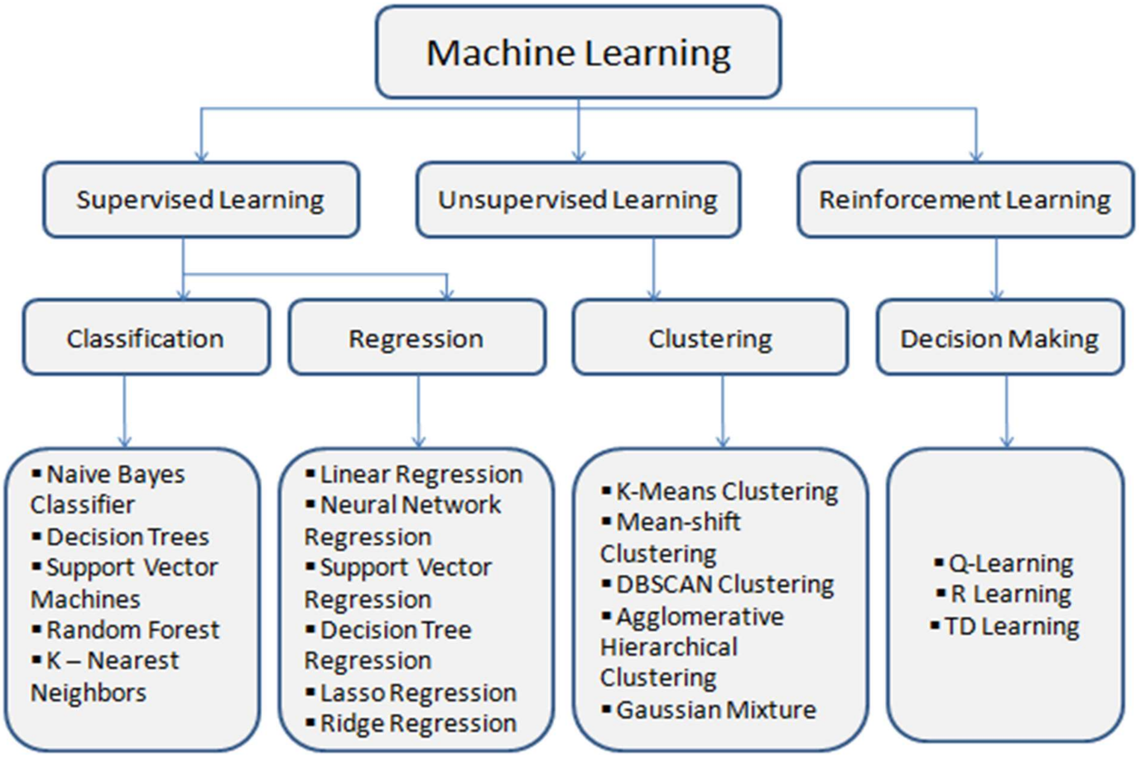
UNIT I

Introduction to Machine Learning:

If a computer program improves its performance at a certain class of tasks T, as measured by a certain performance measure P, with experience E, then it is said to learn from experience E with respect to that class of tasks T.

- Machine learning is a type of computer programming that uses example data or past experience to optimize performance.
- It involves defining a model with certain parameters and then optimizing those parameters using training data.
- Machine learning uses statistics to build models and computer science to efficiently solve optimization problems and process large amounts of data.
- It is useful when there is no human expertise, when humans are unable to explain their expertise, when the solution changes over time, or when the solution needs to be adapted to particular cases.
- The goal of machine learning is to build a model that is a good approximation of the data, as data is abundant while knowledge is scarce.
- An example of machine learning in retail would be analysing customer transactions to predict their behaviour, such as suggesting "antivirus software" to customers who buy "computers".

Different types of learning:



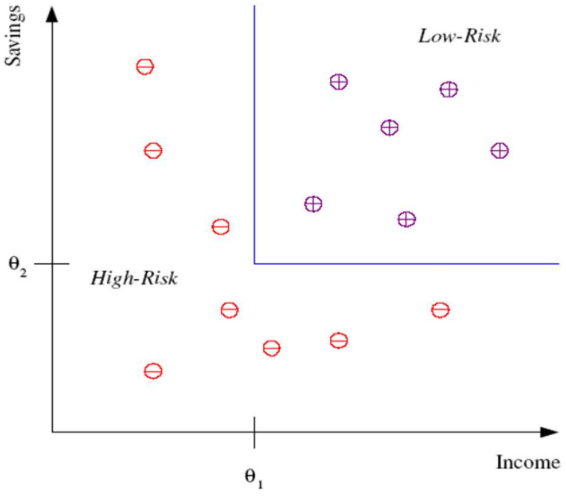
Examples of Machine Learning Applications:

Association Rules:

- In the retail industry, machine learning can be used for basket analysis to find associations between products purchased by customers
- The goal is to calculate the conditional probability of P(Y|X), where Y is the product we want to condition on and X is the product or set of products the customer has already purchased
- For example, if P(chips|beer) = 0.7, we can define a rule that 70% of customers who buy beer also buy chips
- Customer attributes such as gender, age, and marital status can be used to further customize the analysis by estimating P(Y|X, D)
- Basket analysis can be applied in various industries, such as recommending movies or suggesting books/authors to customers based on their past history

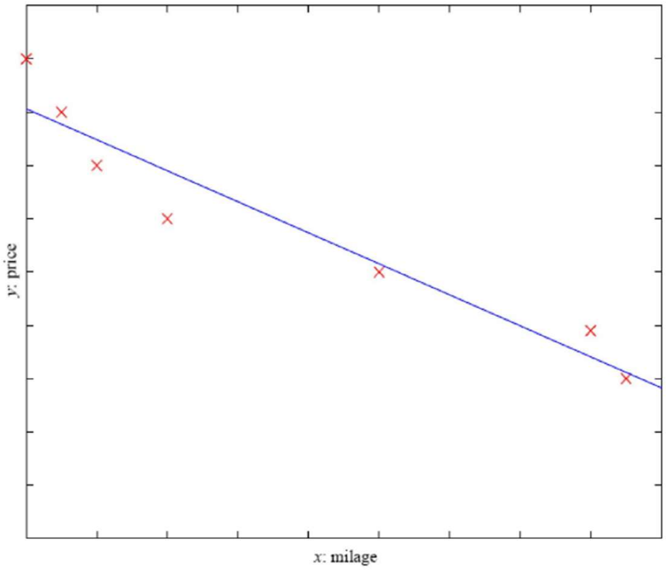
Classification:

- A credit is an amount of money loaned by a financial institution to be paid back with interest.
- The bank needs to predict the risk associated with a loan to make sure that the customer can repay it.
- Credit scoring is used by the bank to calculate the risk based on customer data such as income, savings, profession, age, and past financial history.
- The bank has a record of past loans containing customer data and loan repayment history.
- Machine learning systems fit a model to past data to calculate the risk for a new loan application and decide whether to accept or refuse it.
- This is a classification problem with two classes: low-risk and high-risk customers.
- The customer's information serves as the input to the classifier, which assigns the input to one of the two classes.
- The learned classification rule may be of the form IF income> θ1 AND savings> θ2 THEN low-risk ELSE high-risk.
- This is an example of a discriminant function that separates examples of different classes.
- Once a rule is learned, it can be used to make correct predictions for novel instances, such as a new loan application with a certain income and savings.



Regression:

- The problem of predicting the price of a used car is an example of a regression problem, where the output is a number.
- The inputs are the car attributes, such as brand, year, engine capacity, and mileage.
- The machine learning program collects a training dataset and fits a function to learn the relationship between the inputs and the output.
- The model is defined up to a set of parameters, and the program optimizes the parameters to minimize the approximation error between the predicted and correct values in the training set.
- The model can be linear or nonlinear, such as a quadratic or higher-order polynomial.
- Another example of regression is the navigation of a mobile robot, where the output is the angle by which the steering wheel should be turned.
- In some cases, the goal is to learn relative positions, such as in a recommendation system for movies where a ranking function is learned based on the user's movie preferences and attributes.



Unsupervised Learning:

- Supervised learning aims to learn a mapping from input to output with the help of a supervisor who provides correct output values for the given inputs.
- Regression and classification are two types of supervised learning problems where the output can be a number or a class label, respectively.
- Unsupervised learning aims to find regularities in the input data where there is no supervisor providing correct output values.
- Clustering is a method for unsupervised learning where the aim is to find clusters or groupings of input data based on their similarity.
- Clustering can be used for customer segmentation, document clustering, and outlier detection, among other applications.
- In document clustering, documents are represented as bag-of-words, and similar documents are grouped based on the words they share.

Reinforcement Learning:

- Reinforcement learning algorithms learn from past good action sequences to generate a policy (It is a function that maps the current state of the system to an action to be taken).
- In applications where the output is a sequence of actions, a single action is not as important as the policy of correct actions.
- Games and robot navigation are examples of applications where reinforcement learning can be used.
- Games are easy to describe, but difficult to play well, making them a good research area for reinforcement learning.
- In robot navigation, the robot must learn the correct sequence of actions to reach a goal state from an initial state, without hitting obstacles.
- Reinforcement learning can be more challenging when the system has unreliable or partial sensory information.
- Multiple agents may need to interact and cooperate to accomplish a common goal, such as a team of robots playing soccer.

Supervised Learning: Learning a Class from Examples:

Class learning is the process of finding a description that distinguishes positive examples from negative examples of a class.

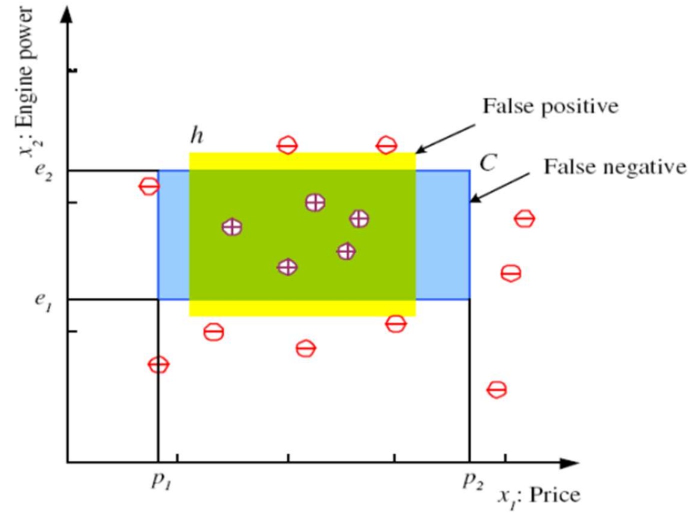
- In the example of learning the class of a "family car", positive examples are cars that people consider as family cars and negative examples are other cars.
- The goal of class learning is to be able to predict whether a new example belongs to the class or not.
- Input representation is important in class learning, and in this example, the input representation includes only two attributes: price and engine power.
- The training set consists of examples of cars labeled as positive or negative based on whether they belong to the "family car" class or not.
- Each car in a training set is represented by an ordered pair (x, r) where x represents price and engine power, and r denotes the type of car (positive or negative).
- The learning process involves finding a description that separates the positive examples from the negative examples based on the input attributes

$$r = \begin{cases} 1 \text{ if } \mathbf{x} \text{ is positive} \\ 0 \text{ if } \mathbf{x} \text{ is negative} \end{cases}$$

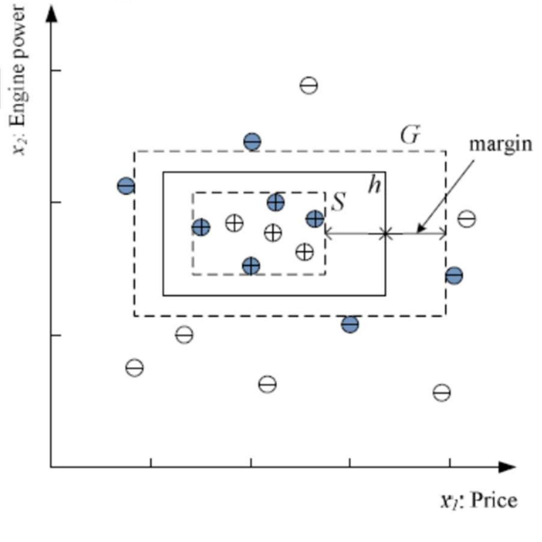
- A hypothesis class H is defined as the set of rectangles in the price-engine power space.
- The learning algorithm finds a hypothesis h from H that approximates the class C as closely as possible, using a quadruple of parameters (p1, p2, e1, and e2) to specify h.
- The aim is to find h that is as similar as possible to C.
- The empirical error is the difference between the predicted labels of the hypothesis h and the actual labels in the training set.
- The error of hypothesis h given the training set X is

$$E(h \mid \mathcal{X}) = \sum_{t=1}^{N} 1\left(h\left(\mathbf{x}^t\right) \neq r^t\right)$$

- The goal of learning is to minimize the empirical error and find a hypothesis h that generalizes well to new, unseen examples.



- The point where C is 1 but h is 0 is a false negative, and the point where C is 0 but h is 1 is a false positive. Other points namely, true positives and true negatives are correctly classified.
- The hypothesis class H in our example is the set of all possible rectangles, and we need to find the best one that includes all the positive examples and none of the negative ones.
- However, there are infinitely many hypotheses that can achieve this, and the problem is how well our hypothesis will classify future examples that are not in the training set.
- One solution is to find the most specific hypothesis, called S, is the tightest rectangle that includes all positive and none of the negative examples.
- The most general hypothesis, called G, is the largest rectangle that includes all positive and none of the negative examples.
- Any hypothesis between S and G is a valid hypothesis with no error and is consistent with the training set, making up the version space.
- Different training sets can lead to different S, G, and version space, affecting the learned hypothesis, h and choosing h halfway between S and G can increase the margin (It is the distance between the boundary and the closest instances).
- To have an error function with a minimum at h and the maximum margin, we need a hypothesis that returns a value carrying a measure of distance to the boundary



Probably Approximately Correct Learning:

The goal is to find the number of examples, N, such that with probability at least $1-\delta$, the hypothesis h has error at most ε. This is achieved through probably approximately correct (PAC) learning.

- In this case, S is the tightest possible rectangle and the error region between C and h = S is the sum of four rectangular strips.
- The probability of a positive example falling in any of these strips and causing an error should be at most $\varepsilon/4$.
- If we take at least $(4/\varepsilon) \log(4/\delta)$ independent examples from C and use the tightest rectangle as our hypothesis h, with confidence probability at least $1-\delta$, a given point will be misclassified with error probability at most ε.
- The number of examples is a slowly growing function of $1/\varepsilon$ and $1/\delta$, linear and logarithmic, respectively.

Linear regression, Multiple Linear regression, Logistic Regression:

Linear Regression:

- Linear regression models the relationship between a dependent variable (y) and one or more independent variables (x).
- The relationship is assumed to be linear, meaning the dependent variable can be modeled as a linear combination of the independent variable(s) with some added noise or error term.
- The goal of linear regression is to find the coefficients of the linear model that best fit the training data.
- This is often done using the method of least squares, where the sum of the squared differences between the predicted values and the actual values is minimized.
- Once the coefficients are found, the model can be used to make predictions on new data by plugging in the values of the independent variables.
- Linear regression is commonly used in statistics, economics, finance, engineering, and other fields.
- The formula for simple linear regression is: y = w0 + w1*x
  Where:
  - y is the dependent variable
  - x is the independent variable
  - w0 is the intercept (the point where the line intercepts the y-axis)
  - w1 is the slope(the rate at which y changes with respect to x)

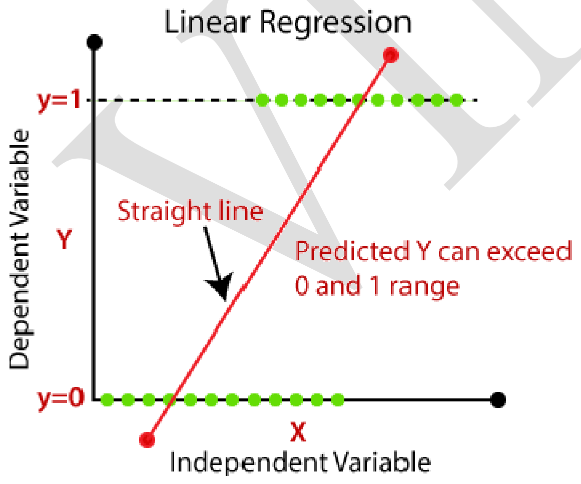
Multiple Linear Regression:

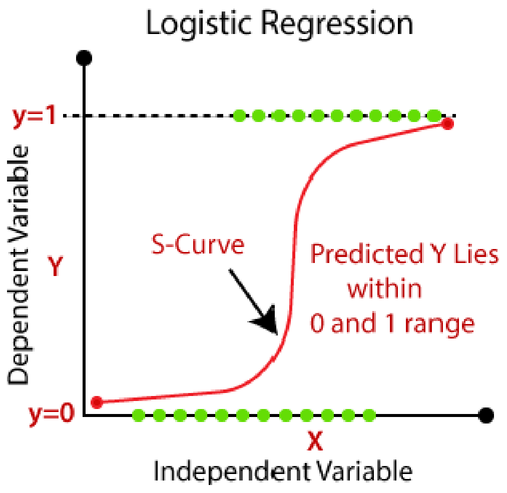
- Multiple linear regression models the relationship between a dependent variable (y) and two or more independent variables (x1, x2, ..., xn).
- The relationship is assumed to be linear, meaning the dependent variable can be modeled as a linear combination of the independent variables with some added noise or error term.
- The goal of multiple linear regression is to find the coefficients of the linear model that best fit the training data.
- This is often done using the method of least squares, where the sum of the squared differences between the predicted values and the actual values is minimized.
- Once the coefficients are found, the model can be used to make predictions on new data by plugging in the values of the independent variables.
- Multiple linear regression is commonly used in various fields including statistics, economics, finance, engineering, and social sciences.

- The formula for multiple linear regression is: $y = w0 + w1x1 + w2x2 + ... + wn*xn$
  Where:
  - y is the dependent variable
  - x1, x2, ..., xn are the independent variables
  - w0 is the intercept (the point where the line intercepts the y-axis)
  - w1, w2, ..., wn are the slopes (the rate at which y changes with respect to each independent variable)

Logistic Regression:

- Logistic regression models the relationship between a dependent variable (often binary or categorical) and one or more independent variables (often continuous or categorical).
- The relationship is modeled using the sigmoid function, which maps any real-valued number to a probability value between 0 and 1.
- The goal of logistic regression is to find the coefficients of the model that best fit the training data and allow for accurate predictions on new data.
- This is often done using maximum likelihood estimation or other optimization techniques.
- Logistic regression is commonly used in fields such as medicine, biology, psychology, and social sciences for classification problems and predicting binary outcomes.
- The formula for logistic regression is: $p = 1 / (1 + e^{-(w0+w1x1+w2x2+ ... + wnxn)})$
  Where:
  - p is the probability of the dependent variable being in a particular category
  - x1, x2, ..., xn are the independent variables
  - b0 is the intercept
  - b1, b2, ..., bn are the coefficients or slopes of the independent variables
  - e is the base of the natural logarithm, approximately equal to 2.71828.





Dimensionality reduction:

In machine learning applications, observation data is used as input for decision making.

- Feature selection or extraction is an important step to reduce the dimensionality of the data.
- Complexity of learning algorithms depends on the number of input dimensions and data sample size, reducing the dimensionality helps to reduce complexity during testing.
- Removing unnecessary features helps to save the cost of feature extraction.
- Simpler models are more robust on small datasets and have less variance.
- When data can be explained with fewer features, it allows knowledge extraction and interpretation of hidden or latent factors.
- Data represented in few dimensions without loss of information can be visualized and analysed for structure and outliers.

Feature Selection:

There are two main methods for reducing dimensionality: feature selection and feature extraction.

- In feature selection, the goal is to find a subset of k dimensions out of the original d dimensions that provide the most relevant information.
- Feature extraction, on the other hand, aims to find a new set of k dimensions that are a combination of the original d dimensions.
- Feature extraction methods can be supervised or unsupervised, depending on whether they use output information or not.
- The most widely used feature extraction methods are principal component analysis (PCA) and linear discriminant analysis (LDA), which are both linear projection methods.
- PCA is an unsupervised linear method that is similar to factor analysis and multidimensional scaling.
- LDA is a supervised linear method that is used for classification problems.
- When we have two sets of observed variables, canonical correlation analysis can be used to find the joint features that explain the dependency between them.
- Examples of nonlinear dimensionality reduction techniques include isometric feature mapping, locally linear embedding, and Laplacian eigenmaps.

Subset Selection:

In subset selection, we aim to find the best subset of features that contributes the most to accuracy and these features should have the least number of dimensions.

- There are 2^d possible subsets of d variables, which can be computationally expensive to test for larger d.
- Forward selection and backward selection are two common approaches for subset selection.
- Forward selection starts with no variables and adds them one by one until any further addition does not decrease the error.
- Backward selection starts with all variables and removes them one by one until any further removal increases the error significantly.
- The error should be checked on a validation set distinct from the training set to test the generalization accuracy.
- With more features, the training error may decrease, but not necessarily the validation error.

- In sequential forward selection, we start with no features and at each step, we select the input that causes the least error when added to the current feature set.
- The error function used depends on the application, and can be mean square error or misclassification error.
- We stop adding features when further additions do not decrease the error or if the decrease in error is too small according to a user-defined threshold.
- Adding features increases the complexity of the classifier/regressor and incurs the cost of observing the new feature.
- This algorithm is known as the wrapper approach, where feature extraction is thought to "wrap" around the learner used as a subroutine.
- Feature selection is a local search procedure and does not guarantee finding the optimal subset that causes the smallest error. It is also costly and time-consuming, especially on large datasets.
- The selected features may depend on the classifier used and the way data is split between training and validation data. Multiple random training/validation splits can be done to decide the selected features by looking at the average validation performance.
- On small datasets, it may be better to use resampling methods to select features.
- Feature selection is supervised and can be used with any regression or classification method.
- Feature selection may not be a good method for dimensionality reduction in applications like face recognition, where feature extraction methods are more effective.

Principal Component Analysis:

PCA is a dimensionality reduction technique that is widely used in machine learning and data analysis. It is used to transform a high-dimensional dataset into a lower-dimensional space while preserving the most important information in the data.

The main steps of PCA are:

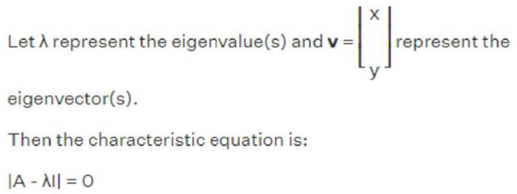
- **Standardization:** PCA assumes that the data is standardized, i.e., it has a mean of 0 and a standard deviation of 1. Therefore, the first step in PCA is to standardize the data if it is not already standardized.

$$\mathrm{Z} = \frac{X - mean(X)}{\sigma(X)}$$

- **Compute the covariance matrix:** Next, we compute the covariance matrix of the standardized data. The covariance matrix is a matrix that measures how two variables are related to each other. It is a square matrix where the diagonal elements are the variances of each variable, and the off-diagonal elements are the covariance between variables.

$$\text{Covariance of data} = \begin{bmatrix} cov(X,X) & cov(Y,X) & cov(Z,X) \\ cov(X,Y) & cov(Y,Y) & cov(Z,Y) \\ cov(X,Z) & cov(Y,Z) & cov(Z,Z) \end{bmatrix}$$

- **Compute the eigenvectors and eigenvalues:** The eigenvectors and eigenvalues of the covariance matrix provide information about the directions of maximum variance in the data. The eigenvectors are the directions in which the data has the most variance, and the eigenvalues represent the amount of variance in each of those directions.

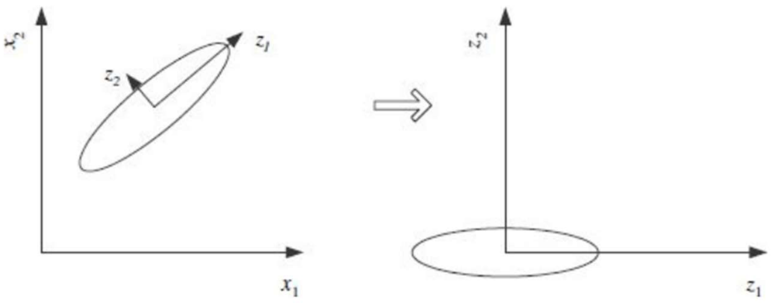


Let λ represent the eigenvalue(s) and $\mathbf{v} = \begin{bmatrix} x \\ y \end{bmatrix}$ represent the eigenvector(s).

Then the characteristic equation is:

|A - λI| = 0

- **Sort the eigenvectors by decreasing eigenvalues:** We sort the eigenvectors based on the corresponding eigenvalues in decreasing order. This is because the eigenvectors with the highest eigenvalues represent the directions in which the data has the most variance.
- **Select the top k eigenvectors:** We select the top k eigenvectors that correspond to the k largest eigenvalues, where k is the desired number of dimensions in the lower-dimensional space.
- **Construct the projection matrix:** We construct a projection matrix from the selected eigenvectors. This matrix is used to project the data onto the lower-dimensional space.
- **Project the data onto the lower-dimensional space**: Finally, we project the data onto the lower-dimensional space using the projection matrix. This gives us a new dataset with k dimensions, where k is the desired number of dimensions in the lower-dimensional space.



Linear Discriminant Analysis:

LDA (Linear Discriminant Analysis) is a supervised learning algorithm used for dimensionality reduction in classification tasks. It is similar to PCA, but it considers the class information of the data and aims to find a projection that maximizes the separation between classes.

The steps of LDA are as follows:

- **Compute the mean vectors of each class**: Calculate the mean vector for each class by averaging the feature vectors for all samples in that class.

$$\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$$

- **Compute the scatter matrices:** Within-class scatter matrix (Sw): measure of the scatter of samples within each class. Between-class scatter matrix (Sb): measure of the scatter of the class means.

$$\mathbf{S}_W = \sum_{k=1}^{K}\sum_{i\in C_k}(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T$$

$$\mathbf{S}_B = \sum_{k=1}^{K} n_k(\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^T$$

- **Compute the eigenvectors and eigenvalues of the matrix (Sw^-1)Sb:** The eigenvectors are the directions that maximize the separation between the classes. The eigenvalues correspond to the magnitude of the separation

$$S_W^{-1} S_B v = \lambda v$$

- **Sort the eigenvectors by decreasing eigenvalues:** choose the k eigenvectors with the largest eigenvalues to form a matrix W.
- **Transform the samples onto the new subspace:** Project the samples onto the new subspace using the matrix W.

$$y = W^T x$$