First Course on Supervised Learning for Practitioners
Raman Kannan
Chapter 1: Course Overview

The three basic questions one should ask to learn anything
what is Machine Learning (ML)?
Why ML?
how ML -- doing starts here

What is ML?

Machine Learning is a particular approach to making sense out of data. ML is vastly different from traditional programming where you construct sequence of statements which are coded to process data in a particular format. When we write software using traditional programming methods, the application you write, cannot and will never do anything you did not explicitly code for. ML is different in that regard. ML will learn from all the data it has seen and learn as it goes along and is able to infer when some new (never seen before data) is presented.

Act of forming general principles from particulars is known as generalization or inductive reasoning. In classification, ML algorithms observe given instances (particulars) and make inferences about scenarios never seen before.

ML applications perform a given task better with more data (or with experience)
because ML programs identify patterns and show signs of leveraging such patterns just as a mammalian brain does. Therefore, in summary, classification is about generalization aka inductive reasoning.

Why ML?
We cannot be writing traditional software given the deluge of data. We need a technique which can learn from data and process data without a programmer anticipating and coding everything in a piece of code. Such a goal would be futile. We need a new approach.

Why now?
ML is an ancient discipline since 1959 --Arthur Samuelson coined this terminology ML-- but it nearly faded away (1) in the 80s, because it required a lot of data (2) significant computational power, and (3) Only DoD and other major gov agencies and large coporations could finance such
projects. In contrast, the human brain learns in real-time, 24x7 for 15 to 20 years, with a capacity to store 4.7 billion books or 670 million web pages [1], before it is useful at all.

Now with the advent of internet, we have more data -- and a smartphone has more computing power -- computing capacity is cheap -- due to vastly improved communication (sharing,disseminating) capabilities, cloud, hw/sw improvements
-and therefore- (1) and (2) are not valid anymore. Even small firms with modest resources can now undertake M/L initiatives --
(3) is not valid anymore because we have IBM, Google, FB,MSFT,AMZN and other such companies that can undertake such projects. DoD is not the only game in town, Elon Musk, Jeff Bezos are also in town and they dont know idling.

M/L is popular and in demand now. It is so successful that there is a prevalence of practitioners labeling everything as M/L or some-sort of learning whether or not there is any learning.

In these essays we will avoid doing such things. We will focus on classical machine learning and exclude NLP, Deep Learning (DL) and Neural Nets. They are important and DL can do what we can do, sometimes better and faster. But we cannot learn anything as we will not know why it did what it did. They are indistinguishable from astrology or magic. These topics NLP and DL deserve a full length semester long study.

Now, then, learning can be achieved in many different ways and
there are three main branches of learning in ML
-- supervised learning (aka classification)
-- unsupervised learning (aka clustering)
-- reinforcement learning
Learning generally involved perceiving the essence beyond what is evident in obvious dimensions -- in that sense dimensionality reduction techniques can also be considered as a major field of ML.
Better claim is that M/L draws from other mature subjects including linear algebra, calculus, statistics, information retrieval and optimization.

We will focus entirely on classical supervised learning techniques. In classification, we are given a set of observations in the form of records. Each observation is composed of set of features and a class variable. ML algorithms learn hidden patterns in the given dataset, and learn to associate a class given feature vector. In ML, therefore, features are independent variables (IV) and the class variable is the dependent variable (DV). There is almost always more than one IV (multi-variate) and exactly one independent variable (DV).

In summary, "Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions." [2]

In supervised learning we seek to find a function or a method to determine the class (DV) given the features (IV).

Classification algorithms seek to find a function f, such that,
DV = f(IV)

Thus, we can abstract the supervised learning in terms of predictor variables (IV) and output variable (DV).
Predictor variables are also known as
  independent variables,
  regressors,
  features,
  covariates,
  input variables.
Predictors can be numerical (height,weight,salary) or categorical (gender, color)
Output variables are also known as
  class variable,
  dependent variable or
  class labels.

When the output variable is quantitative OLS -- we use regression techniques.

When the output variable is qualitative (ordinal,nominal) -- it is called classification.

The scope of this course are Classification techniques -aka- Supervised Learning.
As an academic in my research, I am interested in fundamental improvements to classification, applications of classification, in particular understanding text corpus -- NLP.

There are many classification methods:
-- probabilistic methods, considering entire dataset, properties of individuals are not a significant factor
-- instance based methods -- individual properties are all that matters -- distance based how similar are two objects
-- logic based decision making -- extract rules to classify
-- topological methods

Examples of probabilistic methods:
  Logistic Regression
  Fishers Discriminant Analysis (LDA,QDA, etc)
  Naive Bayes

Example of instance based classification:
  kNN -- nearest neighbor

Example of Logic Based classification:
  Decision Tree

topological methods
  Suppor Vector Machines -- (kernel methods)

These are individual learners. Individual learners are subject to the bias having seen a particular set of observations and variability in the inferences they make.
Therefore, one can seek to improve by combining them.
One fundamental method to improve anything is the
Iterative method -- do something, find out where it went wrong (what makes it inefficient), optimize (compensate) and repeat until no further improvements can be made or a target performance has been achieved.
Practitioners refer to this technique as boosting.
batch method -- learn from several subsets of data and average them -- aka bagging
ensemble method -- in bagging all features are considered -- what if we learn using different features and average them -- when we apply this to Decision tree -- we get random forest
stacking -- all three combination methods above, perturb a single kind of learner. Therefore, we could imagine combining learners of different kinds (probabilistic, logic-based, instance based and topological learners) and such a method is called stacking.

First Task, Mind your data.
Avoid GIGO. Quality data has the potential to result in quality analysis. Without quality nothing of value can be harnessed. Data Quality is paramount. Inappropriate use, results in useless inferences.
Therefore, we must seek to understand data at a deeper level
-- relating to data
  -- how many features
    -- are all features equally important

-- are they collinear or correlated with each other,
-- are they correlated with the outcome variable
-- how many classes
-- binary or multiple
-- imbalanced dataset some classes are rare and under-represented
-- Splitting data into training and testing
-- reproducibility and repeatability
-- scale effects -- some features have wider range, exhibit much different dispersion,
-- requires standardization either scaling or normalization
-- curse of dimensionality (merely mentioned here)
-- a dataset is a matrix, columns as features (length, width, color, blood type) and observations as rows.
-- we are mostly dealing with rectangular matrices in machine learning
-- dimensionality reduction using well known linear algebra techniques
Principal Component Analysis (aka in physics Eigen Value Decomposition),
Cholesky (positive definitive matrices) and
SVD (rectangular matrices)
-- how to interpret the model output in a transformed coordinate space back
in the real world in which the business operates.

Classifier Performance

-- relating to performance
-- metrics to assess performance of a supervised learner
-- classifiers
-- true positives, true negatives -- model is consistent with the reality
-- false positive -- models says it is positive but in reality that observation is negative
-- false negative -- model says it is negative but in reality that observation is positive
-- accuracy, sensitivity, specificity, recall, precision and F1
-- RoC curves and Area Under the Curve (AUC)
-- Null deviance and residual deviance
-- regression
-- R-squared, adjusted R-squared, RMSe
-- F-statistic, AIC,BIC
-- irreducible error, bayes factor
-- theoretical bound on the smallest error
-- source of errors
-- inability to estimate population parameters given a sample
-- bias and underfitting
-- sample is NOT representative of the population
-- simpler models are more prone to bias
-- we have observed 44 presidents all of whom happen to be males
-- that is a biased sample not good for making sound inference
-- until 41 all were caucasian ancestory until Obama
-- under-fitting occurs where the algorithm is unable to learn
-- if the model is unable to achieve 65% accuracy during training phase,
one may conclude the model is unable to learn
-- variance and overfitting
-- due to the variability in the measuring or observation methods
-- both variance and bias cannot be simultaneously reduced
-- Variance/Bias manifest as over-fitting/under-fitting respectively

-- that is, when a learner is unable to classify correctly even when it has seen the data

-- over-fitting occurs when a learner has been over trained or the model is overly complex and it has captured noise in the training data along with the signal

-- when the performance of a classifier degrades when presented with never seen before data -- it is a sign of inability to generalize -- result of overfitting

-- be suspect of AUC=1 or perfect scores or error free learning tools.

-- If your model is classifying with 100% accuracy it is not necessarily a good thing.

-- And on occasion, classifier performance remains in the low teens or below 30% and that is not -- necessarily a poor reflection of our process. Some processes do not conform to any process or pattern --for example volcano eruptions, and earth-quakes are hard to predict --.

-- there is no such thing as perfect -- please reject the pursuit of perfection. There AINT no such thing.

-- Variance reduction strategies

-- Cross Validation

-- Ensemble methods are a reliable way to reduce overfitting, i.e.. reduce variability as they aggregate several individual classifiers susceptible to variance. In general, aggregation results in smoothening.

[1] https://www.telegraph.co.uk/news/science/science-news/12114150/Human-brain-can-store-4.7-billion-books-ten-times-more-than-originally-thought.html

[2] https://emerj.com/ai-glossary-terms/what-is-machine-learning/