MACHINE LEARNING

- Machine learning is a subfield of artificial intelligence (AI).
- The goal of machine learning generally is to understand the structure of data and fit that data into models that can be understood and utilized by people.
- Machine learning is programming computers to optimize a performance criterion using example data or past experience.

Traditional Programming

Machine Learning

- Kinds of Machine Learning
    - ▷ Supervised Learning
        - ○ Classification
        - ○ Regression
    - ▷ Unsupervised Learning
    - ▷ Reinforcement Learning

Supervised Learning

- ▷ A majority of practical machine learning uses supervised learning.
- ▷ In supervised learning, the system tries to learn from the previous examples that are given.
- ▷ Train the machine using data which is well labeled
- ▷ A machine is provided with new set of examples(data)
- ▷ supervised learning algorithm analyses the the training data(set of training examples) and produces a correct outcome from labeled data.

- Supervised learning classified into two categories of algorithms:
  - **Classification:** A classification problem is when the output variable is a category, such as “Red” or “blue” and “disease” or “no disease”.
  - **Regression:** A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

Classification

- the problem of identifying to which of a set of categories a new observation belongs, based on a training set of data containing observations (or instances) whose category membership is known.

Score1	29	22	10	31	17	33	32	20
Score2	43	29	47	55	18	54	40	41
Result	Pass	Fail	Fail	Pass	Fail	Pass	Pass	Pass

- If we have some new data, say “Score1 = 25” and “Score2 = 36”, what value should be assigned to “Result” corresponding to the new data?

- There are several machine-learning algorithms for classification. The following are some of the well-known algorithms.
  - Logistic regression
  - Naive Bayes algorithm
  - k-NN algorithm
  - Decision tree algorithm
  - Support vector machine algorithm
  - Random forest algorithm

Regression

- the problem of predicting the value of a numeric variable based on observed values of the variable.

Price (US$)	Age (years)	Distance (KM)	Weight (pounds)
13500	23	46986	1165
13750	23	72937	1165
13950	24	41711	1165
14950	26	48000	1165
13750	30	38500	1170
12950	32	61000	1170
16900	27	94612	1245
18600	30	75889	1245
21500	27	19700	1185
12950	23	71138	1105

Suppose we are required to estimate the price of a car aged 25 years with distance 53240 KM and weight 1200 pounds.

- Let x denote the set of input variables and y the output variable.
- In machine learning, the general approach to regression is to assume a model, that is, some mathematical relation between x and y, involving some parameters say, θ, in the following form:

$$y = f(x, \theta)$$

For example, if the input variables are “Age”, “Distance” and “Weight” and the output variable is “Price”, the model may be

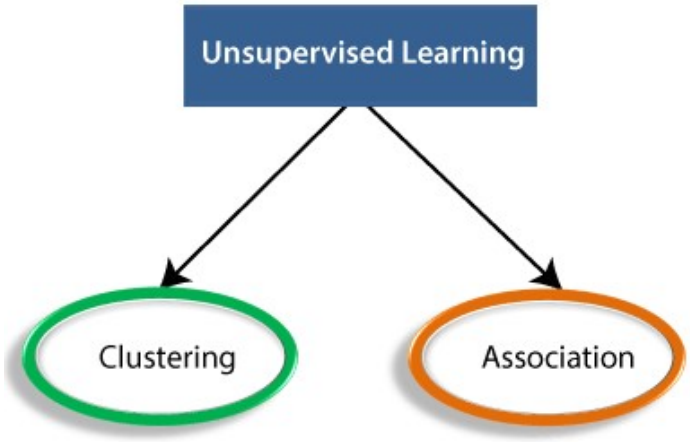
$$y = f(x,\theta)$$

$$\text{Price} = a_0 + a_1 \times (\text{Age}) + a_2 \times (\text{Distance}) + a_3 \times (\text{Weight})$$

where x = (Age, Distance, Weight) denotes the set of input variables and $\theta = (a_0, a_1, a_2, a_3)$ denotes the set of parameters of the model.

Unsupervised Learning

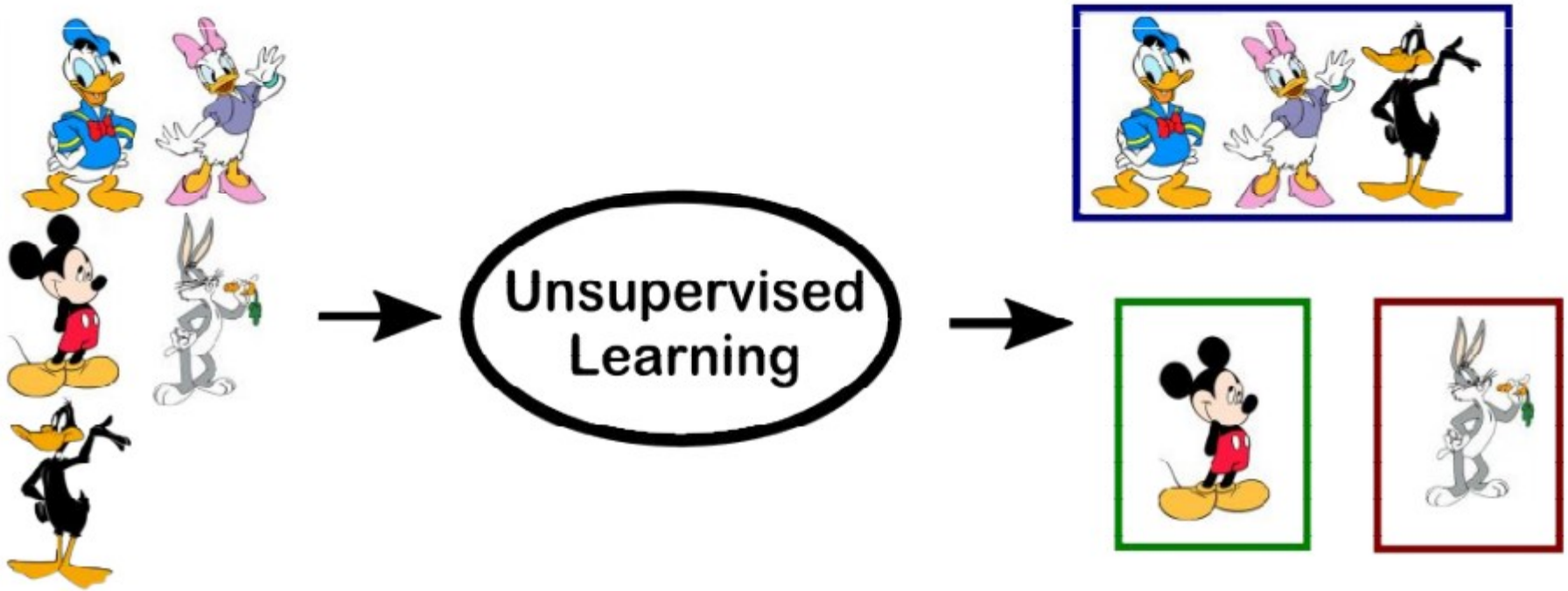
- Training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance.
- Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data.
- Used in clustering (task of grouping a set of objects in such a way that objects in the same group are similar to each other than to those in other groups.)



Below is the list of some popular unsupervised learning algorithms:

- **K-means clustering**
- **KNN (k-nearest neighbors)**
- **Hierarchal clustering**
- **Anomaly detection**
- **Neural Networks**
- **Principle Component Analysis**
- **Independent Component Analysis**
- **Apriori algorithm**
- **Singular value decomposition**

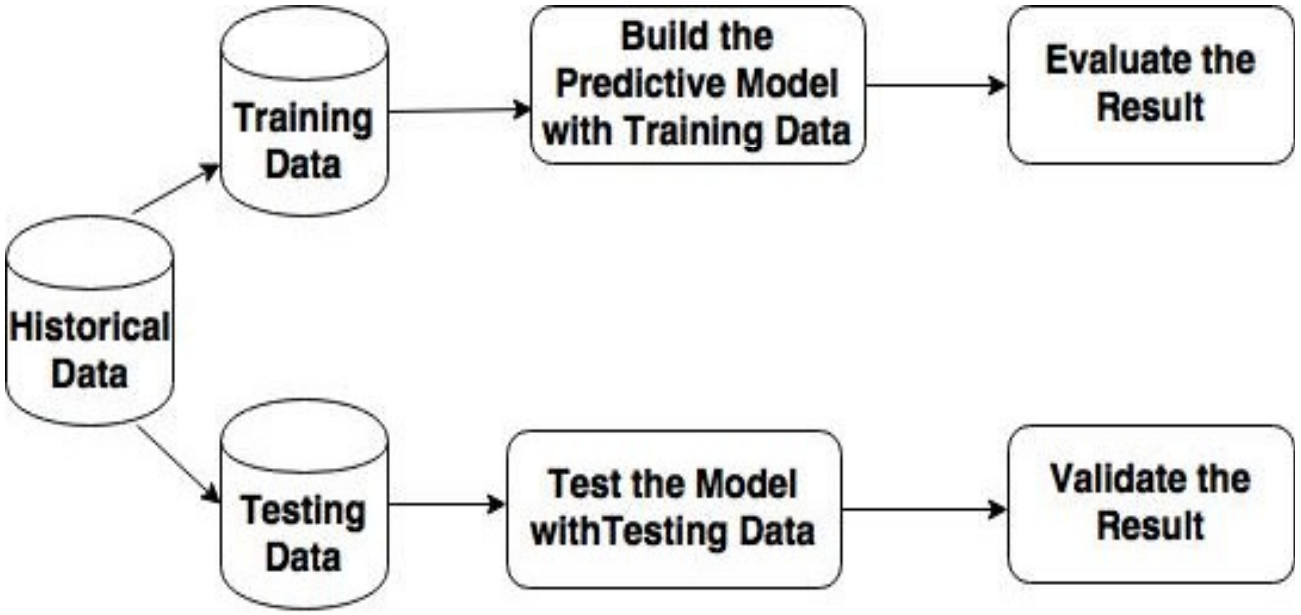
Unsupervised Learning



Reinforcement Learning

- a type of Machine Learning algorithms which allows software agents and machines to
  - automatically determine the ideal behavior within a specific context, to maximize its performance.
- A reinforcement learning algorithm, or agent, learns by interacting with its environment.
- The agent receives rewards by performing correctly and penalties for performing incorrectly.
- The agent learns (without intervention from a human) by maximizing its reward and minimizing its penalty

Steps in machine learning life cycle



Regression Analysis

- Generally, regression analysis is used to determine the relationship between the dependent and independent variables of the dataset.
- Regression analysis helps to understand how dependent variables change when one of the independent variables changes and other independent variables are kept constant.
- This helps in building a regression model and further, helps in forecasting the values with respect to a change in one of the independent variables.

- **Types of Regression Analysis**
  - **Simple Linear Regression**
  - **Multiple Linear Regression**
  - **Logistic Regression**

Simple Linear regression

- One of the most frequent used techniques in statistics is linear regression where we investigate the potential relationship between a variable of interest (often called the response variable) and a set of one of more variables (known as the independent variables).
- $y = \beta 0 + \beta 1x$

- Linear regression is performed with the lm() function.
- lm(y~x)
- The model above is achieved by using the lm() function in R and the output is called using the summary()function on the model.

Sample code is in **slr-example1.r**

slr-example1.r program result analysis

```
> print(relation)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
   -38.4551       0.6746
```

➢**Simple Regression model**:

Parameters – y as height and x as weight
The performance are evaluated with the following aspects

$y=\beta_0+\beta_1 x$

Evaluate regression coefficients $\beta_{0=}$ -38.4551 and β_1= 0.6746

Hypothesis test H0: β1 = 0
Versus
Ha: β1 ≠ 0

```
> summary(relation)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-6.3002 -1.6629  0.0412  1.8944  3.9775

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -38.45509    8.04901  -4.778  0.00139 **
x             0.67461    0.05191  12.997 1.16e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.253 on 8 degrees of freedom
Multiple R-squared:  0.9548,    Adjusted R-squared:  0.9491
F-statistic: 168.9 on 1 and 8 DF,  p-value: 1.164e-06
```

- Hence, if we see a small p-value then we can infer that there is an association between the predictor and the response. We reject the null hypothesis—that is, we declare a relationship to exist between X and Y —if the p-value is small enough.

- The **residual standard error** is used to measure how well a regression model fits a dataset.
- The smaller the residual standard error, the better a regression model fits a dataset. Conversely, the higher the residual standard error, the worse a regression model fits a dataset.
- A regression model that has a small residual standard error will have data points that are closely packed around the fitted regression line:

Here is how to interpret the rest of the model summary:

Pr(>|t|): This is the p-value associated with the model coefficients. Since the p-value for *x* (1.16e-06) is significantly less than .05, we can say that there is a statistically significant association between *x* and *y*.

Multiple R-squared: This number tells us the percentage of the variation in the y can be explained by the x. In general, the larger the R-squared value of a regression model the better the explanatory variables are able to predict the value of the response variable. In this case, **95.5%** of the variation in x can be explained by y.

- **Residual standard error:** This is the average distance that the observed values fall from the regression line. The lower this value, the more closely a regression line is able to match the observed data. In this case, the average observed y falls **3.253** points away from the y predicted by the regression line.
- **F-statistic & p-value:** The F-statistic (**168.9**) and the corresponding p-value (1.16e-06) tell us the overall significance of the regression model, i.e. whether explanatory variables in the model are useful for explaining the variation in the response variable. Since the p-value in this example is less than .05, our model is statistically significant and *x* is deemed to be useful for explaining the variation in *y*.

- Sample code is in **slr-example2.r**

Multiple Linear Regression

- read.csv("filename.csv",header=TRUE, sep=",")
- read.csv("filename.csv")
- Then use the functions
  - dim() :- find the dimension as nxp (number of rows x number of attributes)
  - head():- display first 6 rows
  - tail():- display last 6 rows
- After this exploring the structure of each attribute using the function str().

Sample code is in read_csv_apply.r

- Here use the data set car_data.csv.
- The explanation about the attributes of this data set is given below

[, 1]	mpg	Miles/(US) gallon
[, 2]	cyl	Number of cylinders
[, 3]	disp	Displacement (cu.in.)
[, 4]	hp	Gross horsepower
[, 5]	drat	Rear axle ratio
[, 6]	wt	Weight (1000 lbs)
[, 7]	qsec	1/4 mile time
[, 8]	vs	V/S
[, 9]	am	Transmission (0 = automatic, 1 = manual)
[,10]	gear	Number of forward gears
[,11]	carb	Number of carburetors

- Here the attributes are numeric (non-integer type) ,integer type , character type etc.
- The integer type attributes may be qualitative one (categorical) or discrete-quantitative.
- The numeric type data are continuous or discrete quantitative.

- The summary() function displays several common summary statistics
- Measuring the mean and median of our data provides one way to quickly summarize the values.
- But these measures of center tell us little about whether or not there is diversity in the measurements.

Exploring categorical variables

- In contrast to numeric data, categorical data is examined using table() function rather than summary statistics
- In this data set the attribute cyl is categorical.
- That is qualitative.
- We can find the frequency of this attribute with table() function.

- Another method is the uses of apply()
  - apply(X, MARGIN, FUN)

    Example :-Return the sum of each of the columns of the matrix m

    apply(m,2,sum)

We can compare the mean and median of each attribute from the apply() function result.

Correlation

- In statistical terms we use correlation to denote association between two quantitative variables.
- We also assume that the association is linear, that one variable increases or decreases a fixed amount for a unit increase or decrease in the other.
- Correlation is calculated using the function cor()

The performance evaluation of the regression models

- **Simple Regression model**:

  - Parameters – mpg and disp
  - The performance are evaluated with the following aspects

    - $mpg=\beta_0+\beta_1 disp$
    - Evaluate regression coefficients $\beta_{0=}$ 29.599 and β_1= -0.041
    - Hypothesis test H0: $\beta 1 = 0$
      Versus
      Ha: $\beta 1 \neq 0$
    - p-value is < 2e-16
    - Residual Standard Error (RSE) is 3.251
    - R^2 statistic is 0.7183

```
> coefficients(L1)
(Intercept)          x1
29.59985476 -0.04121512
> summary(L1)

Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8922 -2.2022 -0.9631  1.6272  7.2305

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 29.599855   1.229720  24.070  < 2e-16 ***
x1          -0.041215   0.004712  -8.747 9.38e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.251 on 30 degrees of freedom
Multiple R-squared:  0.7183,    Adjusted R-squared:  0.709
F-statistic: 76.51 on 1 and 30 DF,  p-value: 9.38e-10
```

- Hence, if we see a small p-value then we can infer that there is an association between the predictor and the response. We reject the null hypothesis—that is, we declare a relationship to exist between X and Y —if the p-value is small enough.

- The **residual standard error** is used to measure how well a regression model fits a dataset.
- The smaller the residual standard error, the better a regression model fits a dataset. Conversely, the higher the residual standard error, the worse a regression model fits a dataset.
- A regression model that has a small residual standard error will have data points that are closely packed around the fitted regression line:

Here is how to interpret the rest of the model summary:

Pr(>|t|): This is the p-value associated with the model coefficients. Since the p-value for *disp* (9.38e-10) is significantly less than .05, we can say that there is a statistically significant association between *disp* and *mpg*.

Multiple R-squared: This number tells us the percentage of the variation in the mpg can be explained by the disp. In general, the larger the R-squared value of a regression model the better the explanatory variables are able to predict the value of the response variable. In this case, **71.8%** of the variation in mpg can be explained by disp.

- **Residual standard error:** This is the average distance that the observed values fall from the regression line. The lower this value, the more closely a regression line is able to match the observed data. In this case, the average observed mpg falls **3.251** points away from the mpg predicted by the regression line.
- **F-statistic & p-value:** The F-statistic (**76.51**) and the corresponding p-value (9.38e-10) tell us the overall significance of the regression model, i.e. whether explanatory variables in the model are useful for explaining the variation in the response variable. Since the p-value in this example is less than .05, our model is statistically significant and *disp* is deemed to be useful for explaining the variation in *mpg*.