

# • KNN & PCA ASSIGNMENT

- 

- **K-Nearest Neighbors (KNN)**

- 1. What is K-Nearest Neighbors (KNN) and how does it work?**


- KNN is a supervised machine learning algorithm used for both classification and regression. It works by storing all available cases and classifying new cases based on a similarity measure (e.g., distance<sup>1</sup> functions).
- For classification, it assigns the class that is most common among its 'k' nearest neighbors.
- For regression, it predicts the value based on the average (or median) of the values of its 'k' nearest neighbors.

- 2. What is the difference between KNN Classification and KNN Regression?**

- KNN Classification:** Predicts a categorical label. The output is a class label determined by the majority vote of the k-nearest neighbors.
- KNN Regression:** Predicts a continuous value. The output is the average (or median) of the values of the k-nearest neighbors.

- 3. What is the role of the distance metric in KNN?**

- The distance metric determines how the "nearest" neighbors are found. Common metrics include:

- Euclidean distance:  $d(p,q)=\sum_{i=1}^n (p_i - q_i)^2$  
- Manhattan distance:  $d(p,q)=\sum_{i=1}^n |p_i - q_i|$
- Minkowski distance:  $d(p,q)=(\sum_{i=1}^n |p_i - q_i|^r)^{1/r}$  (generalization of Euclidean and Manhattan)

- The choice of distance metric can significantly impact the performance of KNN.

- 4. What is the Curse of Dimensionality in KNN?**

- As the number of features (dimensions) increases, the data becomes increasingly sparse. In high-dimensional spaces, the concept of "nearest" becomes less meaningful, and the distance between any two points tends to converge. This can lead to poor performance for KNN.

- 5. How can we choose the best value of K in KNN?**

- Use cross-validation techniques (e.g., k-fold cross-validation).
- Try different values of K and evaluate the model's performance on a validation set.

- c. A small K can lead to noisy results, while a large K can smooth out decision boundaries and potentially ignore local patterns.
- d. Typically odd values of k are used in binary classification to avoid ties.

**6. What are KD Tree and Ball Tree in KNN?**

- a. **KD Tree (K-Dimensional Tree):** A space-partitioning data structure that organizes points in a k-dimensional space. It is used to efficiently find nearest neighbors.
- b. **Ball Tree:** Another space-partitioning data structure that organizes points into nested hyperspheres (balls). It is often more efficient than KD Trees in high-dimensional spaces.

**7. When should you use KD Tree vs. Ball Tree?**

- a. **KD Tree:** Works well for low to medium-dimensional data. It can become inefficient in very high-dimensional spaces.
- b. **Ball Tree:** Generally performs better in higher-dimensional spaces and for non-uniformly distributed data.

**8. What are the disadvantages of KNN?**

- a. Computationally expensive for large datasets.
- b. Sensitive to irrelevant features and the scale of features.
- c. Requires storing the entire training dataset.
- d. Does not work well with high dimensional data.

**9. How does feature scaling affect KNN?**

- a. KNN relies on distance calculations. Features with larger scales can dominate the distance, leading to biased results. Feature scaling (e.g., standardization or normalization) is crucial to ensure that all features contribute equally.

**Principal Component Analysis (PCA)**

**10. What is PCA (Principal Component Analysis)?**

- a. PCA is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional<sup>2</sup> representation while preserving as much variance as possible. It identifies the principal components, which are<sup>3</sup> orthogonal linear combinations of the original features.

**11. How does PCA work?**

- a. It calculates the covariance matrix of the data.
- b. It finds the eigenvectors and eigenvalues of the covariance matrix.
- c. It sorts the eigenvectors by their corresponding eigenvalues in descending order.

- d. It selects the top k eigenvectors (principal components) to form a new feature space.
- e. It transforms the original data into this new feature space.

**12. What is the geometric intuition behind PCA?**

- a. PCA finds the directions (principal components) in which the data varies the most. The first principal component is the direction of maximum variance,<sup>4</sup> the second is the direction of maximum variance orthogonal<sup>5</sup> to the first, and so on.

**13. What is the difference between Feature Selection and Feature Extraction?**

- a. **Feature Selection:** Chooses a subset of the original features.
- b. **Feature Extraction:** Creates new features by combining or transforming the original features. PCA is a feature extraction technique.

**14. What are Eigenvalues and Eigenvectors in PCA?**

- a. **Eigenvectors:** Directions or axes along which the data varies the most.
- b. **Eigenvalues:** Magnitudes that represent the amount of variance explained by each eigenvector. Larger eigenvalues correspond to more important principal components.

**15. How do you decide the number of components to keep in PCA?**

- a. Use the explained variance ratio. Keep enough components to capture a significant portion of the total variance (e.g., 95%).
- b. Use scree plots to visualize the eigenvalues and look for an "elbow" point where the eigenvalues start to level off.

**16. Can PCA be used for classification?**

- a. Yes, PCA is primarily used for dimensionality reduction, which can improve the performance of classification algorithms by reducing noise and redundancy. The reduced data can then be used as input for a classification model.

**17. What are the limitations of PCA?**

- a. Assumes linear relationships between features.
- b. Can lose information if the variance is not a good measure of importance.
- c. Not scale-invariant (feature scaling is important).
- d. Difficult to interpret the new features.

**18. How do KNN and PCA complement each other?**

- a. PCA can be used to reduce the dimensionality of the data before applying KNN. This can mitigate the curse of dimensionality and improve the efficiency and accuracy of KNN.

**19. How does KNN handle missing values in a dataset?**

- a. KNN itself does not inherently handle missing values.

- b. Common approaches include:
  - i. Imputation: Replace missing values with the mean, median, or mode of the feature.
  - ii. Distance-based imputation: Impute based on the values of the nearest neighbors.
  - iii. Removing the rows with the missing values.

**20. What are the key differences between PCA and Linear Discriminant Analysis (LDA)?**

- a. **PCA:**
  - i. Unsupervised dimensionality reduction.
  - ii. Maximizes variance.
  - iii. Does not consider class labels.
- b. **LDA:**
  - i. Supervised dimensionality reduction.
  - ii. Maximizes class separability.
  - iii. Uses class labels to find the best features.