Machine Learning [ETCS-402]

Dr. A K Yadav
Amity School of Engineering and Technology
(affiliated to GGSIPU, Delhi)
akyadav1@amity.edu
akyadav@akyadav.in
www.akyadav.in
+91 9911375598

April 15, 2021

What is Learning?

- How to Learn?
  - By example.
  - By experience.

What is Machine Learning?

Definitaion: A computer progragram is said to learn from experience E with respect to some class of tasks T and performance P if its performance P improves with experience E on tasks T.
A checkers learning problem:

- Task T: playing checkers
- Performance measure P: percentage of games won against opponents
- Training experience E: playing practice games against itself.

A handwriting recognition learning problem:

- Task T: recognizing and classifying handwritten words within images
- Performance measure P:percent of words correctly classified
- Training experience E: a database of handwritten words with given classifications

A robot driving learning problem:

1. Task T: driving on public four-lane highway using vision sensors
2. Performance measure P: average distance travelled before an error
3. Training experience E: a sequence of image and steering commands recorded while observing a human driver.

How to make Machine to Learn?

- Direct i.e. By experience.
- Indirect i.e. By example.

Design a Learning System

- Choosing the Training Experience.
- Choosing the Target Function.
- Choosing a Representation for the Target Function.
- Choosing a Function Approximation Algorithm.

Choosing the Training Experience

- Type of training experience.
- Degree of Control of Training Examples.
- Distribution of Examples.

Choosing the Training Experience

- Type of training experience:
  - Direct feedback
  - Indirect feedback

    *Learning from direct training feedback is typically easier than learning from indirect feedback.
- Degree of Control of Training Examples.
  - Action as per Teacher.
  - Challenges to Teacher.
  - Complete control on both.
- Distribution of Examples.
  - Actual Game
  - Played against itself
  - Both.

Issues in Machine Learning

The field of machine learning is concerned with answering questions such as the following:

- What algorithms exist for learning general target functions from specific training examples?
- In what settings will particular algorithms converge to the desired function, given sufficient training data?
- Which algorithms perform best for which types of problems and representations?
- How much training data is sufficient?
- What general bounds can be found to relate the confidence in learned hypotheses to the amount of training experience and the character of the learner's hypothesis space?

- When and how can prior knowledge held by the learner guide the process of generalizing from examples? Can prior knowledge be helpful even when it is only approximately correct?
- What is the best strategy for choosing a useful next training experience, and how does the choice of this strategy alter the complexity of the learning problem?
- What is the best way to reduce the learning task to one or more function approximation problems? Put another way, what specific functions should the system attempt to learn? Can this process itself be automated?
- How can the learner automatically alter its representation to improve its ability to represent and learn the target function?

Goal and Applications of Machine Learning

Goal: Designing a system which can improve their performance themselves without human interference.
Applications:

- Image Classification
- Anti Virus
- Spam Filter
- Face Recognition
- Face Detection
- Medical Diagnosis
- Speech Recognition
- Customer Segmentation
- Fraud Detection
- Weather Predictions
- News Spotting

Types of Learning

- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
  - Clustering
  - Association
- Reinforcement Learning: feedback wrong result but not how to correct.
- Semi supervised Learning is a combination of supervised and unsupervised learning

Areas of Machine Learning

Machine learning algorithms are used primarily for the following types of output:

- Clustering (Unsupervised)
- Two-class and multi-class classification (Supervised)
- Regression: Univariate, Multivariate, etc. (Supervised)
- Anomaly detection (Unsupervised and Supervised)
- Recommendation systems (aka recommendation engine)

The value of machine learning technology has been recognized by companies across several industries that deal with huge volumes of data. By leveraging insights obtained from this data, companies are able work in an efficient manner to control costs as well as get an edge over their competitors. This is how some sectors / domains are implementing machine learning :-

- **Financial Services:** Companies in the financial sector are able to identify key insights in financial data as well as prevent any occurrences of financial fraud, with the help of machine learning technology. The technology is also used to identify opportunities for investments and trade. Usage of cyber surveillance helps in identifying those individuals or institutions which are prone to financial risk, and take necessary actions in time to prevent fraud.
- **Marketing and Sales:** Companies are using machine learning technology to analyze the purchase history of their customers and make personalized product recommendations for their next purchase. This ability to capture, analyze, and use customer data to provide a personalized shopping experience is the future of sales and marketing.

- **Government:** Government agencies like utilities and public safety have a specific need for machine learning, as they have multiple data sources, which can be mined for identifying useful patterns and insights. For example sensor data can be analyzed to identify ways to minimize costs and increase efficiency. Furthermore, ML can also be used to minimize identity thefts and detect fraud.
- **Healthcare:** With the advent of wearable sensors and devices that use data to access health of a patient in real time, ML is becoming a fast-growing trend in healthcare. Sensors in wearable provide real-time patient information, such as overall health condition, heartbeat, blood pressure and other vital parameters. Doctors and medical experts can use this information to analyze the health condition of an individual, draw a pattern from the patient history, and predict the occurrence of any ailments in the future. The technology also

empowers medical experts to analyze data to identify trends that facilitate better diagnoses and treatment.

- **Transportation:** Based on the travel history and pattern of traveling across various routes, machine learning can help transportation companies predict potential problems that could arise on certain routes, and accordingly advise their customers to opt for a different route. Transportation firms and delivery organizations are increasingly using machine learning technology to carry out data analysis and data modeling to make informed decisions and help their customers make smart decisions when they travel.
- **Oil and Gas:** This is perhaps the industry that needs the application of machine learning the most. Right from analyzing underground minerals and finding new energy sources to streaming oil distribution, ML applications for this industry are vast and are still expanding.

Aspects of Training Data

Data pre-processing is an integral step in Machine Learning as the quality of data and the useful information that can be derived from it directly affects the ability of our model to learn; therefore, it is extremely important that we preprocess our data before feeding it into our model. The process for getting data ready for a machine learning algorithm can be summarized in these steps:

1. Select Data
2. Pre-process Data
3. Transform Data

Select Data

Selecting the right dataset for Machine learning is very important to make the AI model functional with right approach. Though selecting the right quality and amount of data is challenging task but there are few rules needs to be followed for machine learning on big data. There is always a strong desire for including all data that is available, that the maxim "more is better" will hold. This may or may not be true.

We need to consider what data we actually need to address the question or problem we are working on. Make some assumptions about the data we require and be careful to record those assumptions so that we can test them later if needed.

Below are some questions to help we think through this process:

a. What is the extent of the data we have available? For example, through time, database tables, connected systems. Ensure we have a clear picture of everything that we can use.

b What data is not available that we wish we had available? For example, data that is not recorded or cannot be recorded. We may be able to derive or simulate this data.

c What data don't we need to address the problem? Excluding data is almost always easier than including data. Note down which data we excluded and why.

It is only in small problems, like competition or toy datasets where the data has already been selected for us. Unstructured data such as images, text or video will eventually need to be converted into a data frame before applying predictive methods, so these metrics listed below apply to all types of data during the machine learning process.

n: Usually the first characteristic of interest in a dataset is its size, N measured as the number of rows or examples.

d: the next descriptor of the data is its dimension, d measured by the number of columns or attributes

k: this descriptor applies only to classification problems. k represents the number of classes. Today most classification problems are binary or k=2. But it is not difficult to envision situations where $k >> 2$.

m: is the ratio of number of samples of the minority class to number of samples of the majority class in a 2-class problem. In a multiclass problem ($k > 2$), this could be the ratio of the number of samples of a given class to the rest of the samples.

Pre-Process Data

The collected data cannot be used directly for performing analysis process. Whenever the data is gathered from different sources, it is collected in raw format which is not feasible for the analysis. Therefore, certain steps are executed to convert the data into a small clean data set. These are the common data pre-processing methods:

1. **Formatting:** The data you have selected may not be in a format that is suitable for you to work with. The data may be in a relational database and you would like it in a flat file, or the data may be in a proprietary file format and you would like it in a relational database or a text file.

2. **Cleaning:** Cleaning data is the removal or fixing of missing data. There may be data instances that are incomplete and do not carry the data you believe you need to address the problem. These instances may need to be removed. Additionally, there may be sensitive information in some of the attributes and these attributes may need to be anonymized or removed from the data entirely.
3. **Sampling:** There may be far more selected data available than you need to work with. More data can result in much longer running times for algorithms and larger computational and memory requirements. You can take a smaller representative sample of the selected data that may be much faster for exploring and prototyping solutions before considering the whole dataset.

4 **Binarize Data (Make Binary):** We can transform our data using a binary threshold. All values above the threshold are marked 1 and all equal to or below are marked as 0. This is called binarizing your data or threshold your data. It can be useful when you have probabilities that you want to make crisp values. It is also useful when feature engineering and you want to add new features that indicate something meaningful.

5 **Standardize Data:** Standardization is a useful technique to transform attributes with a Gaussian distribution and differing means and standard deviations to a standard Gaussian distribution with a mean of 0 and a standard deviation of 1.

6 **Dimensionality Reduction:** When data sets become large in the number of predictor variables or the number of instances, data mining algorithms face the curse of dimensionality problem. It is a serious problem as it will impede the operation of most data mining algorithms as the computational cost rise. This section will underline the most influential dimensionality reduction algorithms according to the division established into Feature Selection (FS) and space transformation-based methods.

Transform Data

The final step is to transform the process data. This step is used to convert the raw data into a specified format according to the need of the model. The methods used for transformation of data are given below :–

1. **Rescale Data:** When our data is comprised of attributes with varying scales, many machine learning algorithms can benefit from rescaling the attributes to all have the same scale. This is useful for optimization algorithms in used in the core of machine learning algorithms like gradient descent. It is also useful for algorithms that weight inputs like regression and neural networks and algorithms that use distance measures like K-Nearest Neighbours.

2. **Decomposition:** There may be features that represent a complex concept that may be more useful to a machine learning method when split into the constituent parts. An example is a date that may have day and time components that in turn could be split out further. Perhaps only the hour of day is relevant to the problem being solved. consider what feature decompositions you can perform.
3. **Aggregation:** There may be features that can be aggregated into a single feature that would be more meaningful to the problem you are trying to solve. For example, there may be a data instances for each time a customer logged into a system that could be aggregated into a count for the number of logins allowing the additional instances to be discarded. Consider what type of feature aggregations could perform.

Data preparation is a large subject that can involve a lot of iterations, exploration and analysis. Getting good at data preparation will make you a master at machine learning. For now, just consider the questions raised in this post when preparing data and always be looking for clearer ways of representing the problem you are trying to solve.

Concept Learning and Concept Representation

- Conceptual learning is an educational method that centres on big-picture ideas and learning how to organize and categorize information.
- Unlike more traditional learning models which concentrate on the ability to recall specific facts, conceptual learning focuses on understanding broader principles or ideas that can later be applied to a variety of specific examples.
- Much of learning involves acquiring general concepts from specific training examples.
- People, for example, continually learn general concepts or categories such as "bird", "car", "situations in which I should study more in order to pass the exam" etc.

- Each such concept can be viewed as describing some subset of objects or events defined over a larger set (e.g., the subset of animals that constitute birds).
- Alternatively, each concept can be thought of as a boolean-valued function defined over this larger set (e.g., a function defined over all animals, whose value is true for birds and false for other animals).
- **Concept learning**: Inferring a boolean-valued function from training examples of its input and output

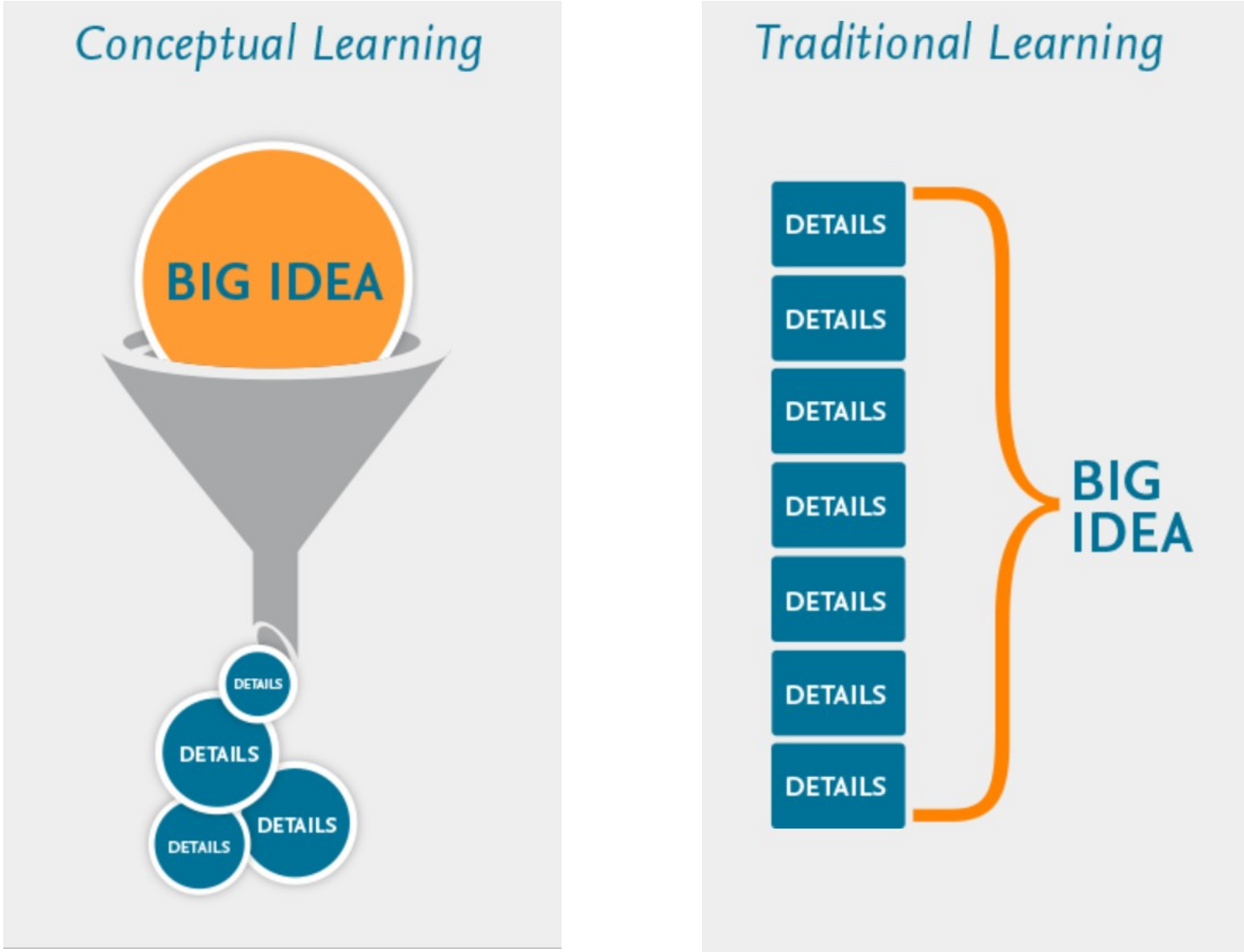

Figure: Conceptual Learning and Traditional Learning

Concepts and Exemplars

Concepts are mental categories for facts, objects, events, people, ideas — even skills and competencies — that have a common set of features across multiple situations and contexts. Concepts can range from simple to complex according to how easily they can be defined.

Examples of concepts :-

- Concrete Concepts
- Abstract Concepts
- Semi-concrete Concepts

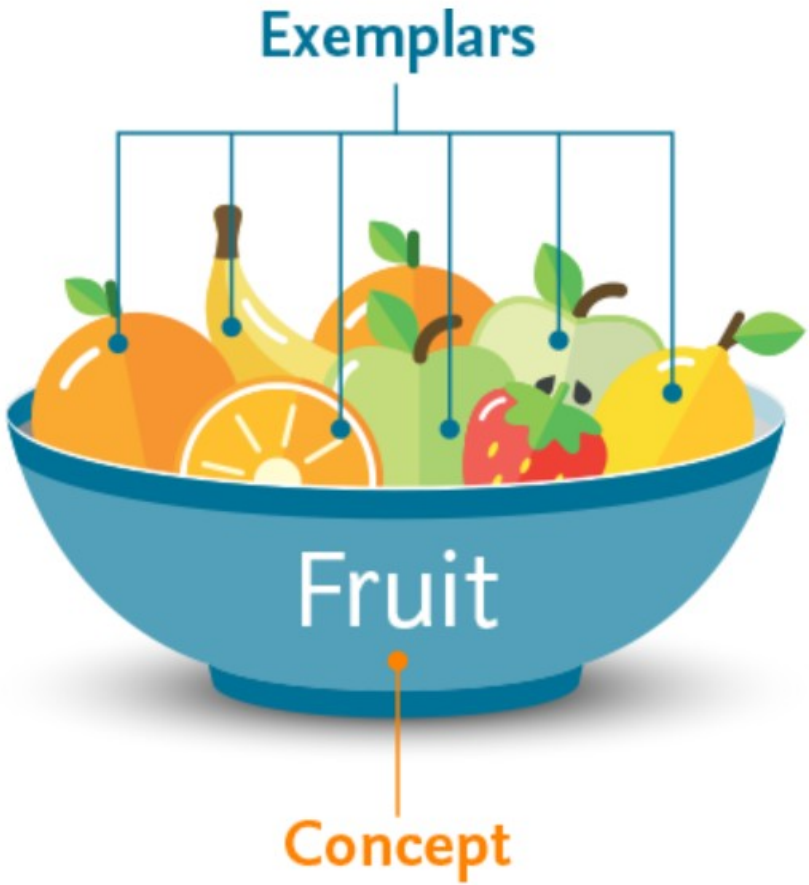

Figure: Concept of Fruit

Function Approximation

Statistical approaches to machine learning are related to function approximation methods in mathematics. In below fig. we have a universe of objects that belong to either of two classes '+' or '-'.

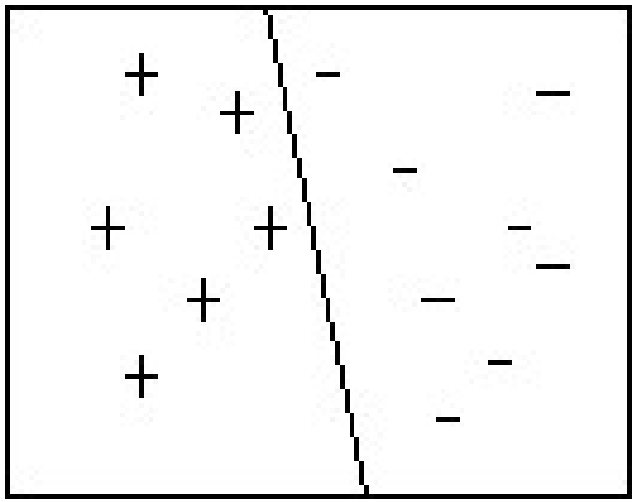

Figure: A linear discrimination between two classes

Sample Space

Sample Space: The set of all possible outcomes of an experiment is called the sample space and is denoted by Ω.
Individual elements are denoted by ω and are termed elementary outcomes.

Examples:

- (Finite) A single roll of an ordinary die. Here, $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- (Countable) Infinite number of coin tosses in order to study, say, the number of tosses before 5 consecutive heads are observed. Here, $\Omega = \{H, T\}^{\infty}$.
- (Uncountable) Speed of a vehicle measured with infinite precision. Here, $\Omega = \mathbb{R}$.

Event

Event: An event is any collection of possible outcomes of an experiment, that is, any subset of Ω.

In most experiments we are generally more interested in observing the occurrence of particular events rather than the elementary outcomes. For example, on rolling a die, we may be interested in observing whether the outcome was even (event $E = \{2, 4, 6\}$) or odd (event $O = \{1, 3, 5\}$).

Over fitting

- Generalization
- Statistical Fit
- A Good Fit in Machine Learning
- Detection of Over fitting
- Prevention of Over fitting

Classification families

- Linear Classifiers: Logistic Regression, Naive Bayes Classifier
- Support Vector Machines
- Decision Trees
- Boosted Trees
- Random Forest
- Neural Networks
- Nearest Neighbor

Combining classifier

- Types of Combined Classifiers
- Bagging
- Boosting
- Evaluating and debugging learning algorithms
- Classification errors

ज्योतिर्वृणीत तमसो विजानन

ज्योतिर्वृणीत तमसो विजानन

Thank you

Please send your feedback or any queries to **akyadav1@amity.edu**
You can contact me on **+91 9911375598**