Machine Learning: Introduction and Overview

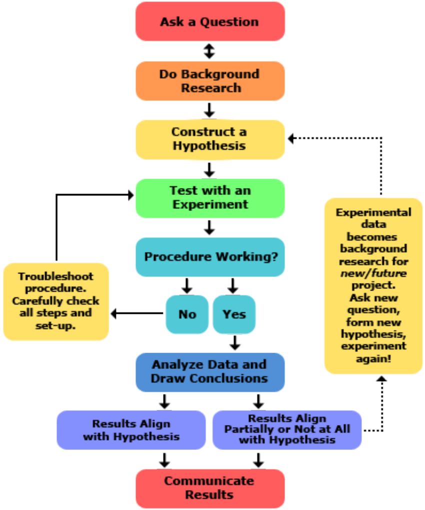
STOR 565

Andrew Nobel

January, 2020

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
- 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

High-profile applications

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

Key steps

- Data acquisition and preprocessing [Stat, CS]
- Model development and implementation [Stat, Math, CS]
- Model fitting and assessment [Stat, Optimization, CS]

Machine Learning

Study and development of general computational methods and models for extracting information from data. Two flavors.

Unsupervised: Finding structure in data

- Dimension reduction, principal component analysis (PCA)
- Finding well-defined subgroups, clustering

Supervised: Building predictive models

- Classification, pattern recognition
- Regression, curve fitting

Machine Learning, cont.

Machine Learning is NOT

- A grab-bag of methods for data analysis
- Magic or computational alchemy

Statistical Caveats

- Use elementary methods before more sophisticated ones
- Don’t forget about uncertainty and noise
- Be aware of multiple testing and correlation vs. causation

Unsupervised Learning

Given: Data $x_1, \ldots, x_n$ taking values in *feature space* $\mathcal{X}$, usually $\mathbb{R}^d$

Clustering

Partition $x_1, \ldots, x_n$ into a small number of disjoint groups (clusters) so that points in the same group are close together, and points in different groups are far apart.

Dimension reduction (PCA)

Find a low dimensional subspace V of $\mathbb{R}^d$ so that the projection of $x_1, \ldots, x_n$ onto V captures most of the variation in the data.

***Mixture modeling**

Fit a mixture of multivariate normal densities to $x_1, \ldots, x_n$

Supervised Learning

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

- Component x_i called *input, predictor, or feature*
- Set $\mathcal{X}$ called *feature space*, usually $\mathbb{R}^d$
- Component y_i called *output or response*
- Set $\mathcal{Y}$ called *response space*

Task: Use data $(x_1, y_1), \ldots, (x_n, y_n)$ to find a rule (function) $f : \mathcal{X} \to \mathcal{Y}$ that will predict the output of a new input $x \in \mathcal{X}$ when the output y is unknown or difficult to obtain.

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

Preliminary Inequalities

The Usual Order Relation

Definition: For $a, b \in \mathbb{R}$ write $a \geq b$ if $a - b \geq 0$.

Basic Properties

(1) If $a \leq b$ then $-b \leq -a$

(2) If $a \leq b$ and $c \leq d$ then $a + c \leq b + d$

(3) If $0 \leq a \leq b$ and $0 \leq c \leq d$ then $ac \leq bd$

Order Relations for Maxima and Minima

Fact: Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be real numbers.

(1) $\min\{a_i\} \leq a_j \leq \max\{a_i\}$ for $1 \leq j \leq n$

(2) $-\max\{a_i\} = \min\{-a_i\}$

(3) $-\min\{a_i\} = \max\{-a_i\}$

(4) $\max\{a_i + b_i\} \leq \max\{a_i\} + \max\{b_i\}$

(5) $\min\{a_i + b_i\} \geq \min\{a_i\} + \min\{b_i\}$

(6) $\max\{a_i\} - \max\{b_i\} \leq \max\{a_i - b_i\}$

Fact: For real numbers a, b we have $2ab \leq a^2 + b^2$.

Order Relations for Maxima and Minima of Functions

Fact: Let $f, g : \mathcal{X} \to \mathbb{R}$ be functions.

(1) $\min_{x\in\mathcal{X}} f(x) \leq f(x_0) \leq \max_{x\in\mathcal{X}} f(x)$ for every $x_0 \in \mathcal{X}$

(2) $-\max_{x\in\mathcal{X}} f(x) = \min_{x\in\mathcal{X}}(-f(x))$

(3) $\max_{x\in\mathcal{X}}\{f(x) + g(x)\} \leq \max_{x\in\mathcal{X}} f(x) + \sup_{x\in\mathcal{X}} g(x)$

(4) If $\mathcal{X}_0 \subseteq \mathcal{X}$ then $\max_{x\in\mathcal{X}_0} f(x) \leq \max_{x\in\mathcal{X}} f(x)$

Euclidean Norm and Inner Product

Given: Vector $x = (x_1, \ldots, x_d)^t \in \mathbb{R}^d$

- Inner product $\langle x, y \rangle \; = \; x^t y \; = \; \sum_{i=1}^d x_i \, y_i$
- Norm $||x|| = (x_1^2 + \cdots + x_d^2)^{1/2} \; = \; (x^t x)^{1/2}$

Basic Properties

- $||x|| \geq 0$ with equality if and only if $x = 0$
- For $a \in \mathbb{R}$, $||a \, x|| = |a| \, ||x||$
- $||x + y|| \; \leq \; ||x|| + ||y||$, the triangle inequality
- $| \, ||x|| - ||y|| \, | \; \leq \; ||x - y||$
- $|x^t y| \; \leq \; ||x|| \, ||y||$, the Cauchy-Schwartz inequality