08 K-NN

- **K-Nearest Neighbors (KNN)** is a simple, yet powerful machine learning algorithm used for classification and regression task
- KNN operates on the principle that similar data points are likely to be found near each other
- Given a data point whose classification or value is unknown, KNN will look at the ‘k’ nearest data points (neighbors) in the training dataset to make a prediction

KNN Steps

1. **Choose the value of ‘k’**
    - The first step is to decide how many neighbors (k) you want to consider when making the prediction
    - Common values for k are small positive integers like 3, 5, or 7
2. **Calculate Distance**
    - To find the nearest neighbors, KNN calculates the distance between the data point in question and all the points in the training data
        - **Euclidean distance :** The most common metric, which is the straight-line distance between two points in Euclidean space

          $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}$
        - **Manhattan distance :** The sum of the absolute differences between coordinates

    $d(\mathbf{x}_i, \mathbf{x}_j) = \sum_{k=1}^{n} |x_{ik} - x_{jk}|$

        - **Minkowski distance :** A generalization that includes both Euclidean and Manhattan distances

    $d(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{n} |x_{ik} - x_{jk}|^p\right)^{\frac{1}{p}}$
    - **Chebyshev Distance** : A special case of the Minkowski distance

      $d(\mathbf{x}_i, \mathbf{x}_j) = \max_{k=1}^{n} |x_{ik} - x_{jk}|$
3. **Find Nearest Neighbors**
    - Once the distances are calculated, KNN identifies the k closest data points (the k-nearest neighbors)
4. **Predict the Outcome**
    - **For classification :** The algorithm assigns the class that is most frequent among the k-nearest neighbors

  - **For regression :** The algorithm averages the values of the k-nearest neighbors to predict the outcome

Choosing ‘k’

- **Small k :** Leads to a model that is sensitive to noise in the data (high variance)
- **Large k :** Leads to smoother decision boundaries but might oversimplify the model (high bias)

Advantages of KNN

- **Simple and intuitive :** No assumptions about the data distribution are required
- **Flexible :** Can be used for both classification and regression tasks

Disadvantages of KNN

- **Computationally expensive :** Especially with large datasets since it calculates distances to all training points
- **Sensitive to the choice of k :** Different k values can lead to different results
- **Affected by irrelevant features :** Feature scaling or dimensionality reduction (like PCA) might be necessary

Practical Considerations

- **Feature scaling :** Standardizing or normalizing features is important because KNN relies on distance metrics
- **Handling large datasets :** Techniques like KD-Trees or Ball Trees can optimize the search for nearest neighbors