useful."

"Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."

George Box (1919-2013) was a British statistician, who worked in the areas of quality control, time-series analysis, design of experiments, and Bayesian inference (Box-Cox transformation was named after him).

Categories of Machine Learning

Supervised learning

Modeling the relationship between measured features of data and some label associated with the data; once this model is determined, it can be used to apply labels to new, unknown data.

- *Classification* and *regression* tasks (labels are discrete vs continuous quantities, respectively)

Unsupervised learning

Modeling the features of a dataset without reference to any label.

- *Clustering* and *dimensionality reduction algorithms*

Semi-supervised learning

Often useful when only incomplete labels are available.

Supervised learning:

- Classification: models predict labels as two or more discrete categories
- Regression models predict continuous labels

Unsupervised learning:

- Clustering models detect and identify distinct groups in the data
- Dimensionality Reduction: models detect and identify lower-dimensional structure in higher-dimensional data

What does a ML project look like?

1. Study the data	Apply any transformations needed for the selected model. Split data into train and test
2. Select model	Create an instance of the selected model
3. Fit the model to data	Use the training dataset to train the model, test the model performance (improve if low)
4. Make predictions	Predict label for unknown data

Model Performance

- **Accuracy:** fraction of correct predictions
- **Confusion matrix:** binary judgement. True positive, false positive (type 1 error), false negative (type 2 error), true negative.
- **Precision:** measures how accurate the positive predictions were.
- **Recall:** measures what fraction of the positives model identified.
- **F1 score:** harmonic mean of precision and recall

Choice of a model is a trade-off between precision and recall!
George Box: "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful."
How many false positives or false negatives are acceptable?