Name: MohammedAmman Chopadiya

Date: 19th December 2023

K-Nearest Neighbours Algorithm

- Non-parametric, supervised learning classifier
- Uses proximity to make classifications or predictions.
- Class labels assigned based on majority vote.
- "Majority voting" requires a majority of greater than 50%
- For multiple classes, a class label can be assigned with a vote of greater than 25%.

KNN Distance Metrics

Calculating distance between query point and other data points helps form decision boundaries:

- **Euclidean distance (p=2):** Measures a straight line between the query point and the other point.
- **Manhattan distance (p=1):** Measures the absolute value between two points.
- **Minkowski distance:** Generalized form of Euclidean and Manhattan distance metrics.
- **Hamming distance:** Identifies points where vectors do not match, also known as the overlap metric.

Defining K in KNN Algorithm

- Defines how many neighbors are checked.
- Lower values: high variance, low bias.
- Larger values: high bias, lower variance.
- Choice of k depends on input data.
- Keep an odd number for k to avoid classification ties.

Applications of KNN

- Data Preprocessing: KNN algorithm estimates missing values in datasets through missing data imputation.
- Recommendation Engines: KNN algorithm provides automatic recommendations based on user behaviour, but may not be optimal for larger datasets due to scaling issues.
- Finance: KNN is used in credit data assessment, stock market forecasting, currency exchange rates, trading futures, and money laundering analyses.
- Healthcare: KNN predicts heart attacks and prostate cancer risk by calculating most likely gene expressions.
- Pattern Recognition: KNN assists in identifying patterns in text and digit classification, particularly handwritten numbers.

KNN Advantages

- Easy to implement due to simplicity and accuracy.
- Adapts easily to new training samples.
- Requires only a k value and a distance metric, making it low-cost compared to other machine learning algorithms.

KNN Disadvantages

- Slow scalability: KNN consumes more memory and data storage, leading to increased business expenses and slower computation.
- "Curse of dimensionality": Higher classification errors with high-dimensional data inputs due to the "curse of dimensionality."
- Overfitting: KNN's behavior can be influenced by the value of k, with lower values overfitting and higher values smoothing out prediction values.

GitHub Repository Link: Diabetes Prediction with KNN Algorithm