Machine Learning: Introduction

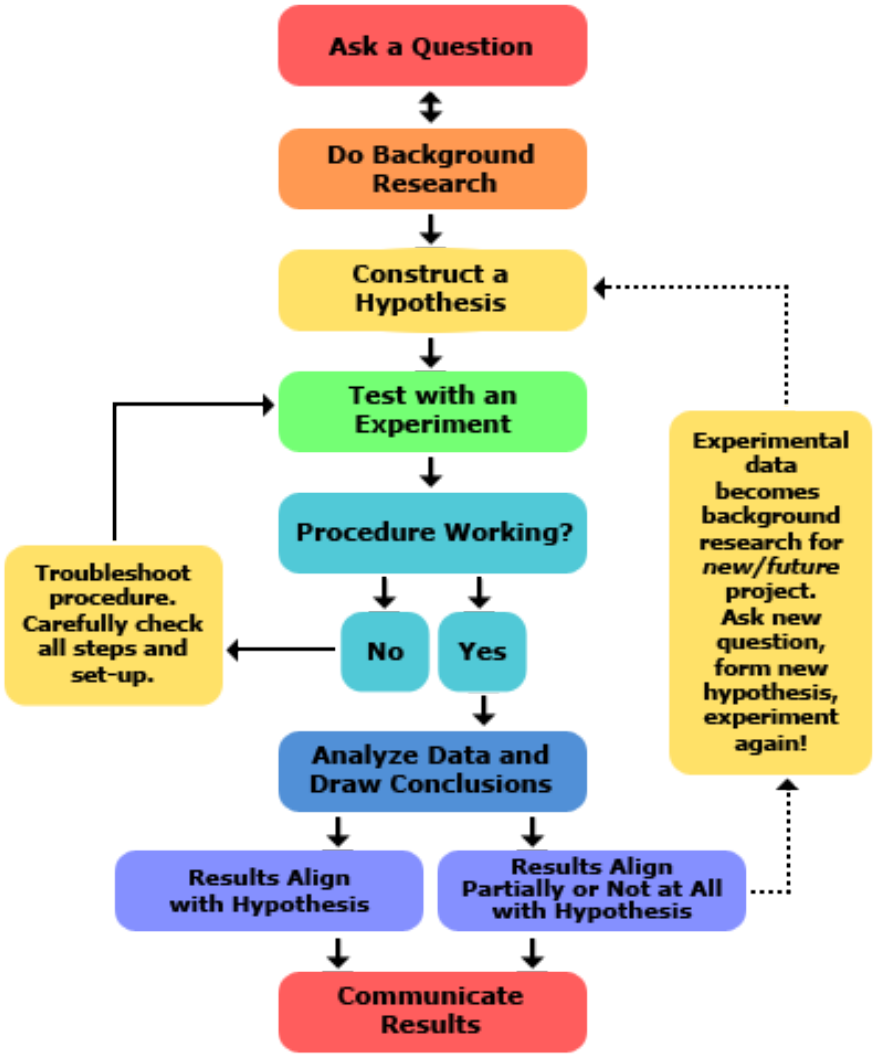
STOR 565

Andrew Nobel

August, 2021

Background: The Scientific Method

The Scientific Method (from science buddies.org)



Paradigm Shift

Traditional Scientific Method: Hypothesis Driven

- Formulate a hypothesis
- Collect data to confirm/refute hypothesis

Modern Scientific Method: Data Driven

- Acquire data from high-throughput measurement technologies
- Mine the data for possible hypotheses
- Often: use the data again to test selected hypotheses

Scientific Discovery: Needles and Haystacks

General Principle: If you have enough data, and you ask enough questions, you are bound to find something interesting, **just by chance.**

Bob: I found a needle in a haystack!

Amy: That's surprising! How many haystacks did you look in?

Bob: A thousand.

Amy: Oh, maybe that's not so surprising.

Overview of Machine Learning

Machine Learning: What is it?

“ML is the study of computer algorithms that improve automatically through experience. It is seen as a part of artificial intelligence.” [Wikipedia]

“Machine learning involves computers discovering how they can perform tasks without being explicitly programmed to do so. It involves computers learning from data provided so that they carry out certain tasks.” [Wikipedia]

“Machine learning is a method of data analysis that automates analytical model building. It is ... based on the idea that systems can learn from data, identify patterns and make decisions with minimal human intervention.” [SAS]

Machine Learning

High-profile applications

- Spam filtering, threat/fraud detection
- Machine translation, facial recognition
- Recommender systems, targeted marketing
- Personalized medicine, automated diagnoses

Under the hood: Statistics, optimization, computer science, mathematics

- Model development and implementation
- Model fitting and assessment
- Data acquisition and preprocessing

Machine Learning

Study and development of general computational methods and models for extracting actionable information from data or experience. Several flavors.

Unsupervised: Finding structure in data

- Dimension reduction: principal component analysis (PCA) and SVD
- Identifying subgroups: clustering

Supervised: Building predictive models

- Classification, pattern recognition
- Regression, curve fitting

Reinforcement: Learning from experience in a dynamic environment

Machine Learning, cont.

Machine Learning is NOT

- Magic or computational alchemy (fairy dust)
- A grab-bag of data analysis methods

A Few Statistical Caveats

- Always visualize your data. Garbage in, garbage out
- Always try simple methods before fancy ones
- Don’t forget about uncertainty and noise
- Double dipping, multiple testing, correlation vs. causation

Unsupervised Learning

Given: Data consisting of points $x_1, \ldots, x_n$ in $\mathbb{R}^d$

Dimension reduction (PCA): Find low dimensional subspace V of $\mathbb{R}^d$ s.t. the projection of $x_1, \ldots, x_n$ onto V captures most of the variation in the data

Clustering: Divide $x_1, \ldots, x_n$ into small number of disjoint groups (clusters) such that points in the same group are close together, and points in different groups are far apart

Supervised Learning

Given: Data $D_n = (x_1, y_1), \ldots, (x_n, y_n) \in \mathcal{X} \times \mathcal{Y}$

- $\mathcal{X}$ called *feature space*; if $\mathcal{X} = \mathbb{R}^d$ components called features
- x_i called *input* or *predictor* for ith observation
- $\mathcal{Y}$ called *response space*
- y_i called *output* or *response* for ith observation

Issue: Responses for other possible inputs are desired, but are difficult to obtain, or simply unknown.

Task: Use data D_n to find a rule (function) $f : \mathcal{X} \to \mathcal{Y}$ that will predict the response of a new input $x \in \mathcal{X}$

Supervised Learning: Classification and Regression

Classification: Response $\mathcal{Y} = \{-1, +1\}$. Use data D_n to predict label y of new input x. Example: email spam detection

- $x_i =$ vector of features extracted from email message
- $y_i = +1$ if email i is spam, $y_i = -1$ otherwise

Task: predict whether new email with feature vector x is spam or not

Regression: Response $\mathcal{Y} = \mathbb{R}$. Use data D_n to predict output value y of a new input x. Example: predicting individual income

- $x_i =$ vector of features regarding education, address, car ownership
- $y_i =$ income of individual

Task: predict income y of new individual with feature vector x

This Course

Emphasis on rigor and mathematical foundations

- Probability and statistics
- Order, minima and maxima
- Matrix algebra
- Convex sets and functions
- Calculus

Features of the course

- Not “drive-by”. We will cover fewer methods in more detail
- Homeworks and exams will be theoretically focused
- Homeworks and exams will not reduce to a few recipes or rubrics