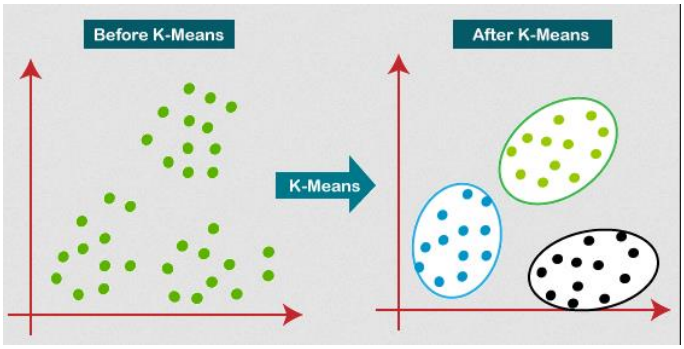
CHAPTER3

REGRESSION, CLASSIFICATION, CLUSTERING, DIMENSION REDUCTION

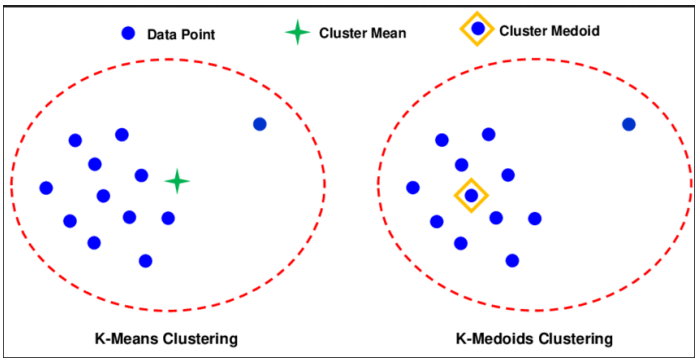
Clustering Algorithms

- **K_means:** that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster.



- **K-medoids Clustering:** clustering that clusters the data points into k clusters. Unlike K-means, which uses the **mean** of the data points as the center of the cluster, K-medoids uses **actual data points** to represent the center of the cluster (medoids).
    - Handles noise and outliers better than K-means.
    - Medoids are actual data points, making the clusters more interpretable.

    - ➢ Generally slower than K-means, especially for large datasets, due to the need to compute pairwise dissimilarities.
    - ➢ Not as scalable to very large datasets as K-means.



Medoids are representative objects of a data set or a cluster within a data set whose sum of dissimilarities to all the objects in the cluster is minimal.

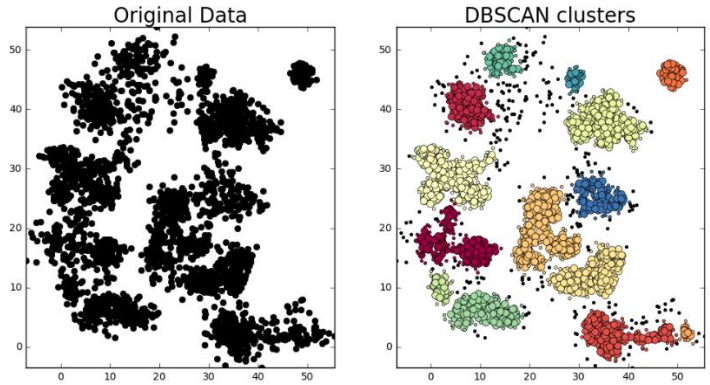
- **Hierarchical Clustering:** method of cluster analysis that seeks to build a hierarchy of clusters. Unlike partitioning methods like K-means or K-medoids, which require the number of clusters to be specified in advance, hierarchical clustering does not. It can produce a dendrogram, a tree-like diagram that records the sequences of merges or splits.

  Types of Hierarchical Clustering

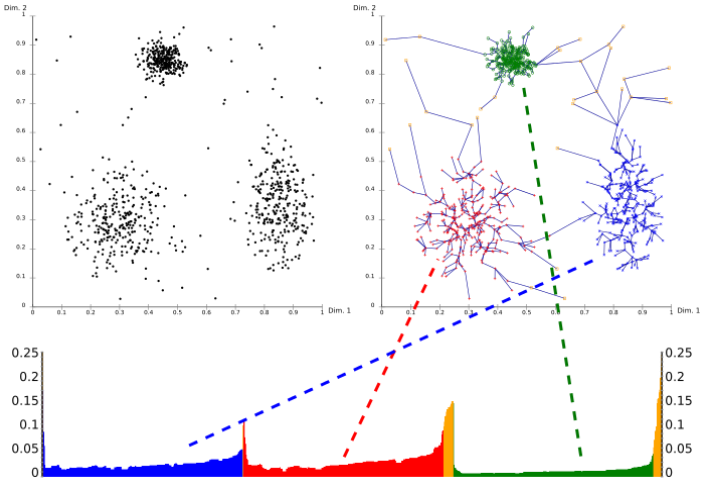
  1. **Agglomerative (Bottom-Up) Clustering:**
     - Starts with each data point as a singleton cluster.
     - Iteratively merges the closest pair of clusters until only one cluster (or the desired number of clusters) remains.
  2. **Divisive (Top-Down) Clustering:**
     - Starts with all data points in a single cluster.
     - Recursively splits clusters until each cluster contains a single data point (or the desired number of clusters).

- **Density-Based Clustering:** methods that identify distinctive clusters in the data, based on the idea that a cluster/group in a data space is a contiguous region of high point density, separated from other clusters by sparse regions.

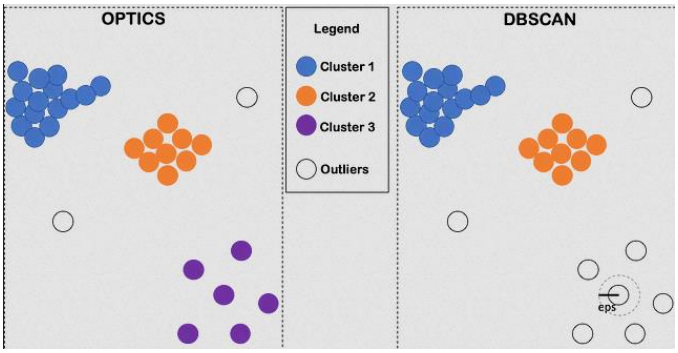
  DBSCAN: It is capable of finding clusters of arbitrary shape and can effectively handle noise in the data.



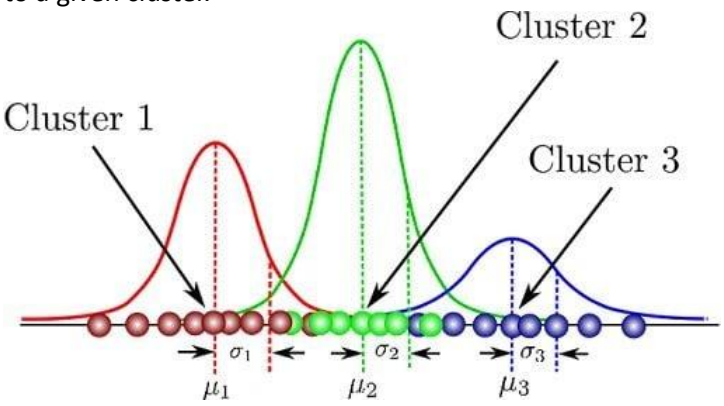
  OPTICS: similar to DBSCAN but with additional capabilities. It generates an ordering of points that reflects the density-based clustering structure of the data, which can be used to extract clusters of varying density.



- ✓ **DBSCAN** is simpler and faster, best suited for datasets with clusters of similar density and where noise identification is important.
- ✓ **OPTICS** is more flexible and powerful for datasets with clusters of varying density, providing a richer clustering structure but at the cost of increased complexity and computation time.



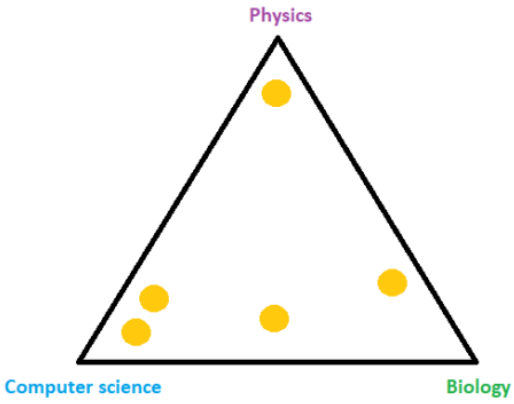
- **Model-based Clustering**
  - **Gaussian Mixture Models (GMM):** ethod used to determine the probability each data point belongs to a given cluster.



- **Latent Dirichlet Allocation (LDA):** technique that is commonly used for text analysis. It's a type of topic modeling in which words are represented as topics, and documents are represented as a collection of these word topics.

  $P(W,Z,\theta,\varphi;\alpha,\beta)=M\prod i=1P(\theta j;\alpha)K\prod i=1P(\varphi;\beta)N\prod t=1P(Zj,t|\theta j)P(Wj,t|\varphi zj,t)$

  Example:



Tasks

When is appropriate to use each algorithm?

1. K-Means Clustering

- The data is numeric and you can define a clear number of clusters (k).
- You want to partition the data into a fixed number of clusters.
- The clusters are roughly spherical and equally sized.

2. Hierarchical Clustering

- The number of clusters is not known in advance.
- You want a dendrogram to understand the data hierarchy.
- You need to find nested clusters.

3. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- The data has noise and outliers.
- You need clusters of arbitrary shapes.
- The data has varying densities.

4. OPTICS (Ordering Points to Identify the Clustering Structure)

- You need to identify clusters of varying densities.
- You want to handle noise and outliers effectively.

- DBSCAN's limitations are too restrictive for your data.

5. Gaussian Mixture Models (GMM)

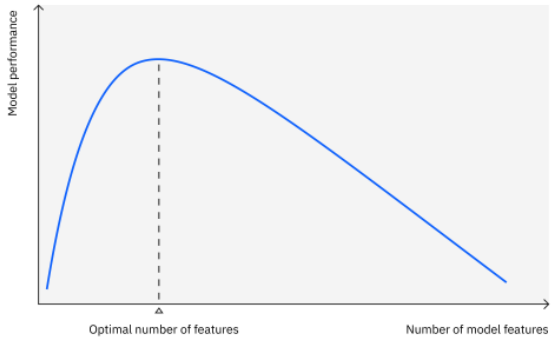
- The data is continuous and follows a Gaussian distribution.
- You need soft clustering, where a data point can belong to multiple clusters with certain probabilities.
- The clusters are elliptical rather than spherical.

6. Latent Dirichlet Allocation (LDA)

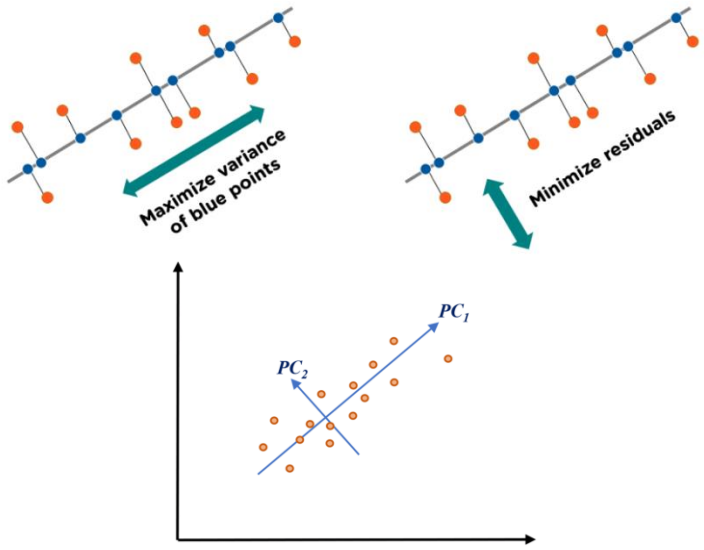
- Use LDA when you want to uncover hidden thematic structures in a large corpus of text. LDA helps in identifying the topics that best represent the content of the documents.
- LDA can be used for document classification based on the discovered topics. Once the topics are identified, documents can be classified into these topics, aiding in organization and retrieval.
- Enhance information retrieval systems by indexing documents based on topics rather than just keywords. This allows for more semantically relevant search results.

Dimension Reduction: Simplifying Complex Data

a method for representing a given dataset using a lower number of features (i.e. dimensions) while still capturing the original data's meaningful properties.



- **Principal Component Analysis (PCA):** that identifies a set of orthogonal axes, called principal components, that capture the maximum variance in the data.



Step 1: Standardization

$$z = \frac{x - \mu}{\sigma}$$

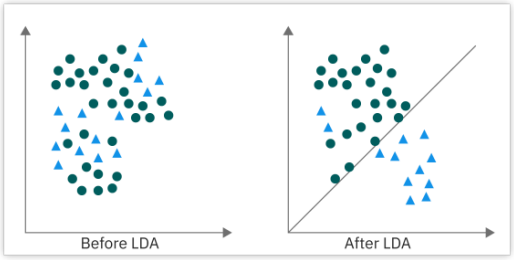
Step2: Covariance Matrix Computation

$$cov(x1, x2) = \frac{\sum_{i=1}^{n}(x1_i - \bar{x1})(x2_i - \bar{x2})}{n-1}$$

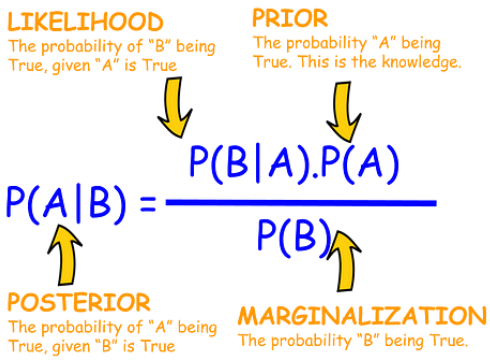
Step 3: Compute Eigenvalues and Eigenvectors of Covariance Matrix to Identify Principal Components

$$AX - \lambda X = 0$$
$$(A - \lambda I)X = 0$$

- **Linear Discriminant Analysis (LDA):** model the data distribution for each class and use Bayes' theorem to classify new data points. Bayes calculates conditional probabilities—the probability of an event given some other event has occurred.



Bayes' Theorem:



t-SNE vs PCA

Both t-SNE and PCA are dimensional reduction techniques that have different mechanisms and work best with different types of data.

PCA (Principal Component Analysis) is a **linear technique** that works best with data that has a linear structure. It seeks to identify the underlying principal components in the data by projecting onto lower dimensions, minimizing variance, and preserving large pairwise distances. Read our Principal Component Analysis (PCA) tutorial to understand the inner working of the algorithms with R examples.

But, t-SNE is a **nonlinear technique** that focuses on preserving the pairwise similarities between data points in a lower-dimensional space. t-SNE is concerned with preserving small pairwise distances whereas, PCA focuses on maintaining large pairwise distances to maximize variance.

Tasks

What is the purpose of dimension reduction in data analysis?

This can be done to reduce the complexity of a model, improve the performance of a learning algorithm, or make it easier to visualize the data.

How does the Curse of Dimensionality affect data analysis?

A. The curse of dimensionality states that **as the number of dimensions or features in a dataset increases, the volume of the data space expands exponentially**. This expansion leads to sparsity in data, making it difficult to analyze effectively.

How to choose between PCA, LDA and t-SNE?

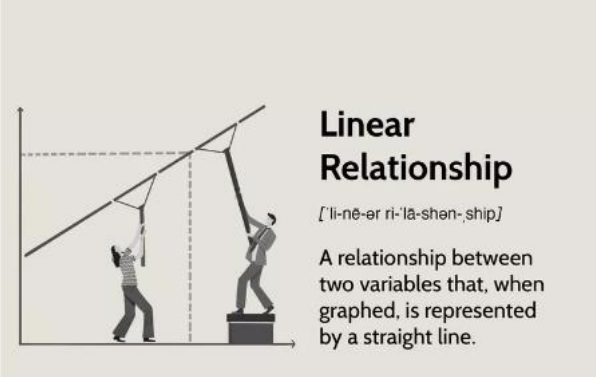
T-SNE is used for designing/implementation and can bring down any number of feature space into 2-D feature space. **Both PCA and LDA are used for visualization and dimensionality reduction but T-SNE is specifically used for visualization purposes only**. It is well suited for the visualization of high-dimensional datasets.

Regression

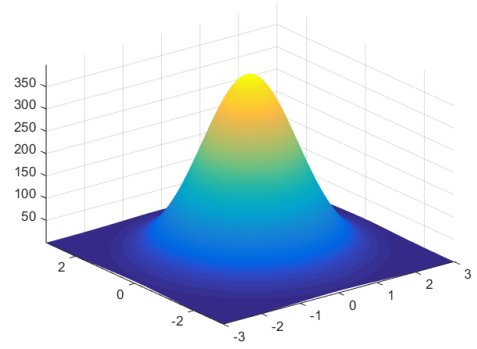
a technique for investigating the relationship between independent variables or features and a dependent variable or outcome.

Assumptions

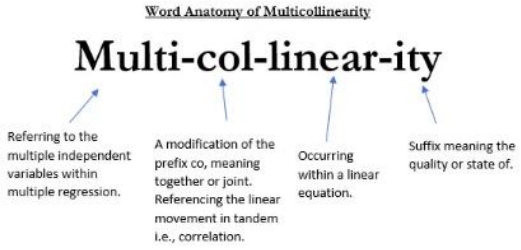
- **Linear Relationship**: The core premise of multiple linear regression is the existence of a **linear relationship between the dependent (outcome) variable and the independent variables**. This linearity can be visually inspected using scatterplots, which should reveal a straight-line relationship rather than a curvilinear one.



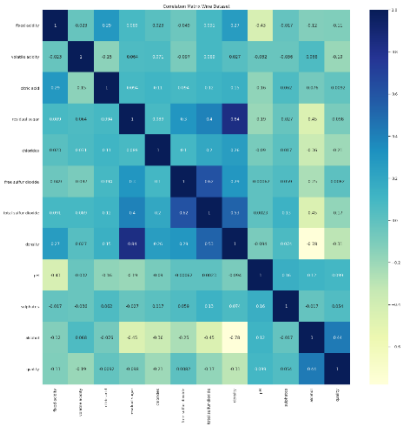
- **Multivariate Normality**: **The analysis assumes that the residuals** (the differences between observed and predicted values) **are normally distributed**. This assumption can be assessed by examining histograms or Q-Q plots of the residuals, or through statistical tests such as the Kolmogorov-Smirnov test.



- **No Multicollinearity**: It is essential that the independent variables are not too highly correlated with each other, a condition known as multicollinearity. This can be checked using:

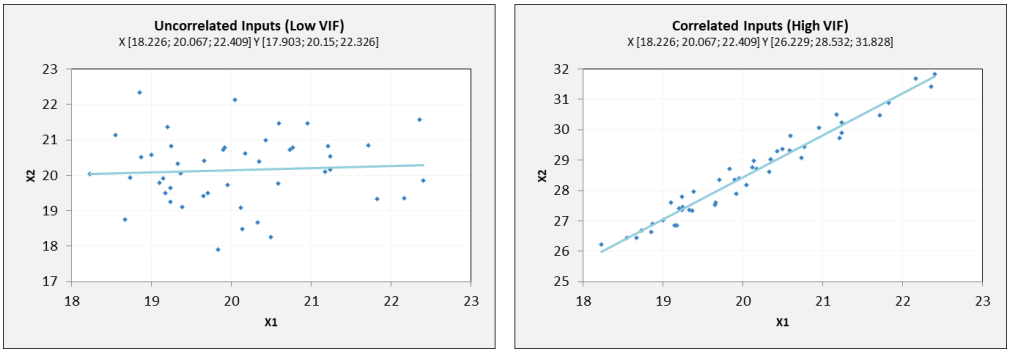


- **Correlation matrices**, where correlation coefficients should ideally be **below 0.80**.



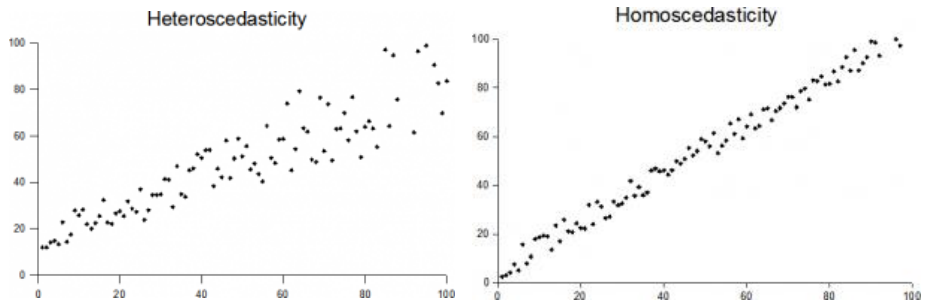
- **Variance Inflation Factor (VIF),** with VIF values above 10 indicating problematic multicollinearity. Solutions may include centering the data (subtracting the mean score from each observation) or removing the variables causing multicollinearity.

$$\widehat{\mathrm{var}}(\hat{\beta}_j) = \frac{s^2}{(n-1)\widehat{\mathrm{var}}(X_j)} \cdot \frac{1}{1-R_j^2},$$



- **Homoscedasticity**: The **variance of error terms** (residuals) should be consistent across all **levels of the independent variables**. A scatterplot of residuals versus predicted values should not display any discernible pattern, such as a cone-shaped distribution, which would indicate heteroscedasticity. Addressing heteroscedasticity might involve data transformation or adding a quadratic term to the model.

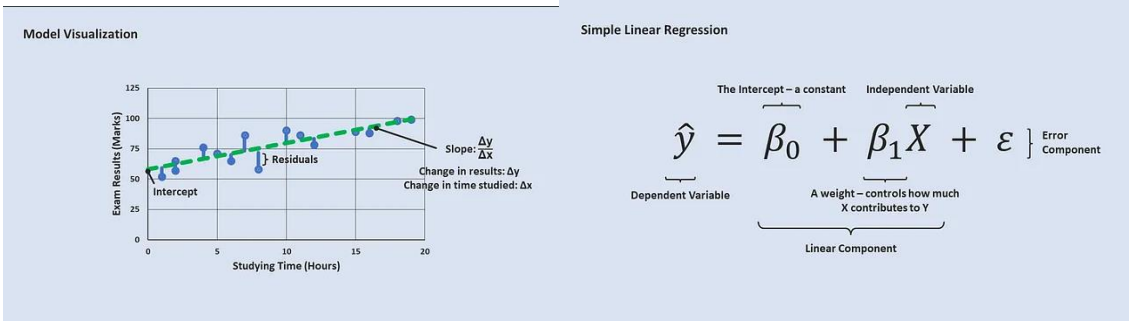
the points must be about the same distance from the line:



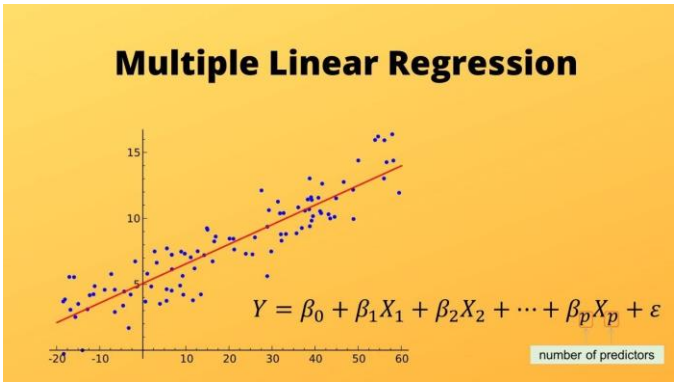
$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

What are the types of linear regression?

- **Simple Linear Regression:** Simple linear regression is used to estimate the relationship between two quantitative variables.



- **Multiple Linear Regression:** refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables.



Tasks

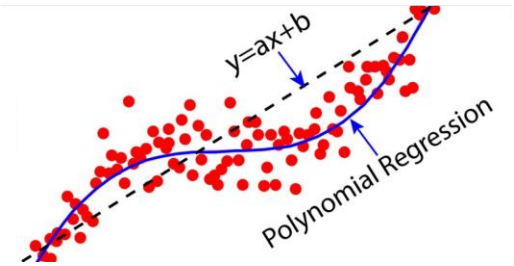
What are the other types of Regression?

1. Linear Regression

- **Simple Linear Regression**: Models the relationship between two variables by fitting a linear equation.
- **Multiple Linear Regression**: Extends simple linear regression by using multiple independent variables to predict the dependent variable.

2. Polynomial Regression

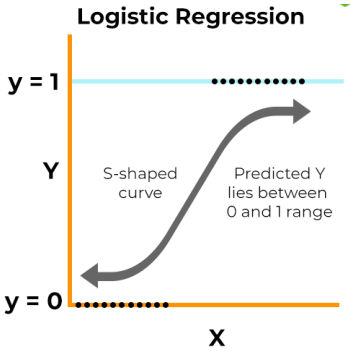
- Models the relationship between the independent variable and the dependent variable as an nth-degree polynomial.



3. Logistic Regression

- Used for binary classification problems. It models the probability that a given input point belongs to a certain class.

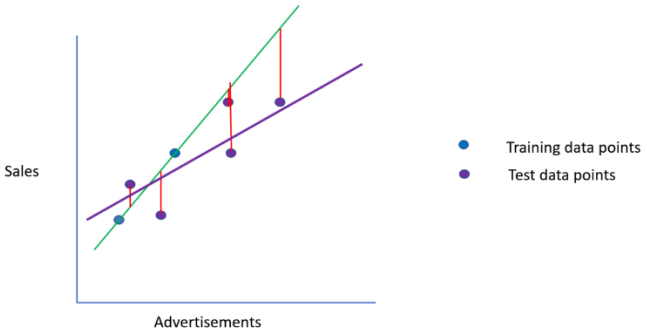
$$y = \frac{e^{(b_0 + b_1 X)}}{1 + e^{(b_0 + b_1 X)}}$$



4. Ridge Regression

- A type of linear regression that includes a regularization term to penalize large coefficients, which helps to prevent overfitting.

$$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

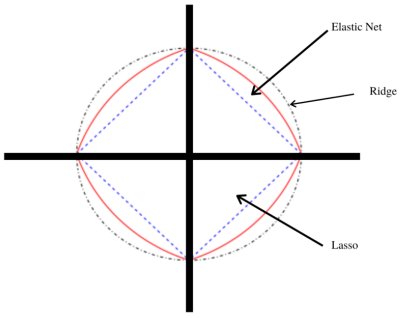


5. Lasso Regression

- Similar to ridge regression but uses L1 regularization, which can shrink some coefficients to zero, effectively performing variable selection.

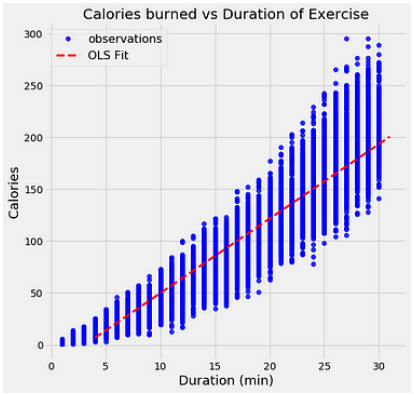
6. Elastic Net Regression

- Combines the properties of both ridge and lasso regression. It uses both L1 and L2 regularization terms.



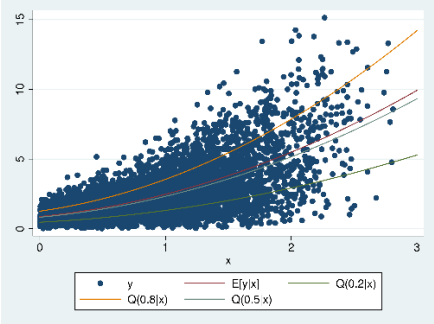
7. Bayesian Regression

- Incorporates Bayesian principles to estimate the regression coefficients by considering prior distributions.



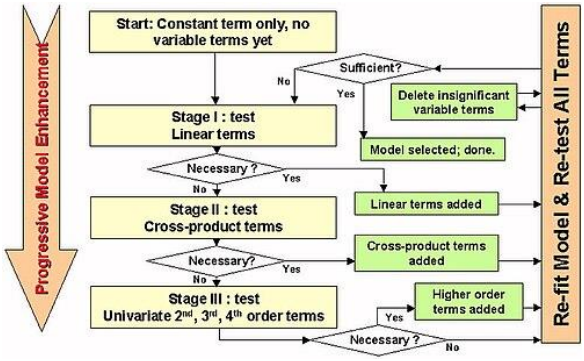
8. Quantile Regression

- Estimates the conditional quantiles of the response variable, providing a more comprehensive analysis than just the mean.



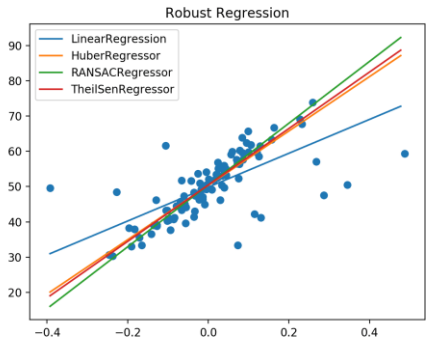
9. Stepwise Regression

- A method of fitting regression models by adding or removing predictor variables, based on their statistical significance.



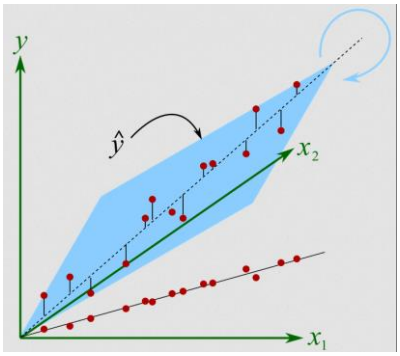
10. Robust Regression

- Techniques designed to be less sensitive to outliers in the data than ordinary least squares regression.



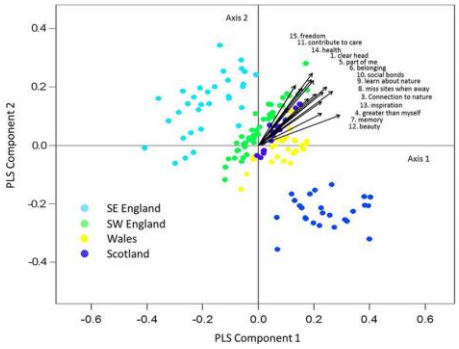
11. Principal Component Regression (PCR)

- A technique that uses principal component analysis to reduce the dimensionality of the data before performing linear regression.



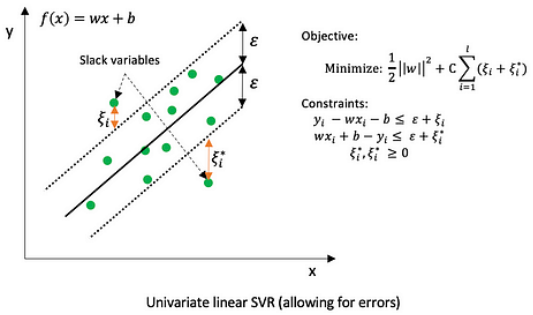
12. Partial Least Squares Regression (PLSR)

- Similar to PCR, but it also takes into account the response variable when determining the principal components.



13. Support Vector Regression (SVR)

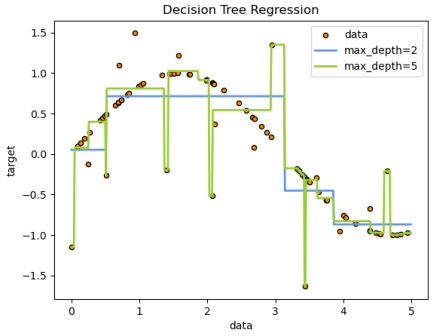
- Uses support vector machine principles for regression problems, offering flexibility in handling non-linear relationships.



Univariate linear SVR (allowing for errors)

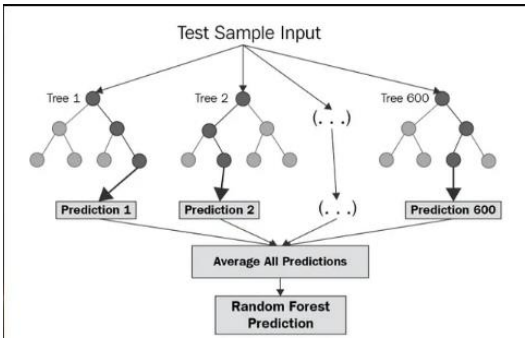
14. Decision Tree Regression

- Uses decision trees to model the relationship between the independent variables and the dependent variable.



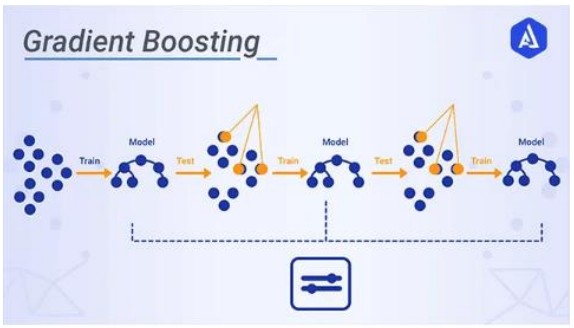
15. Random Forest Regression

- An ensemble method that uses multiple decision trees to improve the prediction accuracy and control overfitting.



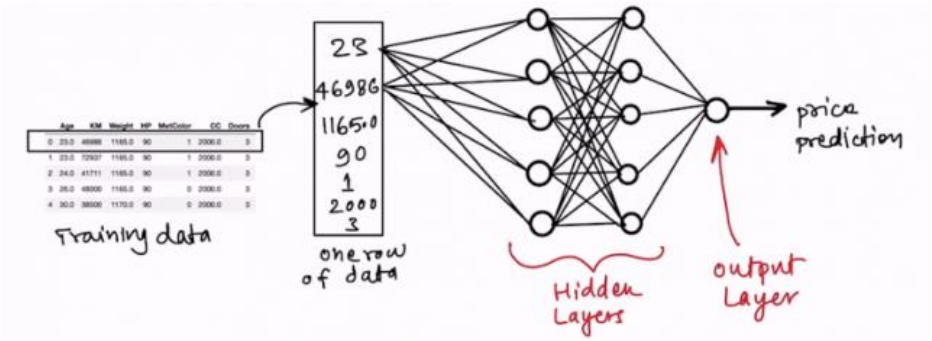
16. Gradient Boosting Regression

- An ensemble technique that builds multiple decision trees sequentially to reduce the prediction error.



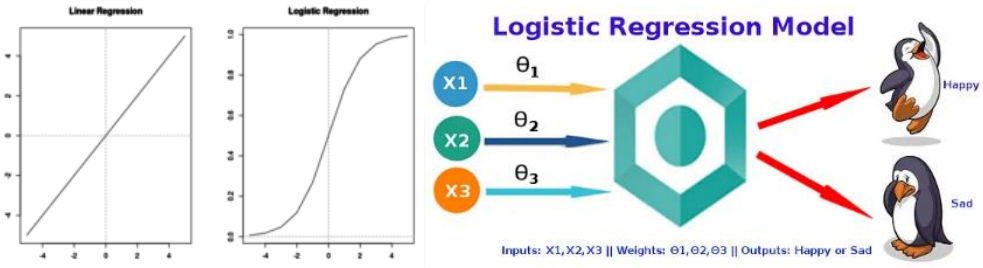
17. Neural Network Regression

- Uses neural networks to model complex relationships between the inputs and the outputs.



Explain about Logistic Regression.

Logistic regression is a supervised machine learning algorithm widely used for binary classification tasks, such as identifying whether an email is spam or not and diagnosing diseases by assessing the presence or absence of specific conditions based on patient test results.



$$\frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$$