Week 5

From the Expert: Introduction to Machine Learning

Machine learning is a subfield of Artificial Intelligence. Machine learning involves algorithms that **learn** to solve problems based on sample data, observations, experiences or environments rather than explicit programing instructions. Machine learning algorithms automatically learn and generalize from data with no human interventions.

Machine learning has spread rapidly with the applications such as spam filtering, fraud detection, customer segmentation, optical character/digit written recognition, medical diagnosis, stock trading and many others.

According to Tom Mitchell (1997), machine learning means "a computer program that improves its performance at some task through experience."

Machine learning algorithms can be categorized into 3 main groups:

- Supervised learning

The learner is trained with known outputs (targets) or labeled examples.
For instance, labels can be categories (i.e. default, not default) or numeric values such as stock prices that we want to predict. The labeled examples will be employed during the training process. However, in the testing process only unlabeled data are used. The model has to predict the label from the given unlabeled data. The responses (predicted) from the learner and their corresponding targets are compared. The appropriate adjustments are performed to ensure that the learner predict outputs with high accuracy. Examples of supervised learning include artificial neural networks, decision trees, Naïve Bayes, Bayesian believe network, logistic regression etc.

- Unsupervised learning

The learner is trained with no label data; therefore, the learner must try to find hidden structures in the data. The input data are clustered into classes based on their feature properties. Unsupervised learning includes self-organizing maps (SOM), K-means.

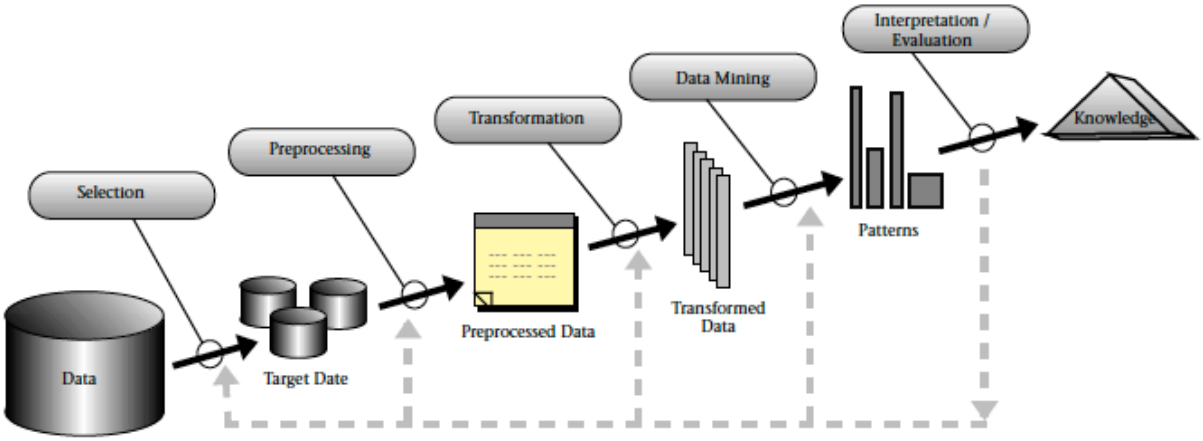
- Reinforcement learning

The learner learns by interaction with the environment. The input from environment is fed to the learner, after that the action/response is determined. Then, the learner learns to adjust its behavior from the environmental feedback. Over time, the best performance is achieved by trial and error. Examples of reinforcement learning are robot control systems, Q-learning, etc.

Machine learning is closely related to data mining. However, the central focus of data mining is on searching for patterns and relationships that may exist in the database. Data mining employs various known techniques from machine learning, pattern recognition, and statistics for analyzing and discovering patterns in the data. Data mining can be considered as a crucial part of the knowledge discovery in databases (KDD). Some other

parts in KDD include data preparation, data selection, and data cleaning (Fayyad et.al 1996).

Piatetsky-Shapiro claimed that the term "knowledge discovery in databases" was first used at KDD workshop in 1989 to emphasize "knowledge is the end product of a data-driven discovery" In the view points of Fayyad et. al. (1996), KDD endeavors to solve data overload problem in this digital age. They further explained that KDD is "the overall process of discovering useful knowledge from data including how the data are stored and accessed, how algorithms can be scaled to massive data sets and still run efficiently, how results can be interpreted and visualized, and how the overall man-machine interaction can usefully be modeled and supported"



An overview of the steps in KDD process (Fayyad et.al 1996)

The two primary goals of data mining are used for predictive modeling and descriptive modeling. Predictive modeling is related to using some variables to predict future or unknown values of the concerned variables. The examples of predictive modeling are classification, regression, time series analysis etc. On the other hand, the descriptive modeling targets on disclosing interpretable patterns in the data or summarized the underlying patterns or relationships in data such as trends, correlations, clusters, and anomalies, etc. In addition to the exploratory process, post-processing techniques are required for validation and result explanation. Examples of descriptive modeling contain clustering, summarization, and association rules.

General tasks of data mining:
Classification is a function that maps input data into one of predefined classes. For instance, classifies clients on the loan requests as (loan, no loan).

Typically, the process of learning can be divided into 2 phases: training phase and testing phase. The purpose of the training phase is to build models with the generalization capability. Generalization can be described as the ability to accurately predict the class labels that have never before seen during the training phase. The testing phase is used to

measure the performance of the model. (i.e. number of records that predicted correctly and incorrectly). In the same manner, inputs can be divided into 2 non-overlapped data sets: training set and testing set. Training set is used in the training phase with the label examples. Testing set with no label examples is used to assess the performance of the model. It is not uncommon to have another data set called validation set, which are examples that used to adjust the architectures and error estimation of the model (after the training phase). Classification techniques include decision trees, rule-based classifier, Naïve Bayes, support vector machine, neural networks, etc.

Regression is the function that maps a data item to a real-valued prediction. The applications are used to estimate the probability of the survival patients based on diagnosis tests, predict the customer default based on income, and predict the revenue from the advertisement cost.

Clustering is the example of descriptive modeling where a set of clusters is identified as the result of data description. Here, objects in the same group are similar in the group but different from objects in other groups. The inputs are grouped according to similarity metrics such as Euclidean distance, Jaccard similarity, Tanimoto coefficient, and cosine similarity on data objects. One of the most often used techniques is K-means.

Summarization is a general overview of the data usually dealing with aggregation information in various ways. There are different abstraction levels, different angles, and different combinations. For example, sale of national brand convenience stores can be summarized into the total sales by day, the total sales by month, and the total sales by year for each individual store. Sales can be summarized by state, or by regions. We may want to see the totals sales for a specific product at specific time period across all stores etc.

Association Analysis signifies the co-occurrence of the items. The central focus is on extracting interesting patterns that underlies the associated features in data. The discovered relationships can be formulated as association rules. The applications consist of market basket analysis (items that customers buy together), genes that co-exist with particular diseases, etc.

Anomaly detection is to pinpoint data points that behave significantly different from other data points. These different data points are called anomalies or outliers. Ideally, anomaly detection algorithms should have high detection rates but low false alarm rates. The applications include credit card fraud, network intrusion, rare disease pattern, etc.

Common Data Mining Methodologies

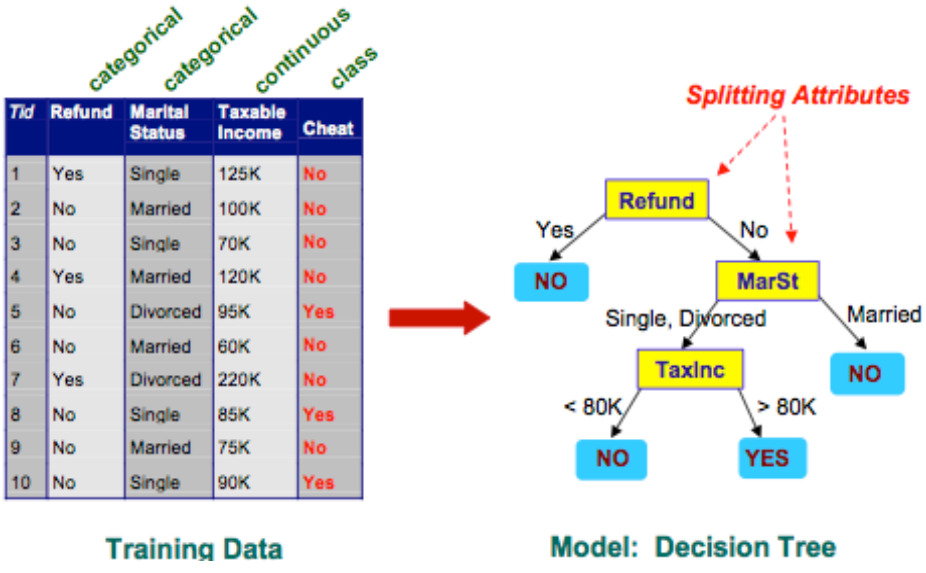
Decision Trees is primary used for predictive modeling in classification and regression. The model has a representative form that can be easily understood and interpreted. Greedy search methods has been involved in growing and pruning tree structures. Decision tree algorithms include ID3, C4.5 (Quinlan, 1993) and CART.

Given a set of records, and each record contain a number of the attributes and their corresponding value. One of these attributes is the category or class attribute. The objective of decision tree is to predict the accurate class attribute from the non-class attributes. The tree structure depends on the chosen features and threshold.

A decision tree is built recursively. The training records (attribute/value pairs) are partitioned into subsets and stop when the records have the same class labels, the same attribute values, number of records fall below minimum threshold, or maximum depth is reached. In the growing tree process, attributes with the most information (goodness of split) among the rest of attributes are considered. Note that selecting the best split often relies on the degree of impurity, the smaller the better. A node with (0,1) class distribution has impurity = 0, whereas a node with (0.5,0.5) uniform distribution has the highest impurity. Here, we want the node to be pure or one class dominates. The impurity measures employed include Entropy and Gini index. The attribute with the highest information gain will be selected. In ID3, Information gain can be calculated from finding the difference between the values of impurity measure (i.e. Entropy) before split and after split. For example, $Gain(X,T) = Info(T) - Info(X,T)$. In C4.5, the gain ratio is utilized, which is defined as $GainRatio(X,T) = \frac{Gain(X,T)}{SplitInfo(X,T)}$ Here, The number of outcomes from the attributes test condition is taking into consideration.

To avoid overfitting, the tree should be pruned. In simple words, the certain depth of the subtree below would be cut off. Overfitting happens when the model does not generalize well or gives less accuracy in a new data set.

This example (Tan, 2006) shows decision tree for tax-cheat classification (yes/no) using refund, marital status and income attributes.



Bayes Theorem (simple form)

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Let X denote the attribute set and Y denote the class variable
$P(Y)$ is the prior probability of Y
$P(X)$ is the prior probability of training data X (evidence)
$P(Y|X)$ is the probability of Y given X (posterior probability for Y)
$P(X|Y)$ is the probability of X given Y (class conditional probability)
In Naïve Bayes, the class conditional probability assumes that the attributes are conditionally independent, given the class label y. Thus, the conditional probability can be represented as:

$$P(X|Y = y) = \prod_{i=1}^{d} P(X_i|Y = y)$$

Here, the attribute set Xconsists of d attributes $\{X_1, X_2, X_3, \ldots, X_d\}$. And, $P(X)$ is fixed for every Y. Therefore, the posterior probability can be formulated as:

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{d} P(X_i|Y = y)}{P(X)}$$

This example illustrates Naïve Bayes classifier for mammals or non-mammals class (Tan, 2006).

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Bayesian Believe Network (BBN)

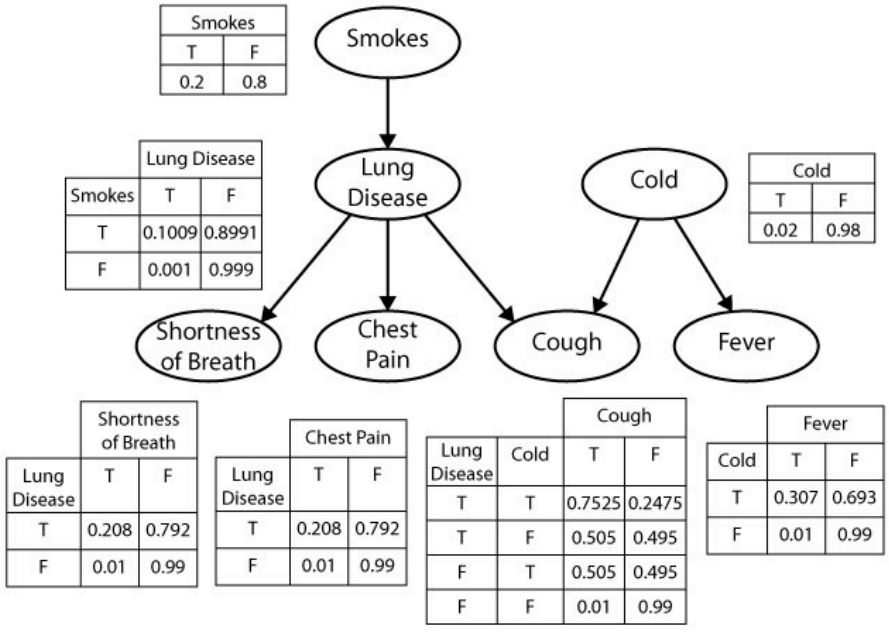
Bayesian network is a graphical model that represents the probabilistic relationships among random variables. However, instead of assuming that all attributes from class conditional probability $P(X|Y)$ are conditionally independent as in Naïve Bayes, the Bayesian network approach can specify which attribute pairs are conditionally independent.

The important components of Bayesian network (Tan, 2006) include:

1) The dependence relationships in a set of variables are presented with a directed acyclic graph (DAG).
2) Each node has its own probability table that associated with its immediate parent nodes.

Bayesian network can be applied in medicine, bioinformatics, and decision support systems.

This figure shows an example of Bayesian network for medical diagnosis (https://probmods.org/patterns-of-inference.html)



Smokes	
T	F
0.2	0.8

Smokes	Lung Disease	
	T	F
T	0.1009	0.8991
F	0.001	0.999

Cold	
T	F
0.02	0.98

Lung Disease	Shortness of Breath	
	T	F
T	0.208	0.792
F	0.01	0.99

Lung Disease	Chest Pain	
	T	F
T	0.208	0.792
F	0.01	0.99

Lung Disease	Cold	Cough	
		T	F
T	T	0.7525	0.2475
T	F	0.505	0.495
F	T	0.505	0.495
F	F	0.01	0.99

Cold	Fever	
	T	F
T	0.307	0.693
F	0.01	0.99

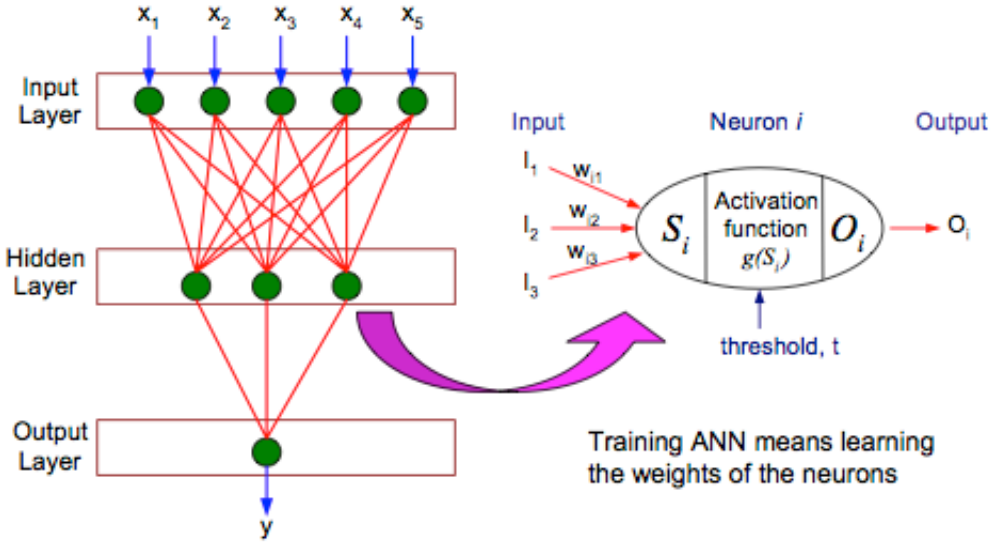
Artificial neural network (ANN)

ANN was inspired by biological neurons. In human nervous system, there are many elements working in parallel and connecting with one another. The neural structure consists of many inputs and one output unit. The output unit can be excited (fired) or not excited (not fired). The signals (inputs) from other neurons are summed up and compared with the threshold value to determine whether the neuron should fire.

To simulate the structure of human brain, ANN consists of many computing units called nodes. Nodes are connected together to form a network. The connection between nodes are called weights, which encode with the learning information. In general, a node receives input from the external source or from other nodes. Each input is associated with the weight. The node computes the weighted sums of its inputs, including applying a function called activation function to this result. The output from the computation can be served as input to other units.

There are many ANN architectures. The simple one called perceptron by Rosenblatt. The perceptron is also known as a single-layer perceptron. It is a linear classifier that consists of only input layer (receive inputs from the environment) and output layer (produce output back to the environment). It is the simplest feed-forward network. The feed-forward network is the network where nodes in one layer connected to the nodes in the next layer. The network, which consists of hidden layers; the intermediate layers between input and output nodes, is known as the multilayer neural networks. The multilayers with at least one hidden layer are universal approximators (Hornik, 1991). This means that the networks can approximate any functions given that they have enough computation units.

Here is the Figure for multilayer neural networks (Tan, 2006).



The most popular algorithm to train the network is back-propagation innovated by Rumelhart and McClelland. In short, back-propagation adjusts weights based on activation functions such as sigmoid. The objective is to reduce the sum of squared error between the desired output (target) and the output calculated by the network. Back-propagation is an example of supervised learning where the target is known.

Applications of neural networks such as stock market prediction, hand written / spoken words recognition, image compression, loan/credit cards approval, and control system (i.e. self driving car).

K-means

K-means is unsupervised learning. It is one of the popular techniques for grouping (clustering) objects because of its simplicity and is reasonably fast. The objective of K-mans is to partition n observation objects into K clusters so that the distance between objects and the cluster centroids are minimized. The term centroid is usually referred to the mean of the objects cluster. At the beginning, the cluster numbers (i.e. constant K) are specified. The algorithm will randomly choose the K objects as the initial centroids. Then, these 2 steps are repeated until centroids converge or do not change 1) assign each object to one of K clusters based on its closet centroid 2) re-compute the new centroid for each cluster.

Assigning objects to the appropriate clusters can be expressed using the similarity measure such as Euclidean distance, Cosine, etc. Using the right measures of closeness will have a great impact on the results. Some disadvantages of this algorithm include result from each run may not be the same; the number of clusters must be specified at the beginning.

References

Decision Trees: sections 8.1-8.4. (n.d.) Retrieved from http://www.cse.msu.edu/~cse802/DecisionTrees.pdf

Domingos, P. (2012). *A few useful things to know about machine learning*. ACM Digital Library. Vol. 55, Issue 10, October. Retrieved from http://dl.acm.org/citation.cfm?id=2347755.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. *From data mining to knowledge discovery in databases*. American Association for Artificial Intelligence, pgs. 37-53. Retrieved from http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf.

Hornik, K (1991). *Approximation Capabilities of Multilayer Feedforward Networks*. ScienceDirect: Neural Networks. Vol. 4, Issue 2, pgs. 251–257. Retrieved from http://www.sciencedirect.com/science/article/pii/089360809190009T.

Meisner, E. (November 22, 2003). *Naïve bayes classifier example.* Retrieved from http://www.cs.rpi.edu/courses/fall03/ai/misc/naive-example.pdf.

Mitchell, T. (1997). *Lecture slides for textbook Machine Learning.* Retrieved from http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/mlbook/ch3.pdf, and http://www.cs.cmu.edu/~tom/mlbook-chapter-slides.html.

Patterns of Inference (n.d.). Probabilistic Models of Cognition, chapter 4. Retrieved from https://probmods.org/patterns-of-inference.html.

Quinlan,J. R. (September 1987). *Simplifying decision trees*. ACM Digital Library. Retrieved from http://dl.acm.org/citation.cfm?id=50008.

Schapire, R. (February 4, 2008). *Lecture COS 511: Theoretical machine learning*. Retrieved from http://www.cs.princeton.edu/courses/archive/spr08/cos511/scribe_notes/0204.pdf.