Machine Learning

Neeta Nain

Malaviya National Institute of Technology
(CST461, CST821 and PHD)

July 25, 2023

Overview

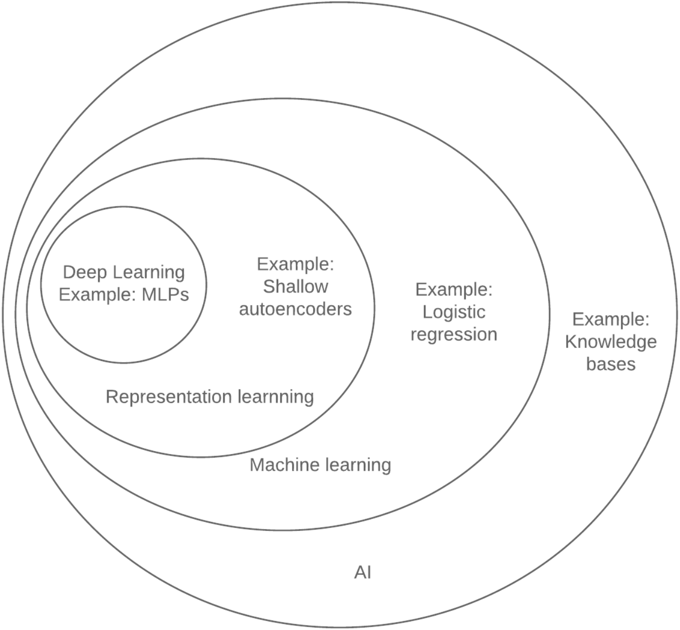
1. About the Course
2. Introduction to Machine Learning
3. The Categorization of Machine Learning

Books and References

1. A First Course in Machine Learning, Simon Rodgers and Mark Girolami, Second Edition CRC Press
2. Learning From Data, Yaser Abu-Mostafa, AML BooksML, PR, Data Mining
3. Machine Learning An algorithmic Perspective, Second Edition, Stephen Marsland, CRC Press
4. Fundamentals of Machine Learning, John D. Kelleher, The MIT Press
5. An Introduction to Machine Learning, Second ed, Miroslav Kubat, Springer
6. An Introduction to Statistical Learning, Gareth James, Robert Tribshirani, Springer

Course Contents

- Prerequisites: Linear Algebra, Calculus and Probability
- Mid Term: 30, CWS: 20, End Term: 50. Mandatory 75% attendance (Lecture and Labs). CWS - quizes will be taken within class. Labs will be evaluated in Lab durations. No Homeworks.
- Introduction to ML, ML Model
- Error based learning: Simple Linear Regression, Gradient descent, Multivariate linear regression, Logistic Regression and Classification
- Information based Learning: Decision trees, random forest, bagging and boosting
- Similarity based learning: Similarity measures, distance metric, evaluation metrics, Nearest Neighbour
- Probability based Learning: Bayesian learning, MLE, MAP
- Accuracy measures: accuracy, classification error, TPR, FPR, FNR, TNR, DET, ROC, AUC, precision, recall, F_1 measure, sensitivity, specificity etc
- Neural Networks and SVM. Feature Selection: PCA, LDA, MDS
- Clustering: Agglomerative, hierarchical and density based.


Deep Learning
Example: MLPs
Example:
Shallow
autoencoders
Example:
Logistic
regression
Example:
Knowledge
bases
Representation learnning
Machine learning
AI

ML, PR, Data Mining

- All three are related concepts - focus on extracting patterns from data
- Tom Mitchell - 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
- Study of methods and algorithms for associating data with objects and classes Pattern Recognition
- Extract patterns from unstructured data Data Mining

What is Machine Learning?

- Learning from data is used in situations where we dont have an analytic solution.
- We have data that we can use to construct an empirical solution. If it is hard to tell a machine exactly how to go about a certain problem, why not provide the instructions indirectly, conveying skills by way the computer will learn - from examples (which rely on the existence of algorithms to do the learning).
- Exploded in 1983, Machine Learning: The AI approach, R, Michalski, J. Carbonell and T. Mitchell - a thick volume of research papers which proposed the most diverse ways of addressing the great mystry.
- Important watershed was Tom Mitchell book on Machine learning (1997) - summarized the state-of-theart both for research students and scientists alike.
- Universities started offering machine learning as an undergraduate class. Not only how to learn but also what to learn and why.

History

- The term ML was coined by Arthur Samuel in 1959, IBM employee - checkers playing program; would learn to play against itself and improve its performance with experience (no strategy). Looking at the patterns in data, analyzes it and makes decision.
- AI - Alan Turing, 1950 to design programs that performs at coginitive level of human intelligence. ML is a subset of AI which use past data to build such stategies/programs, where performance improves with time. DL made a lot of progress in last 5/10 years
- CV and image recognition - ImageNet + CNN revolutionized CV since 2012 - 2015. Navigation
- Voice recognition - Apple assistant 2011, Google assistant 2016, etc
- Language translation - Google translate 2006- just by looging at two language documents it can translate.
- Reinforcement learning - Google Deepmind Atari game 2013, just feed the CNN model with a snap shot (raw pixels) and evaluate the future rewards. AlphaGo 2015 - the first to defeat a professional human world champion, and is arguably the strongest Go player in history.

Examples of Learning?

- Text: document classification, spam detection.
- Language: NLP tasks (e.g., morphological analysis, POS tagging, context-free parsing, dependency parsing).
- Speech: recognition, synthesis, verification.
- Image: annotation, face recognition, OCR, handwriting recognition.
- Games (e.g., chess, backgammon, go).
- Unassisted control of vehicles (robots, car).
- Medical diagnosis, financial forecasting, fraud detection, network intrusion, computer vision, recommender system ...they have successfully utilized learning from data

Some Broad Machine Learning Tasks

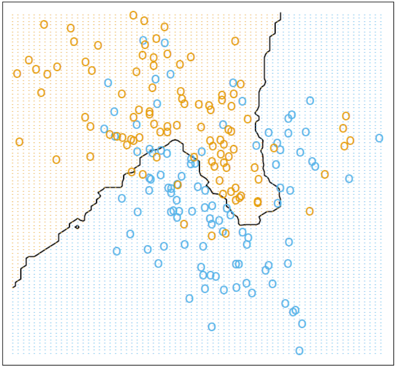
- Classification: assign a category to each item (e.g., document classification).
- Regression: predict a real value for each item (prediction of stock values, economic variables).
- Ranking: order items according to some criterion (relevant web pages returned by a search engine).
- Clustering: partition data into ‘homogenous’ regions (analysis of very large data sets).
- Dimensionality reduction: find lower-dimensional manifold preserving some properties of the data.

Learning from Data - Notations

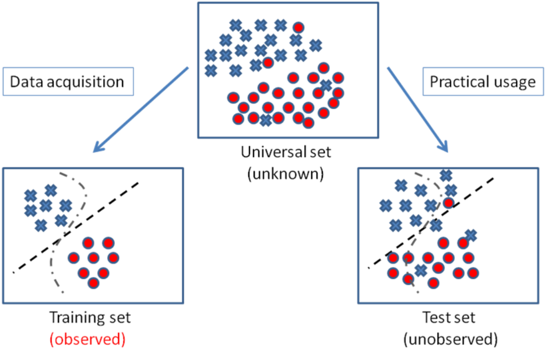
- There are two general dataset types. One is labeled and the other one is unlabeled.
- Labeled dataset: $\mathbb{D} = \{x^{(n)} \in R^d\}_{n=1}^N, Y = \{y^{(n)} \in R\}_{n=1}^N$
- Unlabeled dataset: $\mathbb{D} = \{x^{(n)} \in R^d\}_{n=1}^N$
  where$X$ denotes the feature set containing N samples. Each sample is a d-dimensional vector $x^{(n)} = [x_1^{(n)}, x_2^{(n)}, \cdots, x_d^{(n)}]^T$ and called a feature vector or feature sample, while each dimension of a vector is called an attribute, feature, variable, or element. Y stands for the label set (the color assigned on each point in figures)
- Another form of labeled dataset is described as: $\{x^{(n)} \in R^d, y^{(n)} \in R\}_{n=1}^N$, where each $\{x^{(n)}, y^{(n)}\}$ is called a data pair.
- Unknown $y_n = f(x_n)$

Data Set: Training / Test Set

- An example of two-class dataset is shown
- Two measurements of each sample are extracted
- In this case, each sample is a $2D$ vector

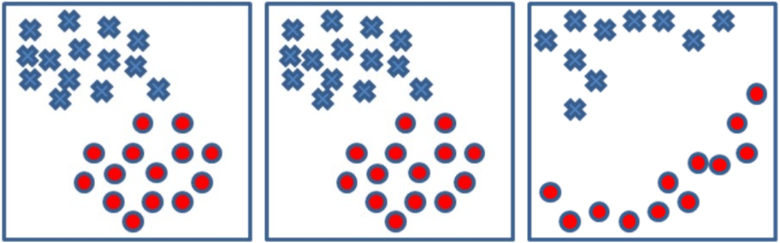


An Explanation of Three Labeled Datasets



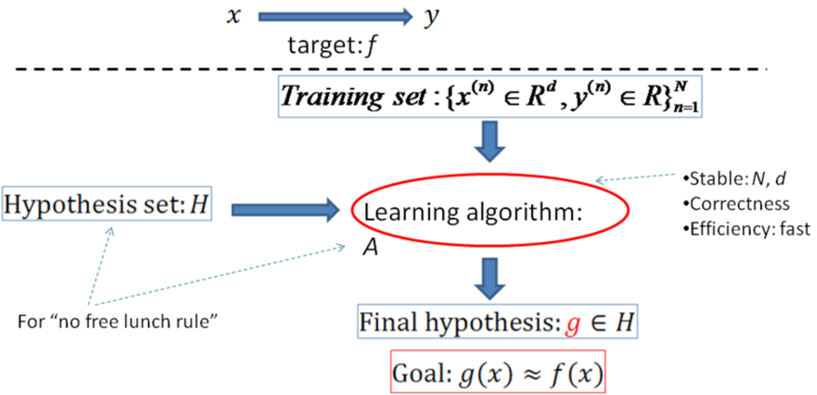
The universal set is assumed to exist but unknown, and through the data acquisition process, only a subset of universal set is observed and used for training (training set). Two learned separating lines are shown in both the training set and test set. As you can see, these two lines definitely give 100% accuracy on the training set, while they may perform differently in the test set (the curved line shows higher error rate)

No Free Lunch Rule



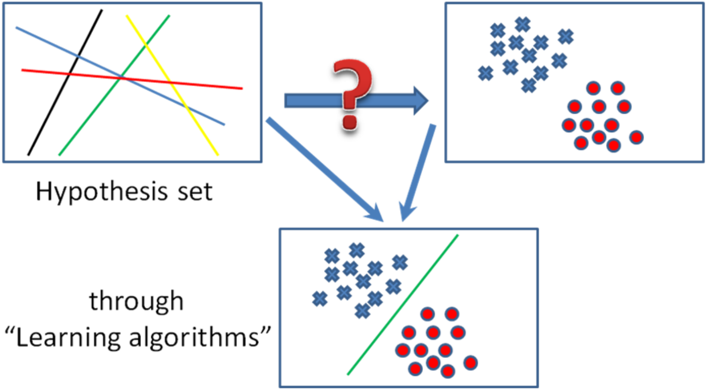
The no free lunch rule for dataset: (a) is the training set we have, and (b), (c) are two test sets.
As you can see, (c) has different sample distributions from (a) and (b), so we cannot expect that the properties learned from (a) to be useful in (c)

Supervised Learning Structure


$x \xrightarrow{\text{target}: f} y$
Training set : $\{x^{(n)} \in R^d, y^{(n)} \in R\}_{n=1}^{N}$
Hypothesis set: H
Learning algorithm: A
•Stable: N, d
•Correctness
•Efficiency: fast
For "no free lunch rule"
Final hypothesis: $g \in H$
Goal: $g(x) \approx f(x)$

The overall illustration of supervised learning structure: The part above the dotted line is assumed but inaccessible, and the part below the line is trying to approximate the unknown target function (f is the true target function and g is the learned function)

An illustration of Hypothesis Set and Learning Algorithm



Hypothesis set

through
“Learning algorithms”

Take linear classifiers as an example, there are five hypothesis classifiers shown in the up-left rectangle, and in the up-right one, a two-class training set in shown. Through the learning algorithm, the green line is chosen as the most suitable classifier

Relationships with Other Disciplines

Machine learning involves the techniques and basis from both statistics and computer science:

- Statistics: Learning and inference the statistical properties from given data
- Computer science: Efficient algorithms for optimization, model representation, and performance evaluation.

In addition to the importance of data set, machine learning is generally composed of the two critical factors, modeling and optimization. Modeling means how to model the separating boundary or probability distribution of the given training set, and then the optimization techniques are used to seek the best parameters of the chosen model.

Machine learning is also related to other disciplines such as artificial neural networks, pattern recognition, information retrieval, artificial intelligence, data mining, and function approximation, etc.

Compared to those areas, machine learning focus more on why machine can learn and how to model, optimize, and regularize in order to make the best use of the accessible training data

Designing versus Learning

- Sky is cloudy, we may decide to bring an umbrella or not.
- For a machine to make these kinds of choices, the intuitive way is to model the problem into a mathematical expression.
- The mathematical expression could directly be designed from the problem background. For instance, the vending machine could use the standards and security decorations of currency to detect false money.
- For problems that we can only acquire several measurements and the corresponding labels, but do not know the specific relationship among them, learning will be a better way to find the underlying connection.
- In many literatures, the knowledge acquired from human understandings or the intrinsic factors of problems are called the domain knowledge. The knowledge learned from a given training set is called the data-driven knowledge.

Machine Learning Paradigms

- Different paradigms of machine learning algorithms:
  - Supervised: Labeled training data (x, y)
  - Unsupervised: Unlabeled training data (x)
  - Reinforcement: No training data (learns with rewards and penalties for every action)
  - Evolutionary algorithms: Optimization technique
- New paradigms:
  - Semi-supervised: Combining supervised and unsupervised
  - Hybrid algorithms: Combining supervised with reinforcement

Learning...

- Supervised - tries to find the relationships between the feature set and the label set, which is the knowledge and properties we can learn from labeled dataset. If each feature vector x is corresponding to a label $y \in L, L = \{l_1, l_2, \cdots, l_c\}$, the learning problem is classification. If each feature vector x maps to a real value $y \in R$, it is regression.
- Unsupervised - aims at clustering, probability density estimation, finding association among features, and dimensionality reduction
- Reinforcement is used to solve problems of decision making (usually a sequence), such as robot perception and movement, automatic chess player, and automatic vehicle driving.
- Semi-supervised attracted increasing attention recently, defined between supervised and unsupervised learning, contains both labeled and unlabeled data, and jointly learns knowledge from them

Learning...

- Supervised - tries to find the relationships between the feature set and the label set, which is the knowledge and properties we can learn from labeled dataset. If each feature vector x is corresponding to a label $y \in L, L = \{l_1, l_2, \cdots, l_c\}$, the learning problem is classification. If each feature vector x maps to a real value $y \in R$, it is regression.
- Unsupervised - aims at clustering, probability density estimation, finding association among features, and dimensionality reduction
- Reinforcement is used to solve problems of decision making (usually a sequence), such as robot perception and movement, automatic chess player, and automatic vehicle driving.
- Semi-supervised attracted increasing attention recently, defined between supervised and unsupervised learning, contains both labeled and unlabeled data, and jointly learns knowledge from them

Learning...

- Supervised - tries to find the relationships between the feature set and the label set, which is the knowledge and properties we can learn from labeled dataset. If each feature vector x is corresponding to a label $y \in L, L = \{l_1, l_2, \cdots, l_c\}$, the learning problem is classification. If each feature vector x maps to a real value $y \in R$, it is regression.
- Unsupervised - aims at clustering, probability density estimation, finding association among features, and dimensionality reduction
- Reinforcement is used to solve problems of decision making (usually a sequence), such as robot perception and movement, automatic chess player, and automatic vehicle driving.
- Semi-supervised attracted increasing attention recently, defined between supervised and unsupervised learning, contains both labeled and unlabeled data, and jointly learns knowledge from them

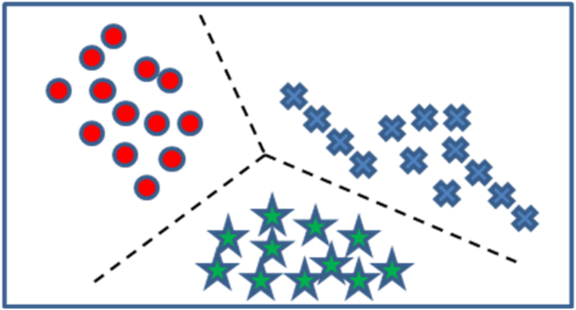
Learning...

- Supervised - tries to find the relationships between the feature set and the label set, which is the knowledge and properties we can learn from labeled dataset. If each feature vector x is corresponding to a label $y \in L, L = \{l_1, l_2, \cdots, l_c\}$, the learning problem is classification. If each feature vector x maps to a real value $y \in R$, it is regression.
- Unsupervised - aims at clustering, probability density estimation, finding association among features, and dimensionality reduction
- Reinforcement is used to solve problems of decision making (usually a sequence), such as robot perception and movement, automatic chess player, and automatic vehicle driving.
- Semi-supervised attracted increasing attention recently, defined between supervised and unsupervised learning, contains both labeled and unlabeled data, and jointly learns knowledge from them

Machine Learning Paradigms...

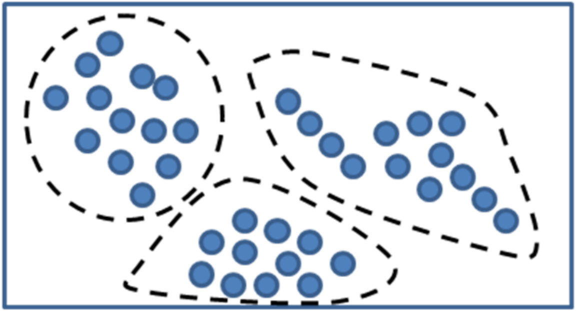
- Another categorization is discriminative vs generative algorithms
    - Discriminative algorithms use the given training data to learn the classification model $p(x|y)$
    - Example: SVM, Neural Network
    - Generative algorithms use the given training data to first compute $p(x, y)$ and then learn the classification model $p(x|y)$
    - Example: Bayes decision theory

Supervised Learning



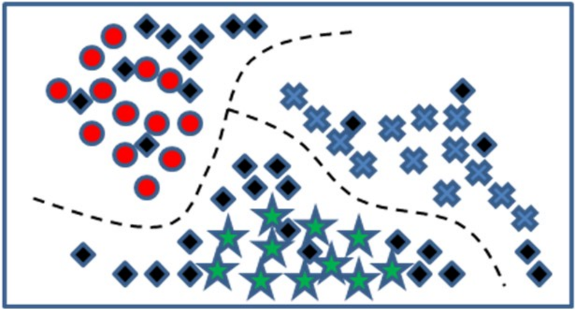
A three-class labeled dataset, where the color shows the corresponding label of each sample. After supervised learning class-separating boundary could be found as the dotted lines

Unsupervised Learning (clustering)



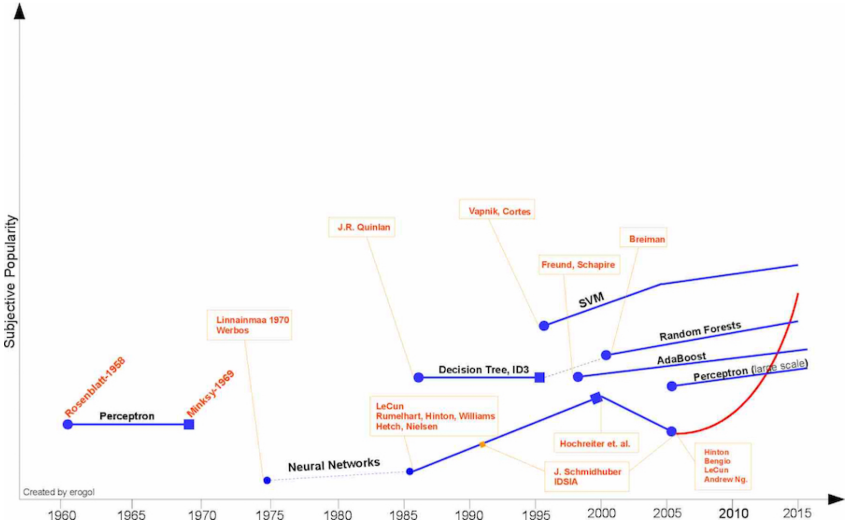
Shows the same feature set as above while missing the label set. After performing the clustering logarithm, three underlined groups are discovered from the data. Also, users can perform other kinds of unsupervised learning algorithm to learn different kinds of knowledge (ex. Probability distribution) from the unlabeled dataset.

Semi-supervised Learning



Presents a labeled dataset (with red, green, and blue) together with a unlabeled dataset (marked with black). The distribution of the unlabeled dataset could guide the position of separating boundary. After learning, a different boundary is depicted against the one in supervised learning.

ML Techniques



- SVM is the most popular technique - should we apply SVM for every classification?
- No
- Why?
- Every algorithm has some assumption and properties
- Choose the algorithm depending on the data characteristics and properties of the algorithm

A simple learning model

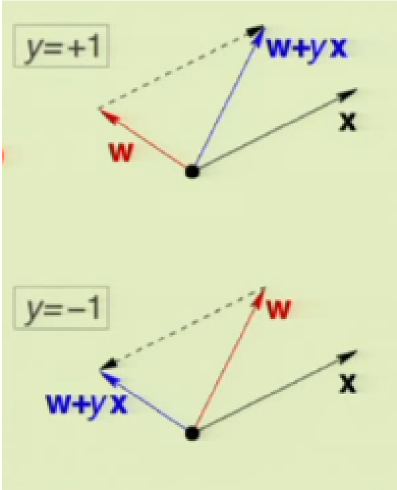
- Specific learning problem (target function and training examples are dictated by the problem)
- Learning model (hypothesis set and learning algorithm)
- $X \in R^d$ and $y = \{+1, -1\}$
- $h \in H$, for example $h(x) = sign((\sum_{i=1}^{d} w_i x_i)) + b$

The Perceptron Learning Algorithm

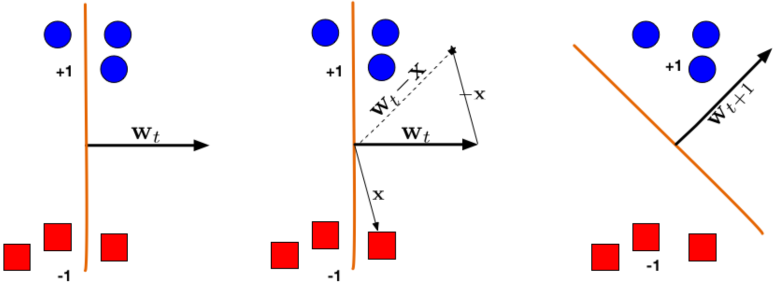
- Aim: Learn what w should be based on data (linearly separable). Learn a vector w that achieve the correct decision $h(x_n) = y_n$ on all the training examples
- Let $w_0 = b$ thus $w = [w_0, w_1, \cdots, w_d]^T$ and $x = [1, x_1, \cdots, x_d]^T$
- Formally, the input space is $X = \{1\} \times R^d = \{[x_0, x_1, \cdots, x_d]^T | x_0 = 1, x_1 \in R, \cdots, x_d \in R\}$
- $w^T x = \sum_{i=0}^{d} w_i x_i$, thus $h(x) = sign(w^T x)$
- Algorithm: At iteration t, there is current weight $w(t)$, the alg. picks up an example from $(x_1, y_1) \cdots (x_N, y_N)$ that is currently misclassified, call it $(x(t), y(t))$
  Thus, $y(t) \neq sign(w^T(t)x(t))$
  The update rule is $w(t+1) = w(t) + y(t)x(t)$
- This rule moves the boundary in the direction of classifying $x(t)$ correctly.
- Continue until there are no misclassified examples in the dataset.

weight update

weight update



weight update


+1
-1
$\mathbf{w}_t$
+1
-1
$\mathbf{w}_t - \mathbf{x}$
$-\mathbf{x}$
$\mathbf{w}_t$
$\mathbf{x}$
+1
-1
$\mathbf{w}_{t+1}$