Foundations of Data Science
Lecture 4, Module 3
Fall 2022

Rumi Chunara, PhD

Goals

- When is machine learning useful?
- Illustrate the need for feature generation/selection
- Describe requirements for causal inference

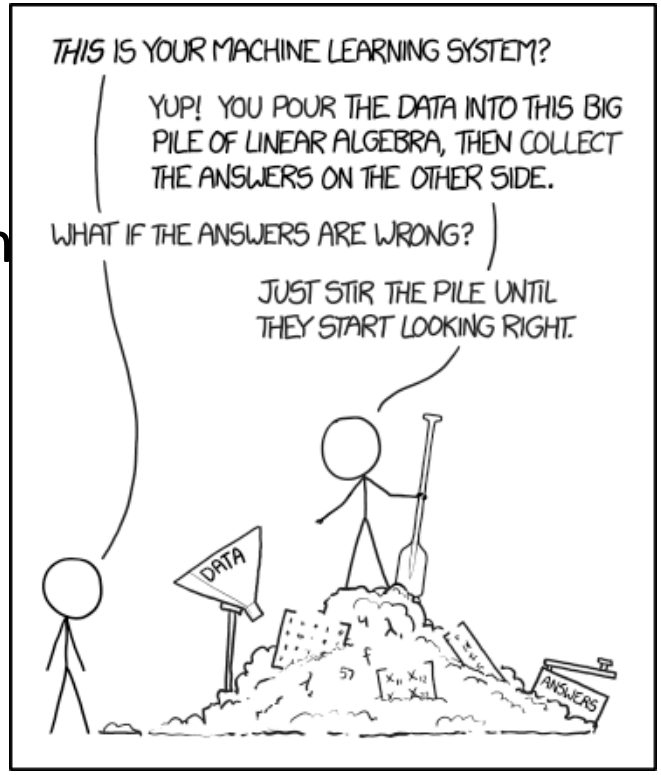
Machine Learning – when is it useful?

- modelling complex non-linear relationships (non-parametric)
  - between outcome variable and predictors (explanatory variables)
  - to produce predicted risk scores
  - making prediction with missing data

- handling a large number of predictors

Machine Learning <> Statistics

- Physics <> mathematics:
  - Physics is built upon mathematics, it is the application of mathematics to understand physical phenomena present in reality
- ML is built upon statistics, e.g. statistical learning theory, empirical risk minimization



Source: XKCD

Selecting an approach

- Regression vs classification
- Performance on new data (training and test data)
- Interpretability

How to pick?

- Which you use depends largely on what your purpose is.
  - If you want to create an algorithm that can predict housing prices to a high accuracy, or use data to determine whether someone is likely to contract certain types of diseases, machine learning is likely the better approach.
  - If you are trying to prove a relationship between variables or make inferences from data, a statistical model is likely the better approach.

Ease of Use vs. Garbage in-Garbage out

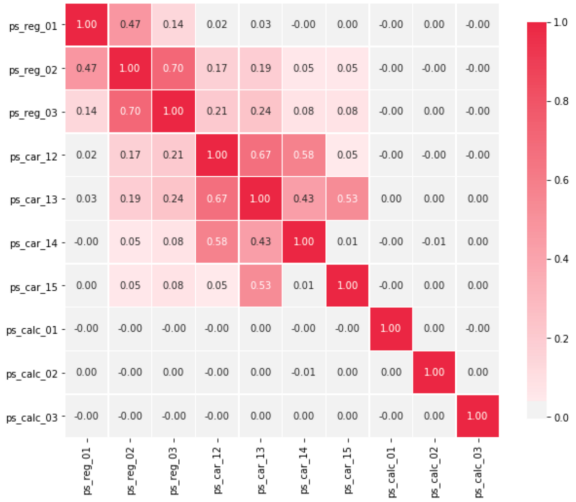
- The abstraction offered by machine learning libraries (e.g. scikit-learn, tensorflow) makes them pretty easy to use them as a non-expert
- An understanding of the underlying statistical ideas is helpful in order to prevent models from overfitting and giving spurious inferences
- An understanding or collaboration with a topical expert is also critical in order to have an idea of the *data generating process (DGP)* and system of interest

Causality

- An understanding of the DGP is critical for creating causal models
- Causal inference from observational data is an active field of research
- While the gold standard are RCTs, challenges of sample size and inclusion, ethics, external generalizability pervade

Feature selection

- **Multicollinearity** happens when one predictor variable in a multiple regression model can be linearly predicted from the others with a high degree of accuracy.
  - This can lead to skewed or misleading results.
  - Decision trees and boosted trees algorithms are immune to multicollinearity by nature. When they decide to split, the tree will choose only one of the perfectly correlated features.
  - Other algorithms like Logistic Regression or Linear Regression are not immune to that problem
  - Feature selection and generation are important in steps before training a model



Evaluation

- Assessment of the machine learning algorithm uses a test set to validate its accuracy.

ML/Stats Case Study 1

- If I am trying to prove that a sensor is able to respond to a certain kind of stimuli (such as a concentration of a gas), then I would use a statistical model to determine whether the signal response is statistically significant.
- I would try to understand this relationship and test for its repeatability so that I can accurately characterize the sensor response and make inferences based on this data.
- Some things I might test are whether the response is, in fact, linear, whether the response can be attributed to the gas concentration and not random noise in the sensor, etc.

Source: https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3

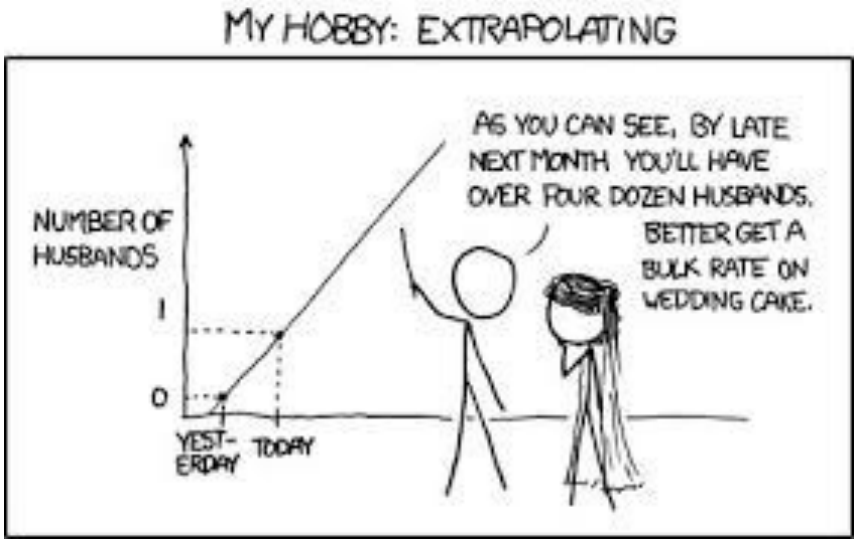
ML/Stats Case Study 2

- I can also get an array of 20 different sensors, and I can use this to try and predict the response of my newly characterized sensor
- A model with 20 different variables predicting the outcome of my sensor is clearly all about prediction, and I do not expect it to be particularly interpretable.
- This model would likely be something a bit more esoteric like a neural network due to non-linearities arising from chemical kinetics and the relationship between physical variables and gas concentrations. I would like the model to make sense, but as long as I can make accurate predictions I would be pretty happy.

Source: https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3

ML/Stats Case Study 3

- If I am trying to prove the relationship between variables to a degree of statistical significance so that I can publish it in a scientific paper, I would use a statistical model and not machine learning.
- This is because I care more about the relationship between the variables as opposed to making a prediction.
- Making predictions may still be important, but learning the relationship between variables is often a goal of scientific endeavor, and just using a model to predict without understanding the relationship can result in spurious inferences



Source: XKCD

Source: https://towardsdatascience.com/the-actual-difference-between-statistics-and-machine-learning-64b49f07ea3