Machine Learning

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

April 16, 2019

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 to some tasks T and performance P if its performance P improves with experience E on tasks T.

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.

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

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 $O = \{1, 3, 5\}$).

Set Theory Notations

$$A \subset B \Leftrightarrow x \in A \Rightarrow x \in B$$

$$A = B \Leftrightarrow A \subset B \text{ and } B \subset A$$

$$A \cup B = \{x : x \in A \text{ or } x \in B\}$$

$$A \cap B = \{x : x \in A \text{ and } x \in B\}$$

$$A^c = \{x : x \notin A\}$$

Properties of Set Operations

Commutativity

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

Associativity

$$A \cup (B \cup C) = (A \cup B) \cup C$$
$$A \cap (B \cap C) = (A \cap B) \cap C$$

Distributivity

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$

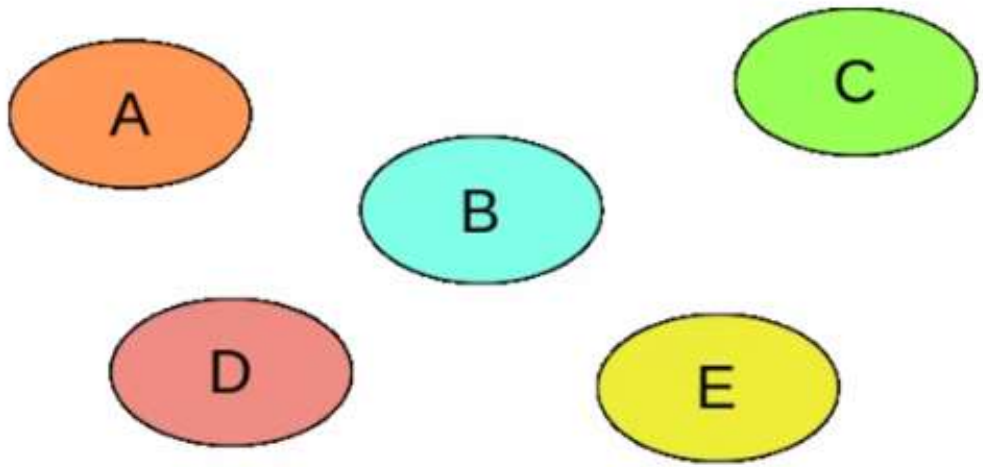
DeMorgan's Laws

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c$$

Disjoint Events

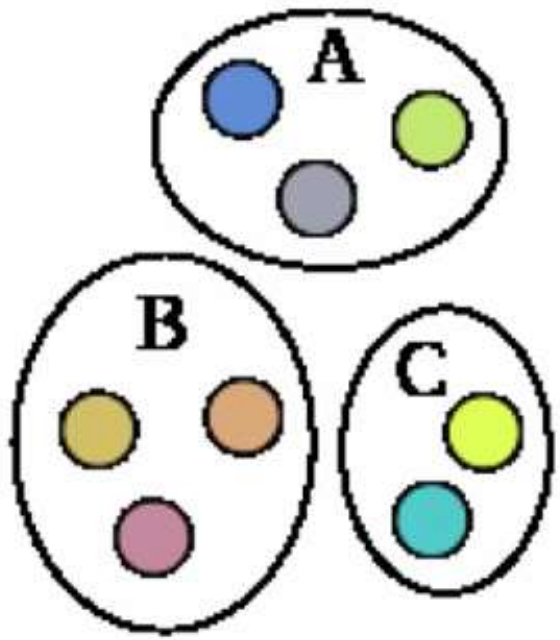
Two events A and B are disjoint (or mutually exclusive) if $A \cap B = \phi$.

A sequence of events $A_1, A_2, A_3, \ldots$ are pair-wise disjoint if $A_i \cap A_j = \phi$ for all $i \neq j$.



Partition

If $A_1, A_2, \ldots$ are pair-wise disjoint and $\cup_{i=1}^{\infty} A_i = \Omega$, then the collection $A_1, A_2, \ldots$ forms a partition of Ω.



Sigma Algebra

Given a sample space Ω, a σ-algebra is a collection $\mathcal{F}$ of subsets of Ω, with the following properties:

(a) $\Phi \in \mathcal{F}$.
(b) If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
(c) If $A_i \in \mathcal{F}$ for every $i \in \mathbb{N}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

A set A that belongs to $\mathcal{F}$ is called an $\mathcal{F}$-measurable set (event).

Example: Consider $\Omega = \{1, 2, 3\}$.
$\mathcal{F}_1 = \{\phi, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\}$.
$\mathcal{F}_2 = \{\phi, \{1, 2, 3\}\}$.

Sample Space Size Considerations

For any Ω (countable or uncountable) 2^{Ω} is always a σ-algebra.

For example, for $\Omega = \{H, T\}$, a feasible σ-algebra is the power set, i.e., $\mathcal{F} = \{\phi, \{H\}, \{T\}, \{H, T\}\}$.

However, if Ω is uncountable, then probabilities cannot be assigned to every subset of 2^{Ω}.

Probability Measure & Probability Space

A probability measure $\mathcal{P}$ on $(\Omega, \mathcal{F})$ is a function $\mathcal{P} : \mathcal{F} \to [0, 1]$ satisfying

(a) $\mathcal{P}(\phi) = 0, \quad \mathcal{P}(\Omega) = 1;$
(b) if $A_1, A_2, \ldots$ is a collection of pair-wise disjoint members of $\mathcal{F}$, then

$$\mathcal{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathcal{P}(A_i)$$

The triple $(\Omega, \mathcal{F}, \mathcal{P})$, comprising a set Ω, a σ-algebra $\mathcal{F}$ of subsets of Ω, and a probability measure $\mathcal{P}$ on $(\Omega, \mathcal{F})$, is called a **probability space**.

Example

Consider a simple experiment of rolling an ordinary die in which we want to identify whether the outcome results in a prime number or not.

$\Omega = \{1, 2, 3, 4, 5, 6\}$

$\mathcal{F} = \{\phi, \{1, 4, 6\}, \{2, 3, 5\}, \{1, 2, 3, 4, 5, 6\}\}$

$\mathcal{P} : \mathcal{F} \rightarrow [0, 1]$

- $\mathcal{P}(\phi) = 0$
- $\mathcal{P}(\{1, 4, 6\}) = 0.5$
- $\mathcal{P}(\{2, 3, 5\}) = 0.5$
- $\mathcal{P}(\Omega) = 1$

Bonferroni's Inequality

$$P(A \cap B) \geq P(A) + P(B) - 1$$

General form:

$$P(\cap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$$

Gives a lower bound on the intersection probability which is useful when this probability is hard to calculate.

Only useful if the probabilities of individual events are sufficiently large.

Bonferroni's Inequality

$$P(A \cap B) \geq P(A) + P(B) - 1$$

General form:

$$P(\cap_{i=1}^{n} A_i) \geq \sum_{i=1}^{n} P(A_i) - (n-1)$$

Gives a lower bound on the intersection probability which is useful when this probability is hard to calculate.

Only useful if the probabilities of individual events are sufficiently large.

Conditional Probability

Given two events A and B, if $P(B) > 0$, then the conditional probability that A occurs given that B occurs is defined to be

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Essentially, since event B has occurred, it becomes the new sample space.

Conditional probabilities are useful when reasoning in the sense that once we have observed some event, our beliefs or predictions of related events can be updated/improved.

Example

Q. A fair coin is tossed twice. What is the probability that both tosses result in heads given that at least one of the tosses resulted in a heads?

Example

Q. A fair coin is tossed twice. What is the probability that both tosses result in heads given that at least one of the tosses resulted in a heads?

Sol. $\Omega = \{HH, TT, HT, TH\}$
$\mathcal{P}(HH) = \mathcal{P}(TT) = \mathcal{P}(HT) = \mathcal{P}(TH) = 1/4$

$$\begin{aligned}
&\mathcal{P}(HH|\text{at least one toss heads}) \\
&\quad = \mathcal{P}(HH|HT \cup TH \cup HH) \\
&\quad = \frac{\mathcal{P}(HH \cap (HT \cup TH \cup HH))}{\mathcal{P}(HT \cup TH \cup HH)} \\
&\quad = \frac{\mathcal{P}(HH)}{\mathcal{P}(HT \cup TH \cup HH)} \\
&\quad = \frac{1}{3}
\end{aligned}$$

Bayes' Rule

We have:

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(A \cap B)}{\mathcal{P}(B)}$$

$$\mathcal{P}(A \cap B) = \mathcal{P}(A|B)\mathcal{P}(B)$$

$$\mathcal{P}(A \cap B) = \mathcal{P}(B|A)\mathcal{P}(A)$$

$$\mathcal{P}(A|B)\mathcal{P}(B) = \mathcal{P}(B|A)\mathcal{P}(A)$$

$$\mathcal{P}(A|B) = \frac{\mathcal{P}(B|A)\mathcal{P}(A)}{\mathcal{P}(B)} \text{ (Bayes' Rule)}$$

Bayes' Rule

Let $A_1, A_2, \ldots$ be a partition of the sample space, and let B be any subset of the sample space. Then, for each i = 1, 2, ...,

$$\mathcal{P}(A_i|B) = \frac{\mathcal{P}(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} \mathcal{P}(B|A_j)\mathcal{P}(A_j)}$$

Bayes' rule is important in that it allows us to compute the conditional probability $\mathcal{P}(A|B)$ from the "inverse" conditional probability $\mathcal{P}(B|A)$.

Example

Q. To answer a multiple choice question, a student may either know the answer or may guess it. Assume that with probability p the student knows the answer to a question, and with probability q, the student guesses the right answer to a question she does not know. What is the probability that for a question the student answers correctly, she actually knew the answer to the question?

Example

Q. To answer a multiple choice question, a student may either know the answer or may guess it. Assume that with probability p the student knows the answer to a question, and with probability q, the student guesses the right answer to a question she does not know. What is the probability that for a question the student answers correctly, she actually knew the answer to the question?

Sol. Let K be the event that the student knows the question, and C be the event that the student answers the question correctly. We have $\mathcal{P}(K) = p$, $\mathcal{P}(\neg K) = 1 - p$, $\mathcal{P}(C|K) = 1$, $\mathcal{P}(C|\neg K) = q$

$$\begin{aligned}
&\mathcal{P}(K|C) \\
&\quad = \frac{\mathcal{P}(C|K)\mathcal{P}(K)}{\mathcal{P}(C)} \\
&\quad = \frac{\mathcal{P}(C|K)\mathcal{P}(K)}{\mathcal{P}(K)\mathcal{P}(C|K)+\mathcal{P}(\neg K)\mathcal{P}(C|\neg K)} \\
&\quad = \frac{p}{p+q(1-p)}
\end{aligned}$$

Independent Events

Two events, A and B, are said to be independent if

$$\mathcal{P}(A \cap B) = \mathcal{P}(A)\mathcal{P}(B)$$

More generally, a family $A_i : i \in I$ is called independent if

$$\mathcal{P}(\cap_{i \in J} A_i) = \prod_{i \in J} \mathcal{P}(A_i)$$

for all finite subsets J of I.

From the above, it should be clear that pair-wise independence does not imply independence.

Conditional Independence

Let A, B, and C be three events with $\mathcal{P}(C) > 0$. The events A and B are called conditionally independent *given* C if

$$\mathcal{P}(A \cap B|C) = \mathcal{P}(A|C)\mathcal{P}(B|C)$$

or equivalently

$$\mathcal{P}(A|B \cap C) = \mathcal{P}(A|C)$$

Example: Assume that admission into the M.Tech. programme at IITM & IITB is based solely on candidate's GATE score. Then

Random Variable

A random variable is a function $X : \Omega \to \mathbb{R}$, i.e., it is a function from the sample space to the real numbers.

Examples:

- The sum of outcomes on rolling 3 dice.
- The number of heads observed when tossing a fair coin 3 times.

Induced Probability Function

Let $\Omega = \{\omega_1, \omega_2, ...\}$ be a sample space and $\mathcal{P}$ be a probability measure (function).

Let X be a random variable with range $\mathcal{X} = \{x_1, x_2, ..., x_m\}$.

We define the induced probability function $\mathcal{P}_X$ on $\mathcal{X}$ as

$$\mathcal{P}_X(X = x_i) = \mathcal{P}(\{\omega_j \in \Omega : X(\omega_j) = x_i\})$$

Induced Probability Function

Consider the previous example experiment of tossing a fair coin 3 times. Let X be the number of heads obtained in the three tosses. Enumerating the elementary outcomes, we observe the value of X as

ω	HHH	HHT	HTH	THH	TTH	THT	HTT	TTT
$X(\omega)$	3	2	2	2	1	1	1	0

Instead of using the probability measure defined on the elementary outcomes or events, we would ideally like to measure the probability of the random variable taking on values in its range.

x	0	1	2	3
$\mathcal{P}_X(X = x)$	1/8	3/8	3/8	1/8

Cumulative Distribution Function

The cumulative distribution function or cdf of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) = \mathcal{P}_X(X \leq x), \text{ for all } x$$

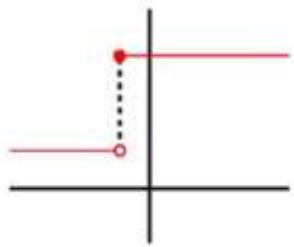
Example:

x	$(-\infty, 0]$	$(-\infty, 1]$	$(-\infty, 2]$	$(-\infty, 3]$	$(-\infty, \infty)$
$F_X(x)$	$1/8$	$1/2$	$7/8$	1	1

Properties of cdf

A function $F_X(x)$ is a cdf iff the following three conditions hold:

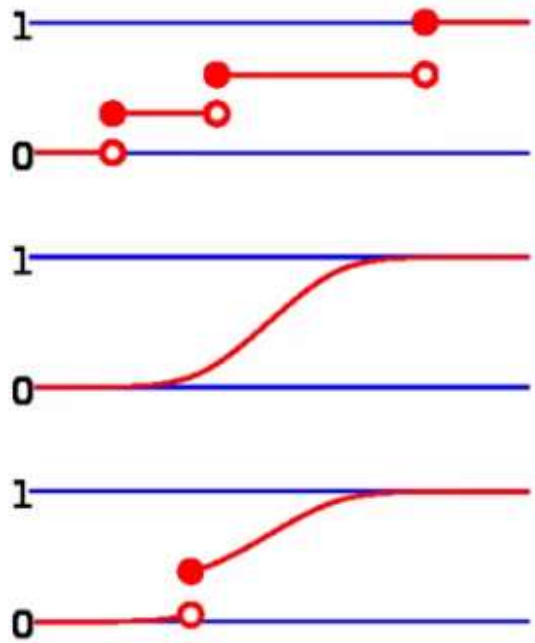
- (Monotonicity) If $x \leq y$, then $F_X(x) \leq F_X(y)$
- (Limiting values) $\lim_{x \to -\infty} F_X(x) = 0$ and $lim_{x \to \infty} F_X(x) = 1$
- (Right-continuity) For every x, we have $\lim_{y \downarrow x} F_X(y) = F_X(x)$



Continuous & Discrete Random Variables

A random variable X is continuous if $F_X(x)$ is a continuous function of x.

A random variable X is discrete if $F_X(x)$ is a step function of x.



Probability Mass Function

The probability mass function or pmf of a discrete random variable X is given by

$$f_X(x) = \mathcal{P}(X = x), \text{ for all x}$$

Example: For a geometric random variable X with parameter p,

$$f_X(x) = \begin{cases} (1-p)^{x-1}p & \text{for } x = 1, 2, \ldots \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Properties:

- $f_X(x) \geq 0$, for all x
- $\sum_x f_X(x) = 1$

Probability Density Function

The probability density function or pdf of a continuous random variable is the function $f_X(x)$ which satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt, \text{ for all x}$$

Properties:

- $f_X(x) \geq 0$, for all x
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$

Expectation

The expected value or mean of a random variable X, denoted by $E[X]$, is given by

- $$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx \text{ (continuous RV)}$$

$$E[X] = \sum_{x:\mathcal{P}(x)>0} x f_X(x) = \sum_{x:\mathcal{P}(x)>0} x\mathcal{P}(X = x) \text{ (discrete RV)}$$

Example

Q. Let the random variable X take values -2, -1, 1, 3 with probabilities 1/4, 1/8, 1/4, 3/8 respectively. What is the expectation of the random variable $Y = X^2$?

Sol. The random variable Y takes on the values 1, 4, 9 with probabilities 3/8, 1/4, 3/8 respectively.
Hence,

$$E(Y) = \sum_{y} y\mathcal{P}(Y = y) = 1 \cdot \frac{3}{8} + 4 \cdot \frac{1}{4} + 9 \cdot \frac{3}{8} = \frac{19}{4}$$

Alternatively,

$$E(Y) = E(X^2) = \sum_{x} x^2\mathcal{P}(X = x) = 4 \cdot \frac{1}{4} + 1 \cdot \frac{1}{8} + 1 \cdot \frac{1}{4} + 9 \cdot \frac{3}{8} = \frac{19}{4}$$

Properties of Expectations

Let X be a random variable and let a, b, c be constants. Then, for functions $g_1(X)$ and $g_2(X)$ whose expectations exist

- $E(ag_1(X) + bg_2(X) + c) = aEg_1(X) + bEg_2(X) + c$
- If $g_1(X) \geq 0$ for all x, then $Eg_1(X) \geq 0$
- If $g_1(X) \geq g_2(X)$ for all x, then $Eg_1(X) \geq Eg_2(X)$
- If $a \leq g_1(X) \leq b$, for all x, then $a \leq Eg_1(X) \leq b$

ANN

- Feed forward
- Feed backward/recurrent : for dynamic
- Associate Network
- Fully connected
- partially connected
- single layer
- multi layer
- Neuron: are the processing or computation unit/elements
- First layer is input layer
- Last layer is output layer
- Perceptron: error e1=d1-o1 , d1=desired output o1=actual output
  w2=w1+n*x1*e1, w2 new weight, w1 previous weight,
  n=learning rate,
  -algorithm for supervised adaptive learning of binary classifiers
  -data pair chosen randomly from data set
  -gradually error rate is reduces to 0 iteratively.

Perceptron

Multi layer Perceptron(MLP)

- Activation Function : Looks like threshold as in perceptron but it is varies smoothly and differentiable. For example sigmoid function or hyperbolic (for classification or pattern recognition task) or Linear (for regression problems).
- $H(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$
- $S(x) = \dfrac{1}{1 + e^{-x}}$
- Error Function $E(t, 0) = \dfrac{1}{2} \sum_i^n (t_i - o_i)^2$

Step in MLP

- Data Preparation
  - Data Selection
  - Data Pre-process
    - Formatting: DBMS to excel or text file, 2D to 1D
    - Cleaning: Removal of duplicate or unbound or fixing of missing data
    - Sampling: take small sample of data to reduce time complexity
    - Normalization: not essential but beneficial
  - Data Transform
    - Scaling :Dollar, rupees, weights etc.
    - Decomposition: split of complex into smaller. split date into three
    - Aggregation: opposite of decomposition , multiple entry of deposits in to acc
- Training , Testing and Validation
- Generalization and Over-fitting

Support Vector Machine

- Supervised learning with learning algo
- Maximize predictive accuracy and automatically avoids over fitting.
- Not possible for unsupervised data
- Used for classification and regression problems
- Binary Linear Classifier
- Non Linear Classifier with Kernel function
- Training data set is divided into two category
- Try to maximize the distance between two category
- If you want to use for unlabelled or semi labelled data set 1_{st} use some algo for clustering such as support vector clustering and the apply SVM
- Linear classifier $wx = b = 0$
- $(w.x_i + b) > +1$ if $y_i = +1$
- $(w.x_i + b) < -1$ if $y_i = -1$
- $y_i(w.x_i + b) \geq +1$ or $y_i(w.x_i + b) - 1 \geq 0$
- Distance from H1 to H is $\frac{|-1-b|}{||w||}$


ज्योतिर्वृणीत तमसो विजानन

Thank you

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