Introduction to Machine Learning

Lecture 1

Manuel Balsera

IEOR, Columbia University

January 18, 2018

Outline

1. Goal Of this Course
2. Machine Learning Road Map
3. Statistical Modeling vs Machine Learning
4. Introduction to Data
5. Bibliography

Machine Learning

From Wikipedia:

Machine Learning

is a field of computer science that gives computers the ability to learn without being explicitly programmed.

By **learn** we mean that the computer behavior in the future will depend on what examples it was exposed to during **training**.
From this point of view **linear regression** is machine learning.

Two Recent Machine Learning Papers

1. Computer-Extracted Texture Features to Distinguish Cerebral Radionecrosis from Recurrent Brain Tumors on Multiparametric MRI: A Feasibility Study [1] (2016).
2. Dermatologist-level classification of skin cancer with deep neural networks [2] (2017).

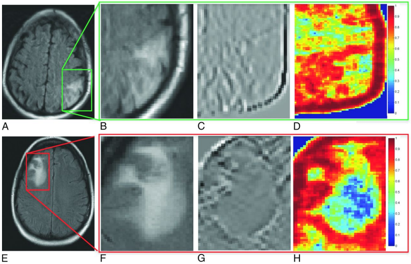


Figure 1: Illustration of Brain image texture features from Tiwari et al

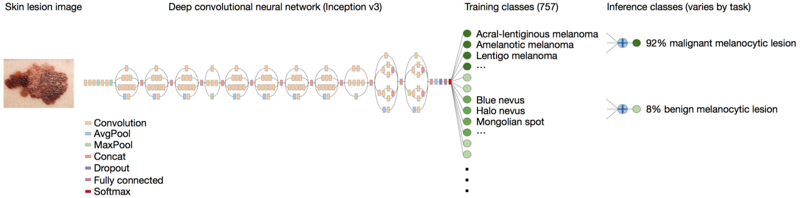
Two Recent Machine Learning Papers II

Both papers are superficially very similar

1. They are about cancer: medical application
2. they are examples of **image recognition**.
3. They both solve a **classification** problem (cancer or no cancer).

The ML methods used are quite different.

1. Tawari's et al use a Support Vector Machine trained over a set of **Texture Features**.
2. Steva's et al use a **deeply connected neural Network** trained on the raw pixels of the images.



Goals of E4525

The goals of IEOR 4525 are:

1. Make you understand the **methodologies** used in ML so that you can **apply** them to practical problems
2. Teach you how to **implement** and **adapt** methodologies to the problems you will face in practice.
3. Understand the **trade-offs** of the different methods and when to use them.

By the end of this course you should understand **why** the authors of [1] used a Support Vector Machine, and the authors of [2] used a Deep Neural Network instead (*it's the amount of data available*).

ML Requirements

For Machine Learning to occur we need

1. Data
2. Mathematical Modeling: Function Approximation, Optimization, Statistics and Probability
3. Computer Software: Efficient implementation of the mathematical models discussed above.
4. Hardware and Information Technology Infrastructure: Fast Computers, large memory and disk space, GPU's, Cloud computing, etc.

we will focus on the first three, and mention very briefly the last one towards the end of the course.

Types of Machine Learning

1. Supervised Machine Learning
    1. Regression
    2. Classification
2. Unsupervised Machine Learning
    1. Clustering
    2. Dimensionality Reduction
3. Reinforcement Learning

We will see a few examples of unsupervised learning later on the course, but we will focus mostly on **supervised learning**. We won't be covering Reinforcement Learning at all.

Supervised learning

the machine learning task of inferring a function from labeled training data.

Supervised learning is traditionally further sub-divided into

Classification Labels take a finite set of non-numeric (categorical) values.

Regression Labels are ordered, numeric values.

Examples of Classification Problems:

- Given an image output `Cancer` or `No Cancer` (This is binary classification).
- Given a news article output one of: `Finance`, `Politics`, `Sports`,`Gardening` or `Science` (this is multi label classification)

Example of Regression Problems:

- Given S&P 500 change and 10 year Treasury yield change on one particular day, predict Apple's stock price change on the very same day.
- Given years of formal education, and zip code, predict a person's yearly income.

Unsupervised Machine Learning

Unsupervised machine learning

the machine learning task of inferring a function to describe hidden structure from "unlabeled" data

This is a much harder and ill-defined problem than supervised learning. Examples of unsupervised learning are

- Clustering: Assign input data to one of K different groups of clusters. This is like classification, but without having access to the original labels.
- Dimensionality Reduction: Find a function that summarizes most of the data variability with only a few variables.
- Latent Variable Analysis: Find hidden variables and their distributions as to explain the properties of the observed data as well as possible:

  Given a set of movie reviews and reviewers, assign movies to particular genre's and reviewers a particular set of preferences.

Reinforcement Learning

Reinforcement Learning

the area of machine learning concerned with how software agents ought to take actions in an environment so as to maximize some notion of cumulative reward.

- Given a state of the environment a **software agent** must take action **repetitively** following a **policy**.
- The agent actions affect the environment's state.
- Each state provides a reward to the agent.
- Goal of agent is to optimize its policy as to maximize the total reward over multiple action turns.

This is the realm of computerized strategy games (Alpha Go) and self driving cars.
We will not be covering reinforcement learning on this course.

Machine Learning vs Statistical Modeling

Traditional Statistics (python's `statmodels` library):

- Focus on the data **model** .
- **Hypothesis testing**, and goodness of **fit**.
- Simplicity and explanatory power are important.
- Emphasis on **parametric** models.

Machine learning (python's `sklearn` library):

- focus on **generalization**. Can we predict the behavior of new data samples?
- model is just a **tool** not and end in itself.
- Predictive power more important that the simplicity or ease of comprehension.
- It is an Engineering field, algorithmic efficiency matters.

But lots of tools and techniques are **shared**. We will use quite a bit of probabilistic and statistical thinking in this class.

Simpson's paradox

Simpson's paradox

A phenomenon in probability and statistics, in which a trend appears in several different groups of data but disappears or reverses when these groups are combined.

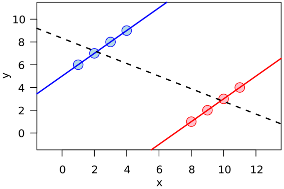


Figure 3: Illustration of Simpson's Paradox. Source: Wikipedia

Occurs when a missing (unobserved) variable has significant correlation with both x and y. **Correlation does not imply causation!**

Collecting Data

In a supervised learning problem (learn y given x) based on a N samples $\{y_i, x_i\}_{i=1}^N$

We face a few difference situations regarding how x_i was collected:

Controlled Experiment For each sample i we have chosen x_i and all its attributes: a scientist on a lab.

Randomized Experiment We can not control all the attributes of sample i, but we can choose some dimensions of x_i and randomly asign values to them: A/B Testing, medical trials.

Observational Data We have no control over x_i: Simpson's paradox is **always a concern**.

We can apply Machine Learning in all cases, but should expect a result ranking like:

$$\text{Controlled} > \text{Randomized} > \text{Observational}. \tag{1}$$

Example

Determine which one of two car is safer in an accident.

Controlled Experiment Crash cars in a lab, at set velocities against a fixed wall and measure damage.

- This is how car safety is measured in practice.
- Simple statistical analysis.

Randomized Experiment Select drivers at random. Give them one of the cars: follow up their accident outcomes.

- Variability of drivers may be larger than variability of car response to damage: **Need lots of cars**.
- Sophisticated analysis needed to distinguish effect for noise.
- Expensive experiment (pharmaceutical companies).

Observational Data We collect statistics on accidents for each kind of car

- Aggressive drivers may prefer one of the cars.
- It is hard to control for **bias** on all variables.
- Cheap experiment (just collect data).

Feature Selection Problem

Supervised Learning Goal

Find optimal function $f(x)$ to predict y **given** $x \in \mathcal{X}$

- Very powerful methods are available.
- But can not do magic is the problem is **miss-specified**.
- Remember **Sympson's paradox**.
- In practice hardest part of job is finding a good space $\mathcal{X}$ of features for prediction.
- There is no general theory on how to select $\mathcal{X}$. Need **domain knowledge**.

Problems suitable to ML

Current ML techniques can learn very good **function approximations**. Best suited from problems where there is a clear (but complicated) correspondence between x and y:

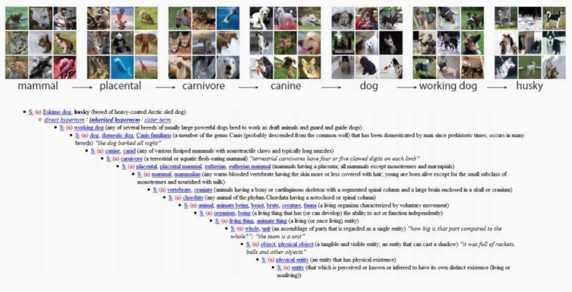
- Image Recognition: given an image, is there a cat on it?
- Speech processing: going from sound to text.
- Machine Translation: Going from English to Chinese.
- Natural Language Processing (NLP): What is the meaning of a text. *Implicit context can make this problem harder to solve*.

Domains where statistics is more suitable

- Predicting the stock market (what is $\mathcal{X}$?, everything could matter in principle)
- Drug discovery: body is very complicated, and not well understood.
- Predicting the outcome of US elections.

In all this problems, the key issue is choosing $\mathcal{X}$. We use statistics for **hypothesis testing**.

ImageNet



- S: (n) Eskimo dog, **husky** (breed of heavy-coated Arctic sled dog)
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) working dog (any of several breeds of usually large powerful dogs bred to work as draft animals and guard and guide dogs)
      - S: (n) dog, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) *"the dog barked all night"*
        - S: (n) canine, canid (any of various fissiped mammals with nonretractile claws and typically long muzzles)
          - S: (n) carnivore (a terrestrial or aquatic flesh-eating mammal) *"terrestrial carnivores have four or five clawed digits on each limb"*
            - S: (n) placental, placental mammal, eutherian, eutherian mammal (mammals having a placenta; all mammals except monotremes and marsupials)
              - S: (n) mammal, mammalian (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)
                - S: (n) vertebrate, craniate (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
                  - S: (n) chordate (any animal of the phylum Chordata having a notochord or spinal column)
                    - S: (n) animal, animate being, beast, brute, creature, fauna (a living organism characterized by voluntary movement)
                      - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
                        - S: (n) living thing, animate thing (a living (or once living) entity)
                          - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) *"how big is that part compared to the whole?"; "the team is a unit"*
                            - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
                              - S: (n) physical entity (an entity that has physical existence)
                                - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Figure 4: Source: David Yanovsky—Quartz

- Large data sets have pushed forward the field.
- Fei-Fei Li's team's Image Net [3] has been instrumental.
- **Big Data**: ≈ 1000 classes, tens of millions of images.
- See David Yanovsky Blog post for background history.

Deep Learning Revolution

- Performance on Imagenet Classification Problem has increased 10× in seven years.
- Many groups can **consistently** get under 5% error rate.
- Human Level performance is now **rutine**.
- AlexNet (2012): Convolutional Neural Network with 8 Layers.
- Microsoft RestNet (2016) has $\approx$ 150 layers (depends how you count)
- This is **Deep Learning**.

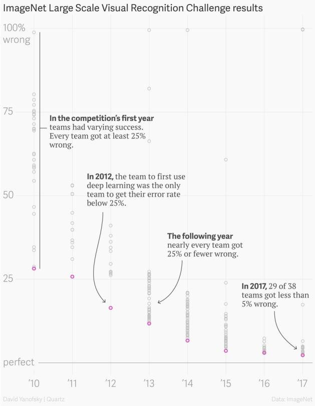


Figure 5: Source: David Yanovsky—Quartz

Why NLP is Hard: Winograd Schema

Contrast this two sentences

Winograd Schema Example

1. The **trophy** would not fit in the brown suitcase because **it** was too **big**.
2. The trophy would not fit in the **brown suitcase** because **it** was too **small**.

- Changing the last word from **big** to **small** changes the referent of pronoun **it** from **trophy** to **brown suitcase**.
- Finding the referent to **it** requires **understanding** how *trophies* fit into *brown suitcases*.
- need to **reason** about the world.
- **Windograd Schema Challenge**: Best accuracy in 2016 was 58% (50% guessing). Next Challenge **this February**.

Machine Learning vs General Artificial Intelligence

Simple Automated Image Caption is feasible

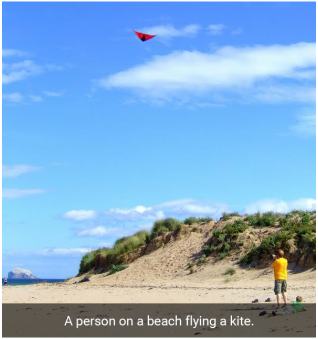


Figure 6: Source: Google Blog

But we still can not build an algorithm that gets this is **funny**.

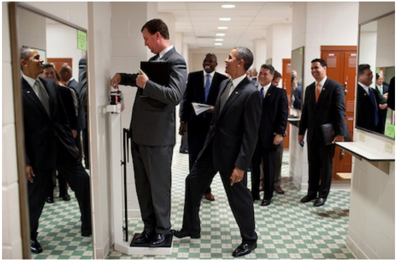


Figure 7: Source: Vanity Fair

Machine Methods We will Cover

If we have time we will cover

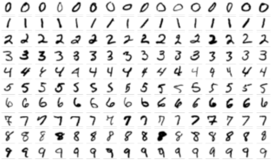
- Nearest Neighbors
- Naive Bayes Classifier
- LDA and QDA Classifier
- Logistic Regression
- Support Vector Machines
- Unsupervised Learning: Clustering and the EM Algorithm
- Neural Networks
    - Dense
    - Convolutional
    - Recurrent

Applications we will cover

Again, if we have time...

Text Classification `A break-in at the U.S. Justice Department's World Wide Web site last week highlighted the Internet's continued vulnerability to hackers...`

Image Recognition



Sound Classification Children's playing

Inputs to Machine Learning (Data)

ML require data examples that describe the information known about the objects of interest.

Feature Engineering

The art of representing information in a computer so that a ML algorithm can consume it.

- one of the **fundamental** problems of machine learning.
- can affect algorithm performance very significantly.
- requires significant **domain knowledge**.
- There are principles, and rules, but at the end its an **art**.

Example: to determine a person's income should we include in our ML model?

- Educational Level.
- Zip code of residence.
- Zodiac Sign.
- Precise measurements of the skull. How?

Scalar Data

The traditional classification for single (scalar) variables is

Nominal (Categorical) can take a finite set of values with no implicit ordering between them: car, bus, truck, or bike.

Ordinal can take a finite number of values which are ordered, but the distance between values is not meaningful: Very Bad, Bad, Indifferent, Good, Very Good.

Interval can takes ordered numerical values, the difference of values make sense, can take an average: A temperature in Fahrenheit or Celsius scale.

Ratio Numerical values make sense, zero and unity are well defined, make sense to take ratios: A temperature on the absolute Kelvin scale.

This classification, is useful, but not complete. For example an angle $0 < \theta < 2\pi$ is a special type of numerical valuable where 0 and 2π represent the same data point. Important in astronomy, mapping, and robotics!

Structured Data

Data can be more complicated structure that a single scalar

- A Homogeneous Vector of scalars all of the same type (categorical, ratio, etc).
- A mixture of categorical and numerical values
- A Gray-scale image represented as a $R \times S$ matrix of integer pixel intensities
- A Color Image represented as a $R \times C \times 3$ volume of RGB Color intensities
- A Text Document Represented as a list of unicode characters.

A **data pipeline** will transform the data from its raw representation into a form suitable for a specific ML algorithm.

Bibliography

P. Tiwari, P. Prasanna, L. Wolansky, M. Pinho, M. Cohen, A.P. Nayate, A. Gupta, G. Singh, K. Hattanpaa, A. Sloan, L. Rogers, and A. Madabhushi.
Computer-extracted texture features to distinguish cerebral radionecrosis from recurrent brain tumors on multiparametric mri: A feasibility study.
American Journal of Neuroradiology, 2016.
https://doi.org/10.3174/ajnr.A4931.

Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun.
Dermatologist-level classification of skin cancer with deep neural networks.
Nature, 542:115–118, 2017.
https://cs.stanford.edu/people/esteva/nature/#!

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei.
ImageNet: A Large-Scale Hierarchical Image Database.