

Machine Learning: An Introduction

A few quotes

“A breakthrough in machine learning would be worth ten Microsofts” (Bill Gates, Chairman, Microsoft)

“Machine learning is the next Internet”
(Tony Tether, Director, DARPA)

“Machine learning is the hot new thing”
(John Hennessy, President, Stanford)

“Web rankings today are mostly a matter of machine learning” (Prabhakar Raghavan, Dir. Research, Yahoo)

“Machine learning is going to result in a real revolution” (Greg Papadopoulos, CTO, Sun)

“Machine learning is today’s discontinuity”
(Jerry Yang, CEO, Yahoo)

“AI is a bigger deal for humanity than fire or electricity” (Sundar Pichai, CEO, Google)

Examples | 1

Database mining

- Machine learning has recently become so big party because of the huge amount of data being generated
- Large datasets from growth of automation web
- Sources of data include
 - Web data (click-stream or click through data)
 - Mine to understand users better
 - Huge segment of silicon valley
 - Medical records
 - Electronic records -> turn records in knowledge
 - Biological data
 - Gene sequences, ML algorithms give a better understanding of human genome
 - Engineering info
 - Data from sensors, log reports, photos etc

Examples | 2

Applications that we cannot program by hand

- Autonomous vehicles
- Handwriting recognition
 - This is very inexpensive because when you write an envelope, algorithms can automatically route envelopes through the post
- Natural language processing (NLP)
 - AI pertaining to language
- Computer vision
 - AI pertaining vision

Self customizing programs

- Netflix
- Amazon
- iTunes genius
- Take users info
 - Learn based on your behavior

Understand human learning and the brain

- If we can build systems that mimic (or try to mimic) how the brain works, this may push our own understanding of the associated neurobiology

What is Machine Learning?

Arthur Samuel (1959)

- ***Machine learning***: "Field of study that gives computers the ability to learn without being explicitly programmed"
- Samuels wrote a checkers playing program
 - Had the program play 10000 games against itself
 - Work out which board positions were good and bad depending on wins/losses

Tom Michel (1999)

- ***Well posed learning problem***: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."

Machine Learning types

Supervised learning

- Teach the computer how to do something, then let it use its new found knowledge to do it
- Learning with teacher
- Inferring a function from labelled training data
- Training data includes desired outputs (labels)

Unsupervised learning

- Teach the computer how to do something, then let it use its new found knowledge to do it
- Learning without teacher
- Finding hidden structure in unlabelled data
- Training data does not include desired outputs (labels)

Semi-supervised learning

- Training data includes a few desired outputs (labels)

Reinforcement learning

- Rewards from sequence of actions

Some examples

house sale price prediction (supervised learning)

image classification (supervised learning)

image clustering (unsupervised learning)

document classification (supervised learning)

document clustering (unsupervised learning)

time-dependent data (speech) analysis

sentiment analysis (supervised learning)

user grouping for collaborative filtering (unsupervised learning)

Some basic problems

Regression: Predict (continuous) value (supervised learning)

- E.g. house value prediction

Classification: Predict target class (binary- or multi-class) (supervised learning)

- identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known
- E.g. filter spam/non-spam, classify tumor as benign, type-1, type-2, type-3

Clustering: Group data/objects (unsupervised learning) in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters) (i.e. high intra-class similarity, low inter-class similarity)

- E.g. assign house addresses to closest pizza delivery store

Some well-known ML algorithms

Regression

- Least Squares Regression
- Linear regression
- Logistic regression

Classification

- Naïve Bayes
- Neural Networks
- Support Vector Machines (SVM)
- Kernel estimation (k-Nearest Neighbors)
- Random forests

Clustering

- K-Means
- Hierarchical clustering
- Spectral clustering
- Density-based algorithms (DBSCAN, OPTICS)