Why & How: Introduction to Machine Learning

Ángel Torrado-Carvajal
atorradocarvajal@mgh.harvard.edu

Postdoctoral Research Fellow – Pain Neuroimaging Laboratory
Athinoula A. Martinos Center for Biomedical Imaging
Massachusetts General Hospital – Harvard University
Index

- Introduction
- Types of Tasks
- Types of Problems
- Approaches
- Cross-Validation
- Software Packages
Introduction

“If it looks like a duck, walks like a duck and sounds like a duck…

…then it’s a duck”
Introduction

- **What is “Learning”?**
  The modification of behavior through practice, training, or experience

- **What is “Machine Learning”?**
  To solve a problem automatically, based on a finite set of examples

- **Objective:**
  To design a system that is able to automatically adapt and optimize its performance in a given environment

- **Mathematically:**
  Find \( f: X \rightarrow Y \)
  Given a set of examples \( x^{(k)} \) called feature vector or just **features**
  To find a desired output \( o^{(k)} = f(x^{(k)}) \)
Introduction

Why & How: Introduction to Machine Learning

$x^{(1)} \rightarrow$ (blood pressure, BMI, ...
  ... glucose level)

$x^{(2)} \rightarrow$ (blood pressure, BMI, ...
  ... glucose level)

$x^{(N)} \rightarrow$ (blood pressure, BMI, ...
  ... glucose level)

$x^{(i)} \rightarrow o^{(i)}$ has diabetes?

$o^{(1)} \rightarrow$ Type 1

$o^{(2)} \rightarrow$ Type 2

$o^{(N)} \rightarrow$ Healthy
Types of Tasks

• Supervised Learning
  \[N\] training examples of the form \[\{(x^{(1)}, y^{(1)}), \ldots, (x^{(N)}, y^{(N)})\}\]
  \[x^{(i)}\] is the feature vector of the \(i\)-th example
  \[y^{(i)}\] is the label (i.e., class)

• We know:
  • The number of classes
  • The labels

• The goal is to “map” the desired output
Types of Tasks

• Unsupervised Learning
  \[ \{x^{(1)}, ..., x^{(N)}\} \]
  \(x^{(i)}\) is the feature vector of the i-th example

• We:
  • (Don’t) know the number of classes
  • Don’t know the labels

• The goal is to find some meaning based on the dataset
Types of Problems

- A **cluster** is a group of objects that:
  - have “something” in common
  - are more similar (in some sense or another) to objects in that group
  - are more different to objects in other groups

- **Clustering** is the process of grouping an heterogeneous set of objects in several subsets that:
  - Are more homogeneous than the original group
  - Have some “meaning”

- Helps us to find and better understand "Natural" groupings in a set of objects (data)
Types of Problems

- Clustering
  - How many groups are there?
Types of Problems

- **Density Estimation** is the construction of an estimate of an unobservable underlying probability density function based on observed data.

  - Why “unobservable underlying” pdf?
    - Because our data is a random sample of the population.
Types of Problems

- **Dimensionality Reduction** simplifies inputs by mapping them into a lower-dimensional space
  - Principal Componen Analysis

- **Feature Selection** is the process of selecting a subset of relevant features for use in model construction
  - Redundant
  - Irrelevant
Types of Problems

- **Classification** is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs
  - Diagnosis of the pathology

- Objects of the same class have some characteristics in common
  - These could be geometric, thematic, etc

- An algorithm implementing this method is known as a **classifier**
Types of Problems

- **Regression** is a set of statistical processes for estimating the relationships among variables.

- A function of the independent variables called the regression function is to be estimated:
  - Inputs are mapped into a function providing continuous outputs.

- An algorithm implementing this method is known as a *regressor*.
Types of Problems

- K means
  - Iterative method

- Assignment
  Assign each observation to the cluster whose mean has the least squared Euclidean distance, this is intuitively the "nearest" mean

- Update
  Calculate the new means to be the centroids of the observations in the new clusters
Types of Problems

- K means
Approaches

• Hierarchical Clustering
  • Organizes data in a hierarchical structure according to their proximity
  • The output is a hierarchy between clusters

• Agglomerative
  “Bottom up" approach
  Each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy

• Divisive
  “Top down" approach
  All observations start in one cluster, and splits are performed recursively as one moves down the hierarchy
Approaches

- Hierarchical Clustering
Approaches

- Expectation-Maximization
  - Iterative Method

- Start with an estimate of the parameters

- Expectation
  Compute the *likelihood* that each parameter produces the data point
  Calculate weights for each data point based on the *likelihood* of it being produced by a parameter

- Maximization
  Combine these weights together with the data to compute a better estimate for the parameters
Approaches

- Expectation Maximization
Approaches

- Decision Tree
  - Tree-like graph
  - Model of decisions and their possible consequences
Approaches

• Random Forest
  • Ensemble method
  • Multitude of decision trees
Approaches

- Neural Networks
  - Biological inspiration

  \[ f(x) \]

- Multi Layer Perceptron (MLP)
- Radial Basis Function Network (RBFN)
Approaches

- Support Vector Machines (SVMs)
Approaches

- Support Vector Machines (SVMs)
  - Linear SVM
  - Non-linear SVM (kernels)

\[ \omega^T x + b = 0 \]
Approaches

- Deep Learning
  - Deep Neural Networks

DEEP NEURAL NETWORK

input layer → hidden layer 1 → hidden layer 2 → hidden layer 3 → output layer

Approaches

- Deep Learning
  - Deep Neural Networks
- Convolutional Neural Networks
- Deep Belief Networks
- Recurrent Neural Networks
Cross-Validation

- The fitting process optimizes the model parameters to make the model fit the training data as well as possible.

- If we take an independent sample of validation data from the same population as the training data, the model does not fit it very well.

- This situation is called overfitting.

- We need to assess how our results generalize.

- Available data
  - Training set
  - Validation/Testing set.
Cross-Validation

- K-fold cross validation
  - Non-exhaustive cross-validation
    All ways of splitting the original sample are not computed

*Image source: Wikipedia*
Cross-Validation

- Leave-one-out cross validation
  - Exhaustive cross-validation
    
    All ways of splitting the original sample are computed

*Image source: Wikipedia*
## Software Packages

- R
- Weka
- Mathematica
- STATISTICA
- Amazon Machine Learning
- IBM Data Science Experience / IBM SPSS Modeler
- Google Prediction API
- Microsoft Azure Machine Learning
Software Packages

- Sci-kit Learn
  - Machine Learning in Python
  - Simple and efficient tools for data mining and data analysis
  - Accessible to everybody, and reusable in various contexts
  - Built on NumPy, SciPy, and matplotlib
  - Open source, commercially usable - BSD license
  - Ready to use from your Martinos Workstation!
  - Plenty of Tutorials and User Guide!
    [http://scikit-learn.org/]
Software Packages

- Statistics and Machine Learning Toolbox
  - Machine Learning in Matlab

- Ready to use from your Martinos Workstation!

- Plenty of Tutorials, User Guide and eBook!

- Why & How Sessions:
  - February 1, 5:30pm
    Demystifying Deep Learning: A practical approach in MATLAB
  - February 8, 5:30pm
    Parallel and Distributed Computing with MATLAB
Readings

• **Machine Learning**
Let’s discuss!

Thank You!