e-PG Pathshala

Subject : Computer Science

Paper: Machine Learning

Module: Type of Learning I

Module No: CS/ML/4

Quadrant I – e-text

Welcome to the e-PG Pathshala Lecture Series on Machine Learning.

Learning Objectives:

The learning objectives of this module are as follows:

- To understand how machine learns from data provided.
- To explore the types of machine learning and their applications.
- To discuss in detail the various representations of data for learning

4. 1 Types of Learning

There are many ways in which types of learning can be categorized. In this section we will discuss the categorization of the types of learning depending on the extent of feedback. There are basically four types of learning, supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. Let us now describe these types of learning in terms of feedback provided.

Supervised Learning: In this type of learning the training data includes the desired outputs.
Unsupervised Learning: In this type of learning the training data does not include the desired outputs.
Semi-supervised Learning: In this type of learning training data includes only some of the desired outputs.
Reinforcement Learning: In this type of learning rewards are received as a result of sequential actions.

4.2 Supervised Learning

Now let us discuss Supervised Learning in detail. This is the type of learning where the training data includes desired outputs. Essentially the system tries to learn a function from examples of its inputs and outputs and then learns to predict output when given an input vector.

Some of the advantages of supervised learning include the ability to learn complex patterns and in general the method exhibits good performance. However the method requires a lot of output labeled data which is time consuming and costly to acquire.

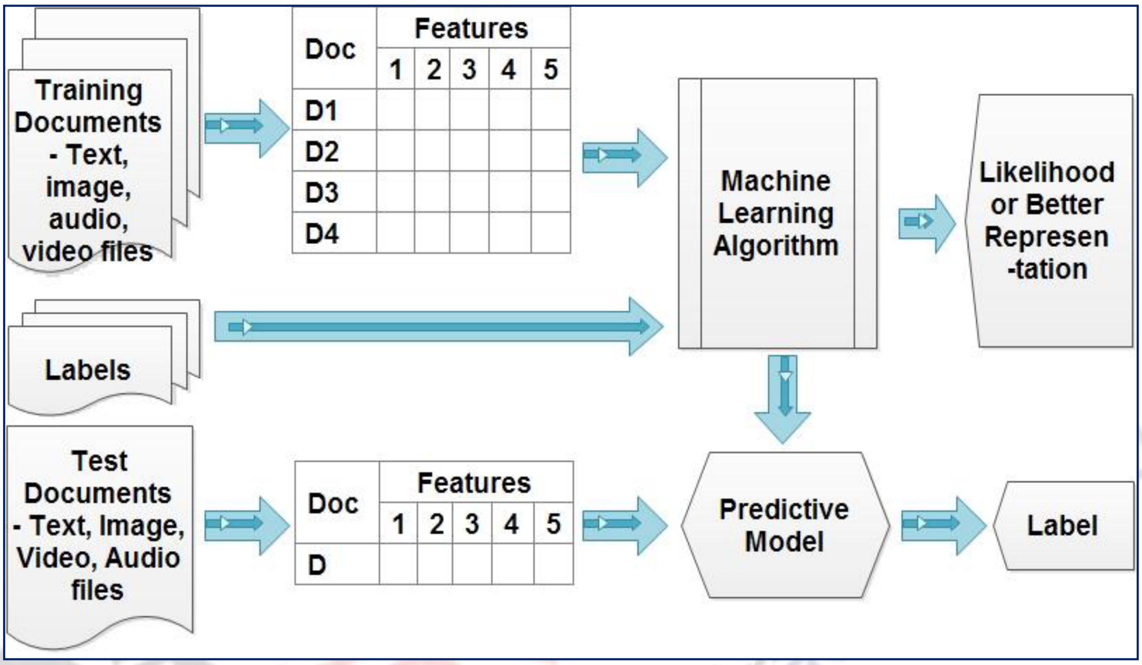
Training Documents - Text, image, audio, video files
Doc
Features
1
2
3
4
5
D1
D2
D3
D4
Machine Learning Algorithm
Likelihood or Better Represen -tation
Labels
Test Documents - Text, Image, Video, Audio files
Doc
Features
1
2
3
4
5
D
Predictive Model
Label

Figure 4.1 Overview of Supervised Learning

Figure 4.1 shows an overview of supervised learning. The input training data consisting of training documents (could be text, image, audio or video files), is represented as a vector of features (in this example five features are shown) along with the label of each document. This feature representation is given to the machine learning algorithm that learns a predictive model and could also learn a better representation model.
Now the test documents that are to be labeled are represented using the same feature space as the training data. These test documents represented as features are given to predictive model that then predicts the labels for the test documents.

4.3 Basic Steps of Supervised Learning

We discussed the steps in the design of a learning system in the previous module. Similarly supervised learning also follows these steps:

- **Data collection** – The first step in supervised learning is the collection of training data for which we know the correct outcome provided by a teacher or oracle. For example: images for which we know the object category.

- **Representation** – The next step is the choice of how to represent the data- Format and values of each feature.
- **Modeling** - The next step is the choice of a hypothesis class - a set of possible explanations for the connection between examples and classes. This is our model of the problem – and this type of approach is specific to supervised learning.
- **Estimation** – Now we need to find best hypothesis we can in the chosen class.
- **Model Selection** - We may also reconsider the class of hypotheses given the outcome.

Remember that each of these steps can make or break the learning outcome**.**

4.4 Examples of Supervised Learning

Here we discuss some examples of supervised learning which help to understand the formulation of the learning problem. The examples shown below illustrate different aspects of learning. In all the examples x represents the input vector and f is the label. While Example 4.1 is a type of recommendation system, Example 4.2 is a tagging problem. On the other hand Example 4.3 is a recognition problem which can be cast as a classification or labeling problem. Irrespective of the different aspects, the key issue is generalization, since mere memorizing of the training set would result in over-fitting and loss of generalization. In the next section we will discuss in detail the different issues associated with generalization.

Example 4.1

Disease diagnosis

- x: Properties of patient (symptoms, lab tests)
- f : Disease / recommended therapy

Example 4.2

Part-of-Speech (POS) tagging

- x: An English sentence (e.g., The can will rust)
- f : The POS of a word in the sentence

.

Example 4.3

Face recognition

- x: Bitmap picture of person’s face
- f : Name of the person / a property

4.5 Generalization

The ability to produce correct outputs on previously unseen examples is possible because of **generalization.** The most crucial part of learning theory is : how to get good generalization with a limited number of examples. As already discussed the idea is to favor simpler classifiers. Simpler decision boundary may not fit ideally to the training data but tends to generalize better to new data.

In supervised learning, given examples (x,f(x)) of some unknown function f, the task here is to find a good approximation of 'f', where 'x' is a representation of the input example. The process of mapping a domain element into a representation is called Feature Extraction.

4.6 Phases of Supervised Learning

The first step of supervised learning is the division of all labeled samples $\mathbf{x}_1,\mathbf{x}_2,\ldots\mathbf{x}_n$ into 2 sets, the *training* set and the *test* set.

Training Phase: The training phase is for "teaching" our machine. An important task is finding the optimal weights **w** such that the function $h(\mathbf{x}_i,\mathbf{w}) = \mathbf{y}_i$ (denoting the hypothesis) fits "as much as possible" to the *training* samples ($\mathbf{x}_i$, $\mathbf{y}_i$). These weighs are determined by the process of optimization, and is usually time consuming.

Testing Phase: In the testing phase the performance of the trained classifier $\mathbf{f}(\mathbf{x},\mathbf{w})$ is validated on the unseen labeled test samples.

4.7 Using Supervised Learning

The supervised learning process is concerned with the instance space which is defined by kind of features we are using, label space which is concerned with the kind of learning task that we are dealing with and hypothesis space which specifies the kind of model we are learning. We also need to decide the learning algorithm we are going to use which determines the way we learn the model from the given data. Finally we need a loss function/evaluation metric that can be used to measure the success of the learning task. Now let us discuss each of these aspects in detail.

4.7.1 The Instance Space X

When we apply machine learning to a task, we first *define* the instance space **X** (Figure 4.2). Instances $\mathbf{x} \in \mathbf{X}$ are defined by features. These features may be Boolean features (an example - Does this email contain the word 'money'?) or Numerical features (an example - How often does 'money' occur in this email?

What is the width/height of this bounding box?). Designing an appropriate instance space **X** is crucial for good prediction of y.

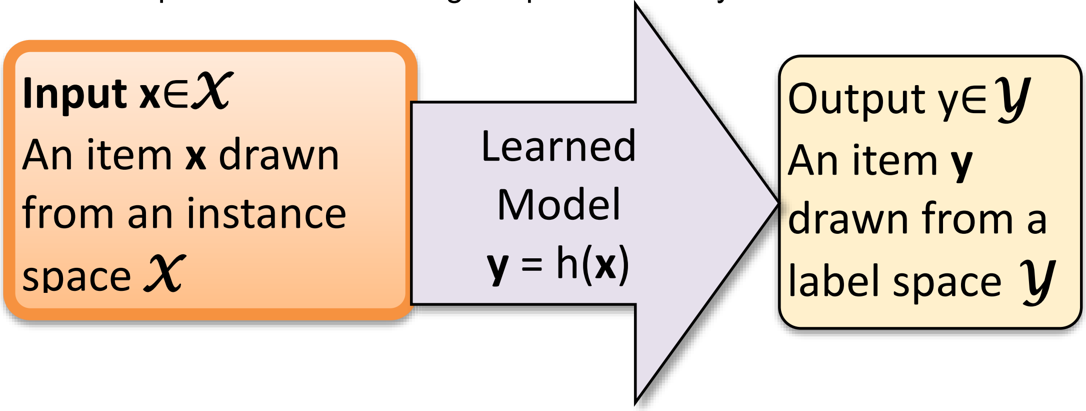
Input x∈$\mathcal{X}$
An item **x** drawn from an instance space $\mathcal{X}$
Learned Model
y = h(**x**)
Output y∈$\mathcal{Y}$
An item **y** drawn from a label space $\mathcal{Y}$

Figure 4.2 Instance and Label Space

4.7.1.1 Possible features

- **Character recognition** – histograms counting the number of black pixels along horizontal and vertical directions, number of internal holes, stroke detection etc,.
- **Speech recognition** - features for noise ratios, length of sounds, relative power, MFCC (Mel-frequency cepstral coefficients), filter matches and many others.
- **Spam detection** - presence or absence of certain email headers, the email structure, the language, the frequency of specific terms, the grammatical correctness of the text.

4.7.1.2 X as a vector space

X is an N-dimensional vector space where each dimension corresponds to one feature. Each **x** is a feature vector and can be represented as **x** = [x_1 … x_N] which can be thought of as a point in **X**. Figure 4.3 shows a 2-D feature vector X = (x_1, x_2).

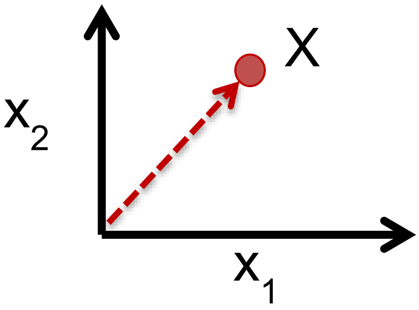

Figure 4.3 2D Feature Vector

The selection of good features is crucial for how well a task can be learned. When designing features, we often think in terms of templates, not individual features that is a set of features which together define the task. In many application areas (language, vision, etc.), a lot of work goes into designing suitable features. This requires domain expertise.

4.7.2 Label Space Y

The label space **Y** (Figure 4.2) determines **what *kind* of supervised learning task** we are dealing with. The output labels can be broadly classified as categorical, numerical or structured objects.

The output labels y∈Y are categorical and can be of the following types:

- **Binary Classification:** Given x find y in {-1, 1}. Example - An e-mail is a spam (1) or not (-1).
- **Multi-category Classification:** Given x find y in {1,2,3,.....,k} generally with the loss of {**y, f(x)**}. Example – a given object is tennis ball, football or cricket ball.

The output labels y∈Y are numerical and can be of the following types:

- **Regression:** Labels are continuous-valued and learn a linear/polynomial function f(x) – Example – Stock market prices.
- **Ranking:** Labels are ordinal and learn an ordering $f(x_1) > f(x_2)$ over input. Example – Search engine ranking

The output labels y∈Y are structured objects

- **Sequence of labels:** Given Sequence $x_1....x_l$ find $y_1....y_l$. Example **–** Given a sequence of words learning the corresponding type of POS Tags.
- **Hierarchical Categorization:** Given x find the point in the hierarchy of y (e.g. a tree)
- **Prediction: G**iven x_{t-1} ….. x_1 and y_{t-1}.... y_1 find y_t

Thus given examples of a function (X, F(X)) we are predicting a function *F(X)* for new examples *X.* This function can be defined in the following ways:

- **Discrete *F(X)*:** Classification -Classification (discrete labels) – Predict e-mail spam or not
- **Continuous *F(X)*:** Regression - Regression (real values) – Predict tomorrow's temperature
- ***F(X)* = Probability(*X*):** Probability estimation.

Thus an important step in supervised learning is choosing the output Label Space and depending on the type of output and the way the function is defined will determine the type of supervised learning that will be used.

4.7.3 Model g(x)

An important component of learning is the model used to map between instance space and label space (Figure 4.2). We need to choose what *kind* of model we want to learn.

4.7.4 Hypothesis Space

ID	X1	X2	X3	X4	Y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

$y = f(x1, x2, x3, x4)$

Figure 4.4 Hypothesis Space

The hypothesis space is defined as the space of all hypotheses that can, in principle, be the output by a particular learning algorithm. That is we learn function from the given samples. For the example shown in Figure 4.4 we have 2^4 =16 as the total sample space. There are $|\mathbf{Y}|^{|\mathbf{X}|}$ possible functions f(**x**) from the instance space **X** to the label space **Y** **i.e** 2^{16} = 65536 possible functions. From this set of functions we need to find the appropriate hypothesis space.

4.7.4.1 Hypothesis Space Reduction

We need to generalize from the few training examples given and identify the function. Since the possible hypothesis space is large we apply prior knowledge or guess to get a small reduced hypothesis space H. The hypothesis space representation can be in the form of simple conjunctive rules, *m*-of-*n* rules, linear functions or multivariate Gaussian joint probability distributions.

4.7.5 Learning Models

There are many types of learning models that can be used for learning. Some of them are :

Linear classifier (numerical functions) where in finding the functions for classifying, we are going to focus on dividing these points with a straight line. This is called linear classification.

Parametric (Probabilistic functions): A parametric model is one that can be parametrized by a finite number of parameters. It reduces the problem of estimating a probability density function (pdf), discriminant, or regression function to estimating the values of a small number of parameters. Parametric estimation: all data instances affect the final global estimate. Examples of such models include Naïve Bayes and Hidden Markov models (HMM).

Non-parametric (Instance-based functions): In this case, assumption is that similar inputs have similar outputs, and functions (pdf, discriminant, regression) change smoothly where similar data instances in the training data are found and their outputs are interpolated/ averaged. Examples include *K*-nearest neighbors, Kernel regression, Kernel density estimation, Local regression

Non-metric (Symbolic functions): In this case, the learner must search the concept space to find the desired concept. The complexity of this concept space is a primary measure of the difficulty of a learning problem. Examples include Classification and regression tree (CART), ID3, etc..

Aggregation: Bagging and Boosting are two examples of this technique. Bagging (bootstrap + aggregation) is the combining of many unstable predictors to produce an ensemble (stable) predictor. Each predictor in ensemble is created by taking a bootstrap sample of the data. The bootstrap sample of N instances is obtained by drawing N example at random, with replacement. Boosting is the combining of many weak predictors (e.g. tree stumps or 1-R predictors) to produce an ensemble predictor. Each predictor is created by using a biased sample of the training data where Instances (training examples) with high error are weighted higher than those with lower error and hence difficult instances get more attention. Adaboost is a typical algorithm under this category.

4.8 Classification – an Example of Supervised Learning

The task can be defined as the assigning of an object/event to one of a given finite set of categories.

Example: **medical diagnosis** – given a set of categories (illness types) and a laboratory data (blood test, MRI, Scan etc.) , assign a patient to the appropriate category (illness)

Other examples include credit card applications or transactions, spam filtering in email, Recommended articles in a newspaper, classify documents classes

Naïve Bayes Classifier is a common type of classifier. Let us consider the example of document classification. The steps involved are:

- Train the program (Building a Model) using a training set with a category for e.g. sports, cricket, news. Classifier Model will compute probability for each word, the probability that it makes a document belong to each of considered categories
- Test with a test data set against this Model

4.9 Classification Target Values

In supervised approaches such as classification, we will have access to both the data point, **x,** and a target value, **t.** The goal is to first Identify a function *h*, s.t. *h*(**x**) = **t.** Now we identify which of the *N* classes a data point, **x**, belongs to. **x** is a column vector of features (Figure 4.5).

$$\vec{x} = \begin{pmatrix} x_0 \\ x_1 \\ \dots \\ x_{n-1} \end{pmatrix} \qquad \vec{x} = \begin{pmatrix} f_0(x) \\ f_1(x) \\ \dots \\ f_{m-1}(x) \end{pmatrix}$$

Figure 4.5 Column Vector of Features

4.9.1 Single feature Classification

It is possible that the classification can be based on a single feature. Example is the classification of fish based only on the single feature - “Length alone” (Figure 4.6). Salmon shorter than Sea bass and the classification based on length (Threshold value L=5) alone gives an error rate of 17/50 = 34%. For threshold value L =9 error rate is still 20%. Similarly classification of fish can be based on another feature - “Lightness” (Figure 4.7). Salmon lighter than Sea bass and the classification based on lightness (Threshold value L=3) alone gives an error rate of 4/50 = 8%. It will be efficient if both the features are used.

Length → Fish ↓	2	4	8	10	12	14
Sea Bass	0	1	3	8	10	5
Salmon	2	5	10	5	1	0

Figure 4.6 Classification based on Length

Lightness → Fish ↓	1	2	3	4	5
Sea Bass	0	1	2	10	12
Salmon	6	10	6	1	0

Figure 4.7 Classification based on Lightness

4.9.2 Linear Classifier

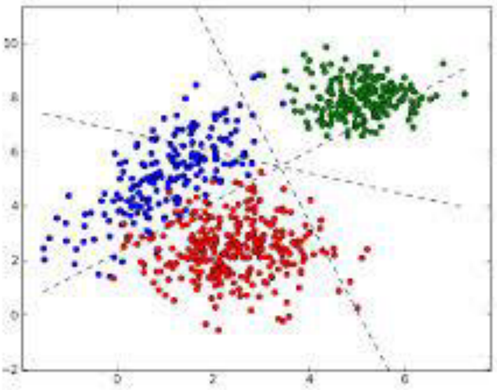

Figure 4.8 Linear Classifier

On the other hand classification can be based on a linear function as shown in Figure 4.8, where the classification is decided by the linear line.

4.10 Under-fitting to Over-fitting

Now a model can under-fit, just fit, or over-fit the data (Figure 4.9). Under-fitting is the case where the simple decision boundary *under-fit* data, i.e. chosen model is not expressive enough. There is no way to fit a linear decision boundary for the data given, so that the training examples are well separated. In this case the model learning error is too high and hence test error is also high.

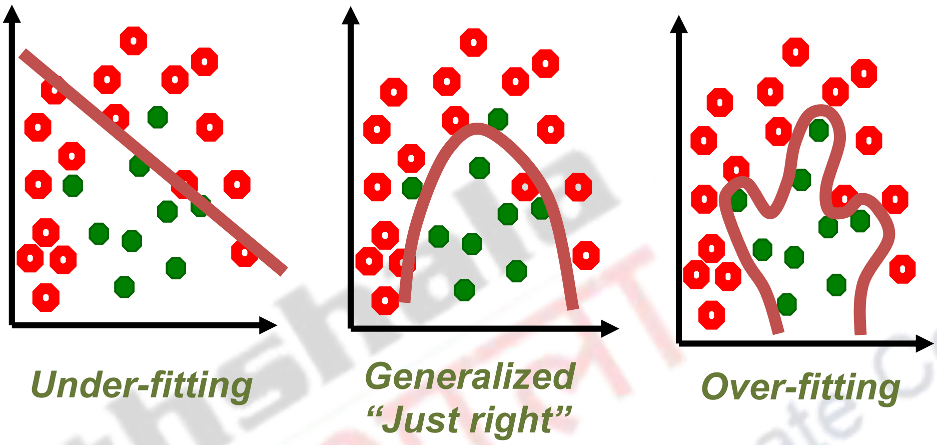

Figure 4.9 Under and Over Fitting

However complicated boundaries can ***over-fit*** the data, they are too tuned to the particular training data at hand. Therefore complicated boundaries tend to not ***generalize*** well to the new data. There will be zero error on stored data and 50% error on test (new) data. We want to find a model that is generalized and at the same time fits the data.

4.11 Regression

Regression is a again a supervised machine learning task where a statistical process is used for estimating the relationships among variables where the target value **t** is continuous, and the process gives a measure of the relation between the mean value of one variable and corresponding values of input variables . One example of regression is Logistic regression (binary regression)

- http://en.wikipedia.org/wiki/Logistic_regression

Examples of Regression as a classification problem include the following:

- Voltage ⇒ Temperature
- Processes, memory ⇒ Power consumed
- Protein structure ⇒ Energy
- Robot arm controls ⇒ Torque at effector

In the example shown in Figure 4.10, y = (2*x)+80 is the best fit for the points given. Given a new point x=50 we can find that y=180 from the regression equation. Regression equation can also be a polynomial which is called Non-linear Regression.

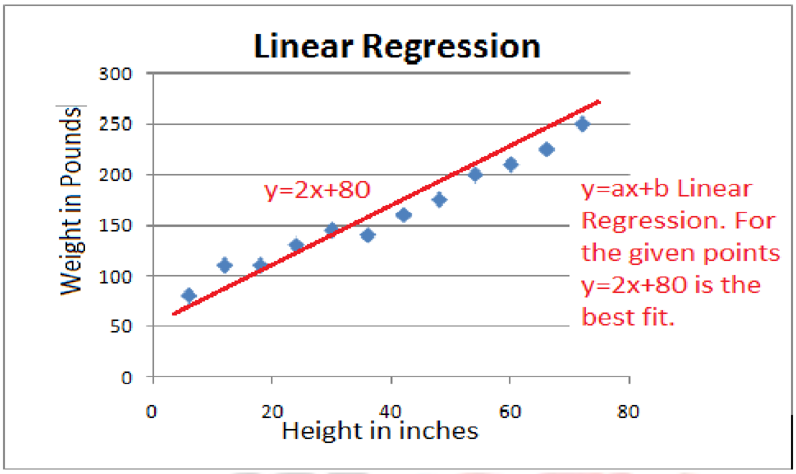

Figure 4.10 Linear Regression

4.12 Popular Frameworks/Tools

There are many tools available for machine learning. Some of them are listed here.

Weka

- Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

  Website: http://www.cs.waikato.ac.nz/~ml/**weka**/index.html

Carrot2

- Collaborative Agent-based Routing and Retrieval of Text and the Carrot2 application suit contains: – Carrot 2 Document Clustering Workbench – Carrot 2 Document Clustering Server – Carrot 2 Web Application • Carrot2 Java API • Integrate carrot code with own software.

  On-line demo: http://www.carrot2.org

Some of the natural language based tools include Gate, OpenNLP, LingPipe, Stanford NLP and Mallet. Gensim are two tools used for Topic Modelling. Apache Mahout is a machine learning tool used in the Big Data Scenario. The links of these tools are given below:

- Gate Tool - https://**gate**.ac.uk/
- OpenNLP - https://**opennlp**.apache.org/
- LingPipe - alias-i.com/**lingpipe**/
- Stanford NLP **- nlp**.**stanford**.edu/software/
- **Mallet - mallet**.cs.umass.edu/
- Gensim - https://radimrehurek.com/**gensim**/
- **Mahout - mahout**.**apache**.org/

Summary

- Discussed about what is Supervised Machine Learning.
- Discussed the steps involved in the Supervised Learning process.
- Explained the different types of supervised learning such as classification and Regression.