classification → classifying observation into labels

Regression → predicting a number

K-Nearest Neighbors (K-NN)

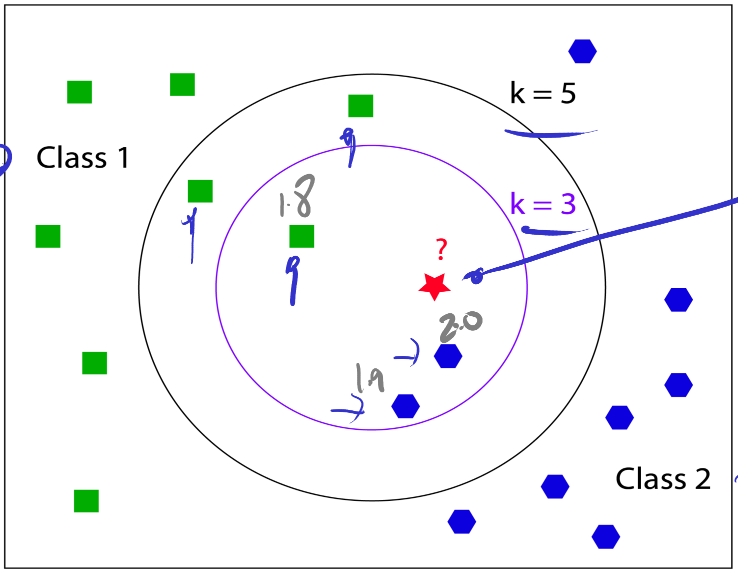
1. Introduction to K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a simple, yet powerful supervised machine learning algorithm used for both classification and regression tasks. It makes predictions based on the 'k' closest data points in the feature space.

2. How KNN Works

1. Choose the number of neighbors: k ✓
2. Compute the distance between the new data point and all points in the training set (commonly using Euclidean distance).
3. Select the k nearest neighbors.
4. For classification, assign the majority label among the k neighbors.
5. For regression, compute the average of the neighbors' values.

Classification task



new data point

if k=3, then the predicted label is blue hexagon

If k=5, then the predicted label is green square

3. Common Applications of KNN in Data Science

- **Customer segmentation:** Classifying users based on purchasing behavior.
- **Medical diagnosis:** Predicting disease based on patient features.

4. Advantages of KNN

- 1 **Simple to understand and implement.**
- 2 **Non-parametric:** No assumptions about data distribution. ✓
- 3 **Versatile:** Works for both classification and regression.

5. Disadvantages of KNN

- 1 **Computationally expensive:** Slow with large datasets since it stores the entire training data.
- 2 **Sensitive to noise and outliers.**
- 3 **Feature scaling is required:** Performance is affected if features are on different scales.
- 4 **Not interpretable:** Doesn't provide insights into feature importance.

6. When to Use KNN

- 1 When your dataset is small to medium in size.
- 2 When the data is labeled and not too noisy.
- 3 When interpretability is not a primary concern.

7. Summary

K-Nearest Neighbors is a valuable algorithm for data scientists to understand due to its simplicity and versatility. While not ideal for large or high-dimensional datasets, it offers a practical introduction to the concepts of distance-based learning and instance-based algorithms.