Data Analytics

Unit - 2 [One-Shot]

Most important topics:

1. Regression modeling and multivariate analysis.
2. Baysian networks
3. SVM and kernel methods.
4. Time - series analysis.
5. PCA and fuzzy decision trees
6. Stochastic search methods.

Regression modeling :

- It is a statistical technique used to understand the relationship between a dependent variable and one or more independent variables.
- The goal is to model the expected value of the dependent variable based on the values of the independent variables.
- Regression modelling is widely used for prediction, forecasting, and statistical strategies.

* Types of regression models :

1. Linear Regression :

→ Simple linear regression : Models the relationship between two variables by fitting a linear equation to observed data.
i.e. b/w an independent and a dependent variable.

→ Multiple linear regression : Models the relationship between more than One independent and single dependent variable.
i.e. b/w one dependent and multiple independent variables.

2. Polynomial regression : Models the the relationship between the dependent variable and independent variable as an

n^{th} degree polynomial.

e.g: predicting the growth of a plant over time, where growth accelarates at different rates.

3. Ridge regression: It includes a penalty term to avoid overfitting. This penalty term discourages large coefficients by adding the sum of their squares to cost function.

i.e. for regularization.

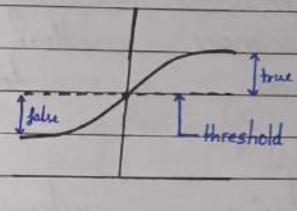
e.g: Predicting a students performance based on many related features like study hrs, activities, etc.

4. Lasso regression: It is also a regularization model that simplifies the complex models and eliminates some specific features for effective analysis.

e.g: Predicting weight on the basis of diet, excercise, lifestyle, etc.

ARFAT Date Page

5. Logistic regression: Used for predicting binary outcomes (yes/no, true/false). Instead of fitting straight line, it fits S-shaped curve to estimate the probability of a particular outcome.



e.g: Predicting whether a customer will buy a product or not based on their age, behaviour, and browsing data.

6. Random forest: It combines multiple decision trees to improve predictive performance and robustness of the model. It averages the predictions of all trees individually.

e.g: Predicting a car's price based on its maintainance, mileage and brand.

@brevilearning

Multivariate analysis :

- It is a statistical approach used to understand the relationships between multiple variables simultaneously.
- It involves observing and analysing more than one outcome variable at a time.

e.g: analysing customer behaviour not only on basis of product he buys but how factors like age, income, location, browsing history, etc. can affect the purchase.

Techniques used / types of MVA :

i) PCA (Principal Component Analysis)
ii) Regression analysis
iii) Clustering
iv) Factor analysis
v) Multivariate regression

- MVA helps in understanding relations, reducing dimensions, and making informed decisions in various fields.

Baysian networks:

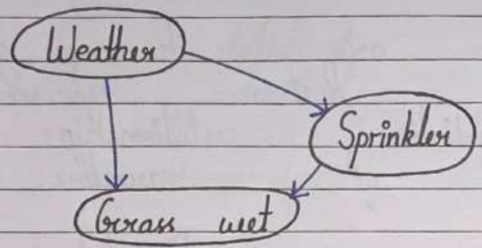
- These are the graphical models that represents the probabilistic relationships among a set of variables.
- They use Directed Acyclic Graphs (DAGs) where nodes and edges ~~variables~~ are used to represent relationship b/w model variables and their conditional probabilities.

Key points :

- Nodes : Each node in a network represents a variable.
- Edges : Directed arrows b/w nodes that indicates conditional dependencies.
- Probabilities : Each node has an associated probabilities that quantifies the likehood of different outcomes.

Like, Share & Subscribe
→ @brevilearning YT

example:



variables:
- Weather (W): Can be sunny or rainy.
- Sprinkler (S): Can be on or off.
- Grass Wet (G): Can be wet or dry.

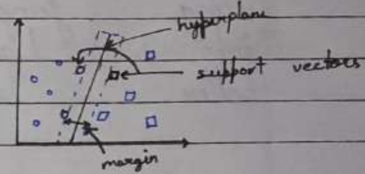
edges:
- $W \rightarrow S$
- $W \rightarrow G$
- $S \rightarrow G$

Conditional probabilities:
- $P(W)$: Probability of weather (Rainy or Sunny).
- $P(S/W)$: Prob. of sprinkler being on or off.
- $P(G/W, S)$: Prob. of grass wet or dry given the weather and sprinkler status.

SVM and Kernel methods:

* SVM (Support Vector Machines):

- These are supervised learning models used for classification and regression.
- They find the optimal boundary (hyperplane) that best separates different classes.



Key points:

- Hyperplane: A flat boundary that separates different classes.
- Support Vectors: Data points closest to the hyperplane that defines the margin.
- Margin: The distance b/w the hyperplane and the nearest support vectors. SVM aims to maximize this margin for better separation.

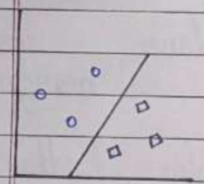
e.g: Classifying emails as spam or not.

* Kernel Methods :

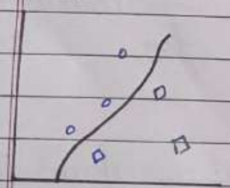
- It allows SVMs to handle non-linear data by transforming it into a high dimensional space without computing the co-ordinates explicitly.
- It enables SVMs to handle complex, non-linear data by using functions to map data into higher dimensions for better separation.

Types of Kernels used in SVM :

1. Linear Kernel: It is ideal for linearly separable data, means it can be best when data can be separated by a straight line.



2. Polynomial Kernel: It handles polynomial relationships in data. It is flexible but computationally more intensive.



3. RBF (Radial Basis Function) Kernel: Also known as Gaussian kernel, it's versatile and great for non-linear data.

4. Sigmoid kernel: Similar to the activation function in neural networks, it's useful for non-linear data but sometimes less effective than RBF kernels.

Time - series analysis :

- It involves the collection, analysis and interpretation of data points collected or recorded at specific time intervals to understand underlying patterns and predict future values.

* Components of time - series :

- Trend: The long - term movement or direction in the data over a period of time.

e.g.: Upward trend of stock - market over several years.

- **Seasonality**: Regular repeating patterns or cycles in data occuring at specific intervals. i.e. daily, monthly, or annually.

e.g: Increased sales of ice-cream during summer.

- **Cyclical**: Long-term wave-like patterns in the data not tied to a fixed calender period.

e.g: Bussiness cycles of economic expansion and contractions.

- **Noise**: Random variations or irregularities in the data that do not follow any pattern.

e.g: Unexpected spikes in temperature readings due to sensor errors.

* Applications of Time Series Analysis:

- Finance
- Economics
- Weather forecast
- Healthcare
- Retail

PCA (Principal Component Analysis):

- It is a statistical technique used for dimentionality reduction.
- It transforms the original variables into a new set of uncorrelated variables called principal components, which capture the maximum variance in the data.
- It helps reducing complexity of the data while retaining as much varience as possible.

Working:

Step 1: Standardization: Ensures that the data is centered and scaled.

Step 2: Covariance matrix: Computes the covariance matrix to understand how variables vary together.

Step 3: Eigen value and Eigenvectors: Compute all Eigenvalue and Eigenvectors of the covariance matrix.

- Eigenvalue represents the variance captured by each principal component.
- Eigenvectors determines the direction of the principal components.

Step 4: Sort and select Principal Components:

- Sorts the eigenvalues in descending order and arrange the corresponding eigenvectors.
- The eigenvector with highest eigenvalue is the first principal component.

Step 5: Transform Data: Project the original data onto the new set of axes defined by the eigenvectors.

★ Applications of PCA:

- Data visualization
- Noise filtering
- Feature extraction
- Image compression

Fuzzy decision trees:

- FDTs combine decision tree learning with fuzzy logic to handle imprecision and uncertainity in data.
- It is used to enhance decision trees by incorporating fuzzy set theory for more flexible and human-like reasoning.

★ Key concepts in FDTs:

1. Fuzzy set: A set with gradual membership than binary. i.e. elements can partially belongs to a set with a membership degree between 0 and 1

   e.g: height → tall, very tall, short, etc

2. Membership function: Defines how each point in the input space is mapped to a membership value between 0 and 1.

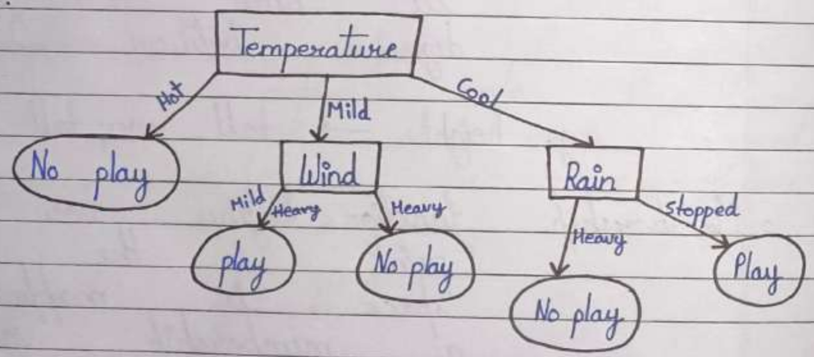
* Structure of FDTs :

i) Nodes : Represents test or conditions based on fuzzy logic. Each node can have multiple branches corresponding to different fuzzy sets.

ii) Branches : Correspond to fuzzy rules derived from the membership functions.

iii) Leaves : Represents the final decision or classification, often with a fuzzy value indicating the degree of certainity.

e.g :



* Advantages of FDTs :

- Robustness : (handles imprecise data effectively)
- Human-like reasoning : (intuitive decision making)
- Verstality : (applicable in various fields)

* Disadvantages of FDTs :

- Complexity : (more complex than traditional decision trees)
- Computationally intensive : (complicated for large dataset)
- Parameter tuning : (requires careful tuning of membership funᶜ for optimal performance)

@brevilearning

⊛ Stochastic Search Methods :

- These are optimization techniques that includes randomness to find solutions in complex search spaces.
- It helps in efficiently exploring large and potentially non-convex search spaces to find optimal ~~and~~ or ~~non~~ near-optimal solutions.
- Helps in exploring new areas in complex landscapes.

* **Some common Stochastic Search methods:**

1. **Genetic Algorithm (GA):**

It is based on principles of natural selection and genetics.

Process:

Initialization
↓
Fitness evaluation
↓
Selection
↓
Crossover
↓
Mutation
↓
Repetition — not found
↓ solun found
Termination

2. **Ant Colony Optimization (ACO):**

Mimics the behaviour of ants finding the shortest path to food.

Process:

Initialization
↓
Solution construction
↓
Updation
↓
Repetition
↓
Termination

3. **Random search:**

It defines the search space and ~~to~~ the no. of iterations that sets the boundaries and duration of search process.

Process:

Initialization
↓
Random sampling
↓
Evaluation
↓
Tracking Best solution
↓
Repetition
↓
Termination

4. Particle Swarm Optimization (PSO):

Inspired by social behaviour of birds flocking.

Process:

Initialization

↓

Velocity and Position Updation

↓

Personal and Global best

↓

Repetition

↓

Termination

Thanks for watching !!

@bravilearning