K-NEAREST NEIGHBORS

INSTANCE-BASED VS MODEL-BASED

- Instance-Based

- Model-Based

K-NEAREST NEIGHBORS

KNN OR K-NN OR KNN

- K Nearest Neighbors is a supervised instance-based ML model
- It is general fairly quick fit and predict
- It often does a surprisingly good job
- Can be used for both regression and classification tasks

K-NN METHOD

PREDICTION ALGORITHM

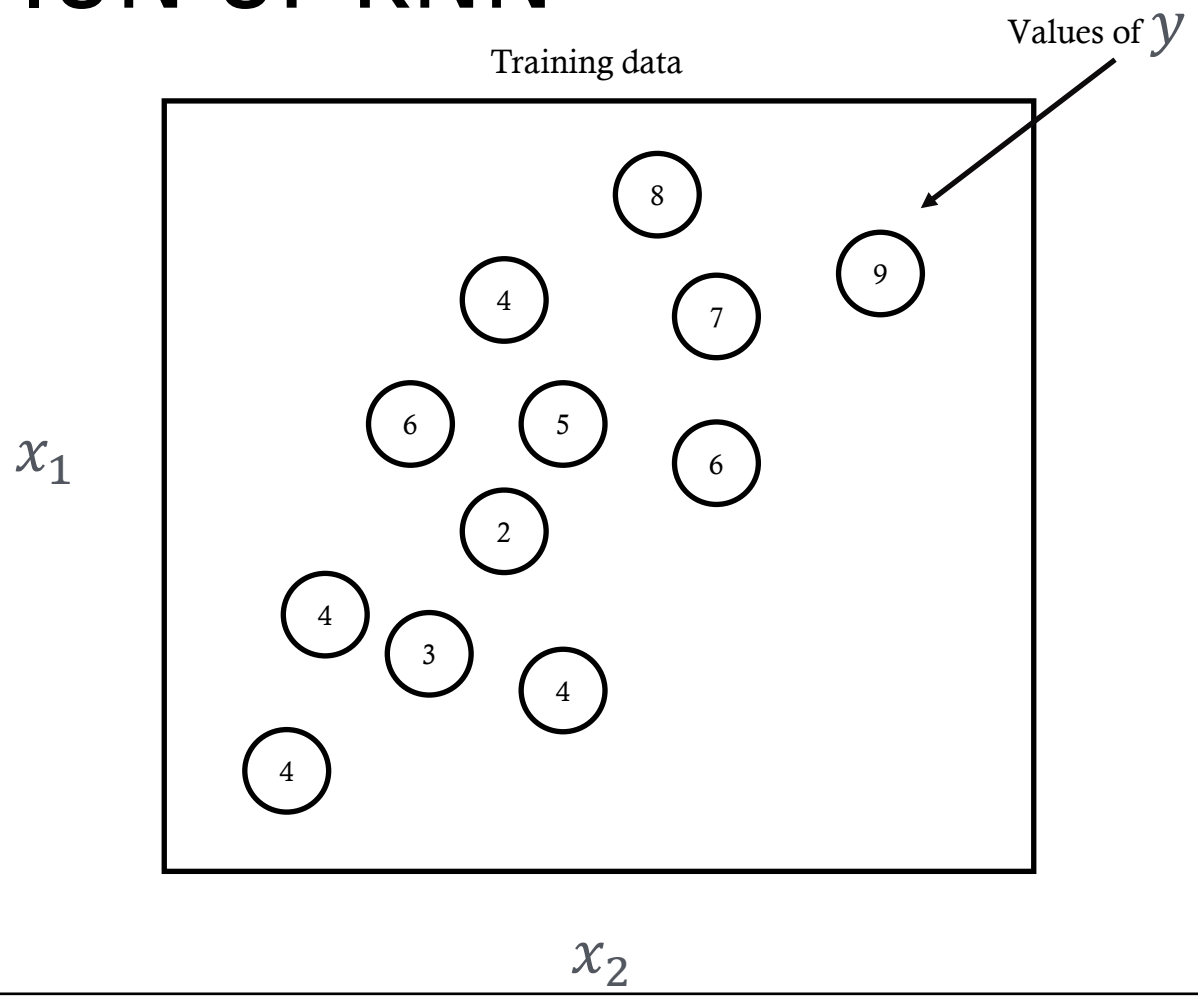
- Calculate the distance from the value to be predicted to all the points in the training data
    - How should we calculate distance?

- Find the k-nearest training data points
- Predict the target value for the new point:
    - Regression:

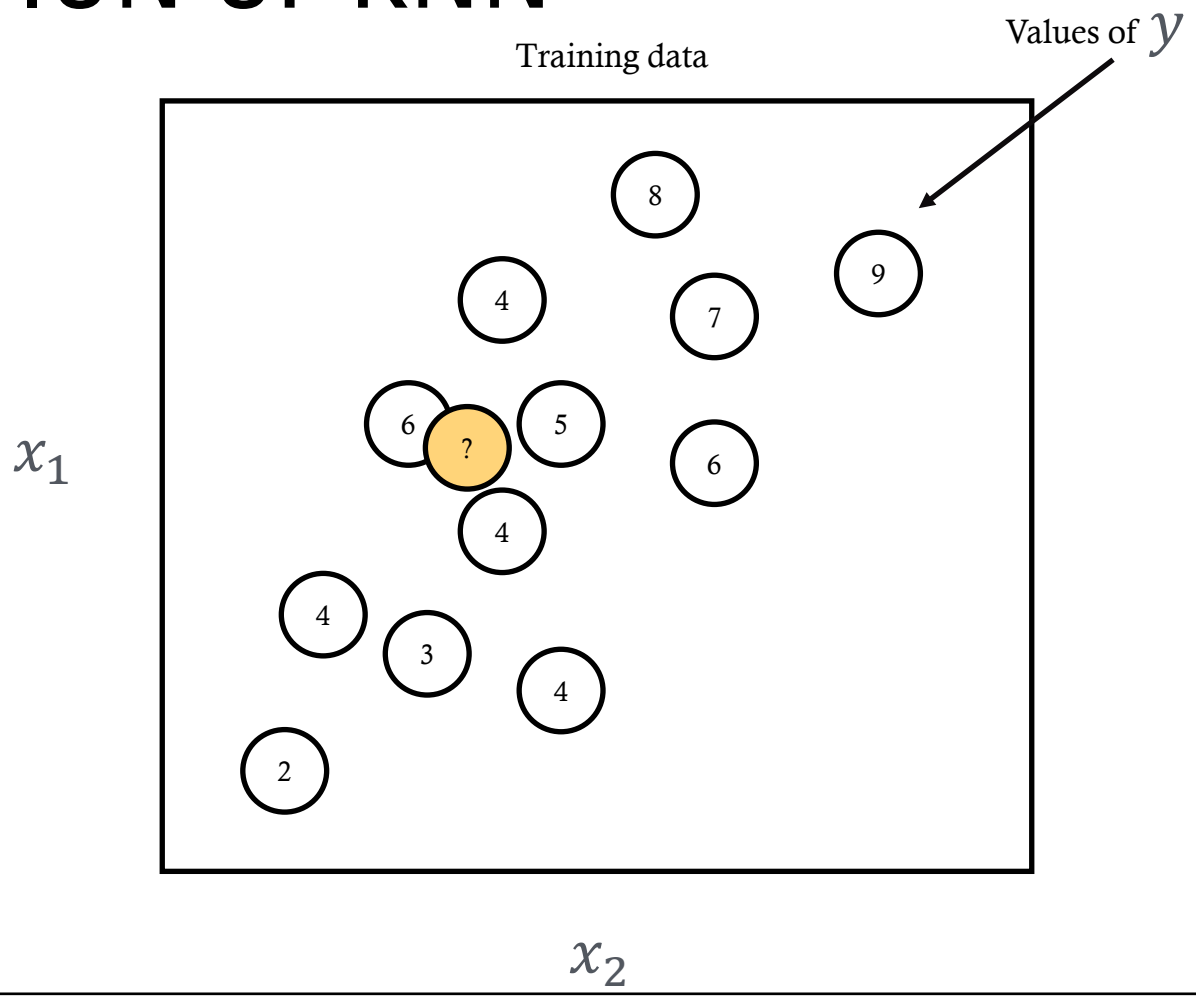
    - Classification:

K-NN REGRESSION

VISUALIZATION OF KNN



VISUALIZATION OF KNN

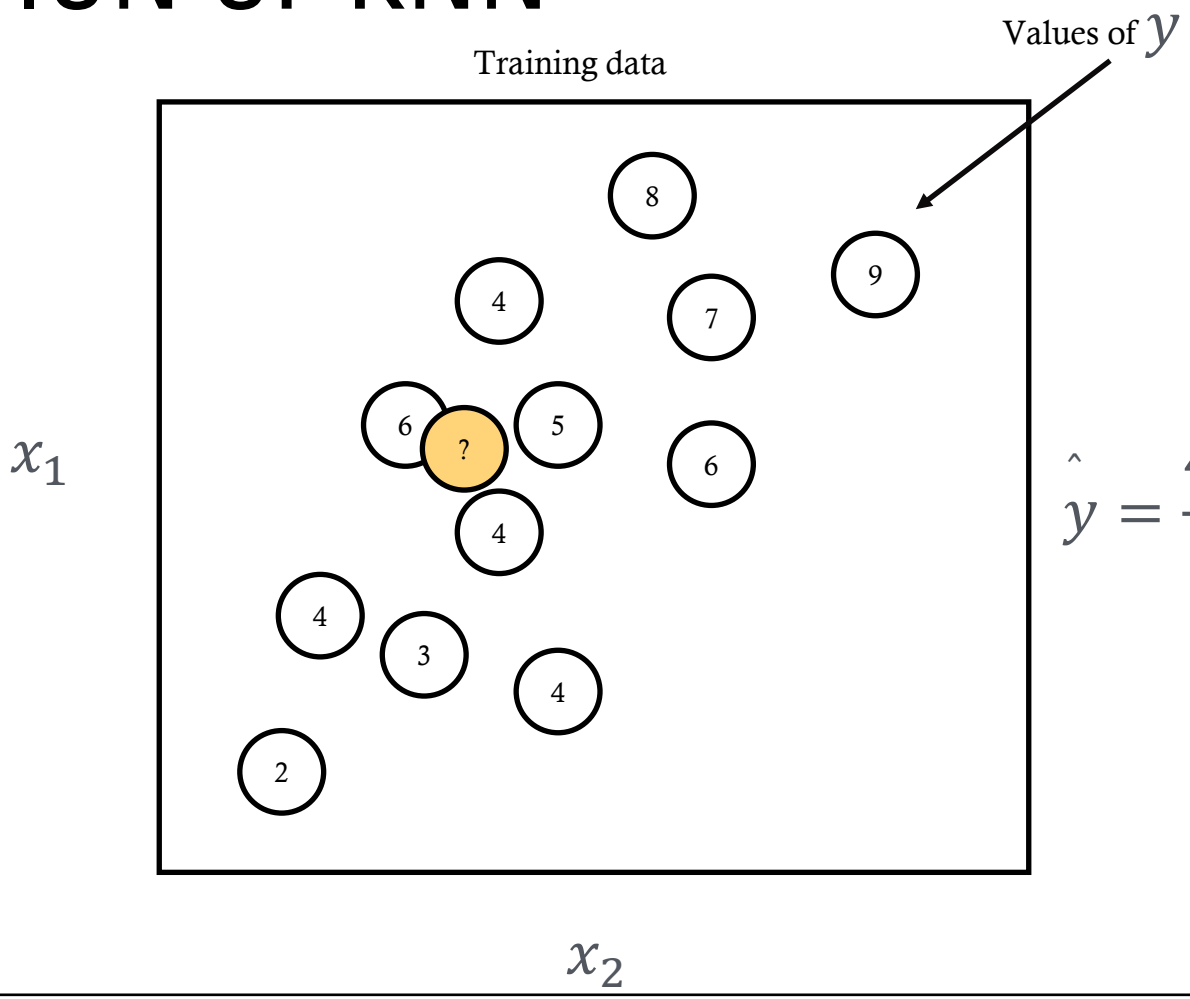


VISUALIZATION OF KNN

$k = 1$

$\hat{y} = 6$

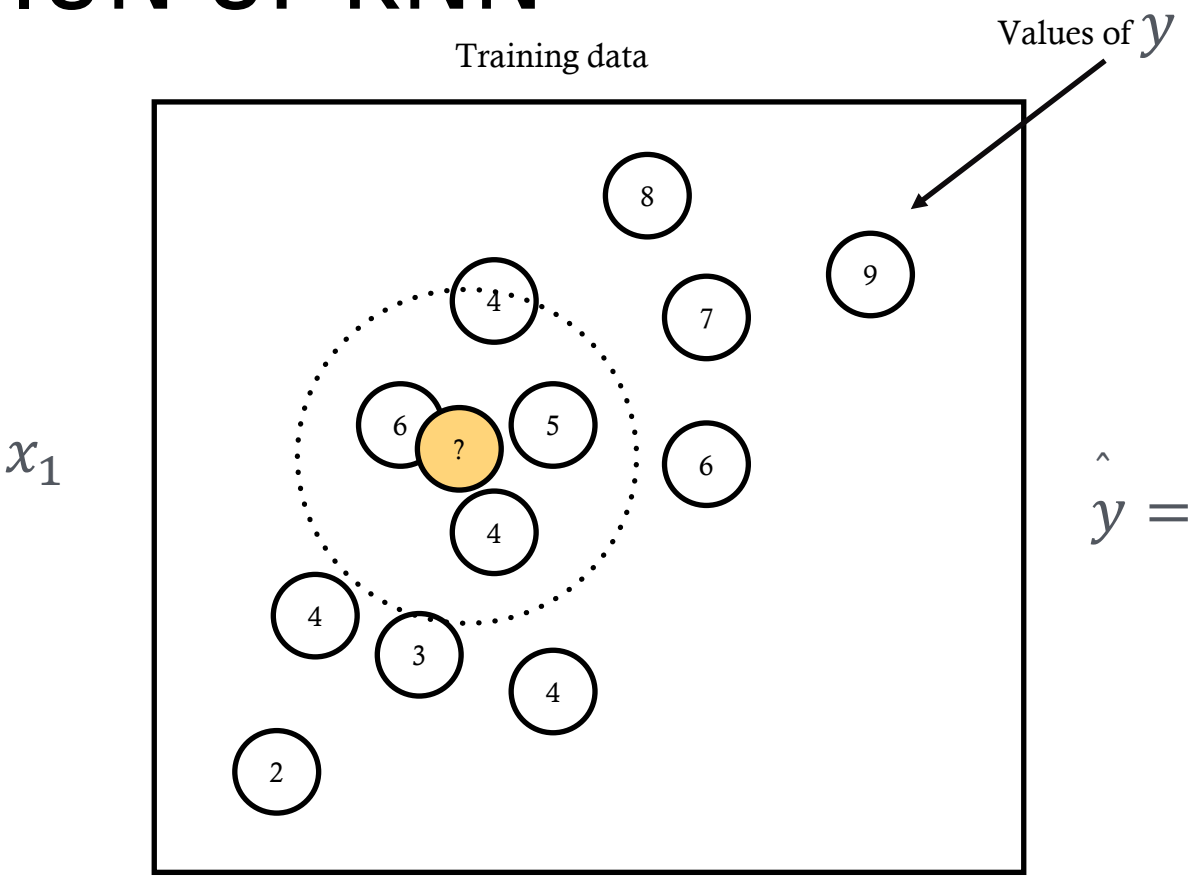
VISUALIZATION OF KNN



$$k = 3$$

$$\hat{y} = \frac{4 + 5 + 6}{3} = 5$$

VISUALIZATION OF KNN

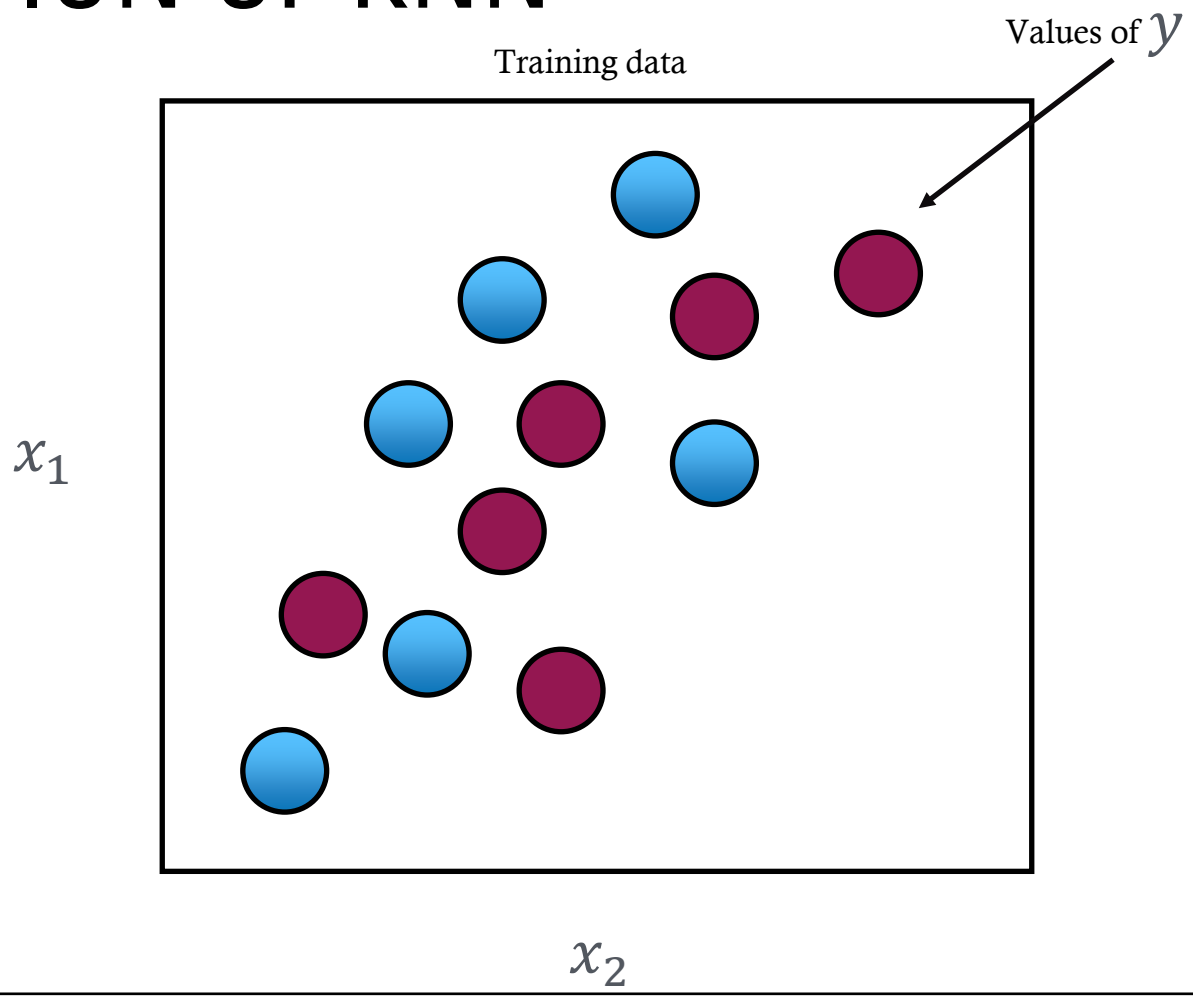


$$k = 5$$

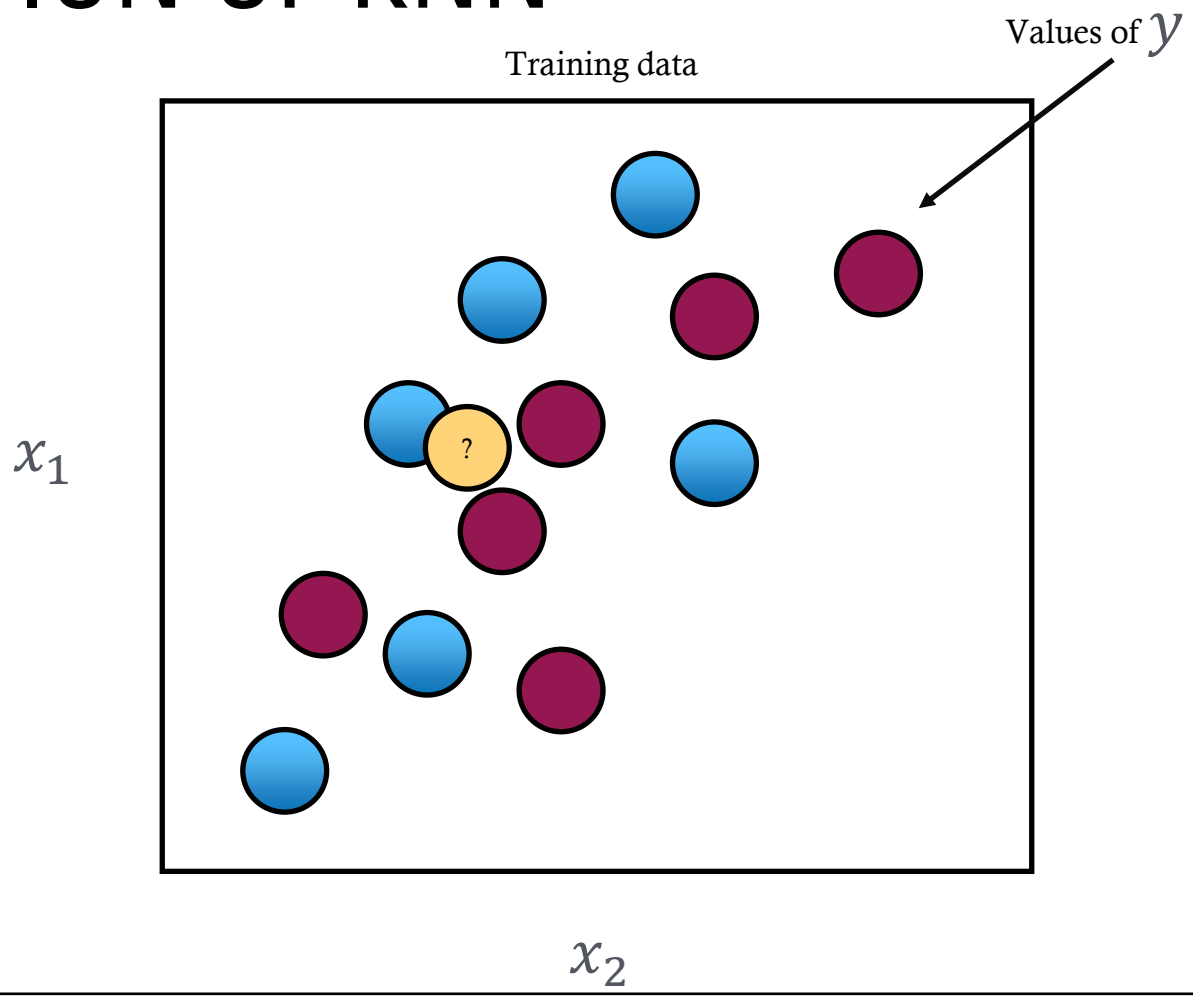
$$\hat{y} = \frac{22}{5} = 4.4$$

K-NN CLASSIFICATION

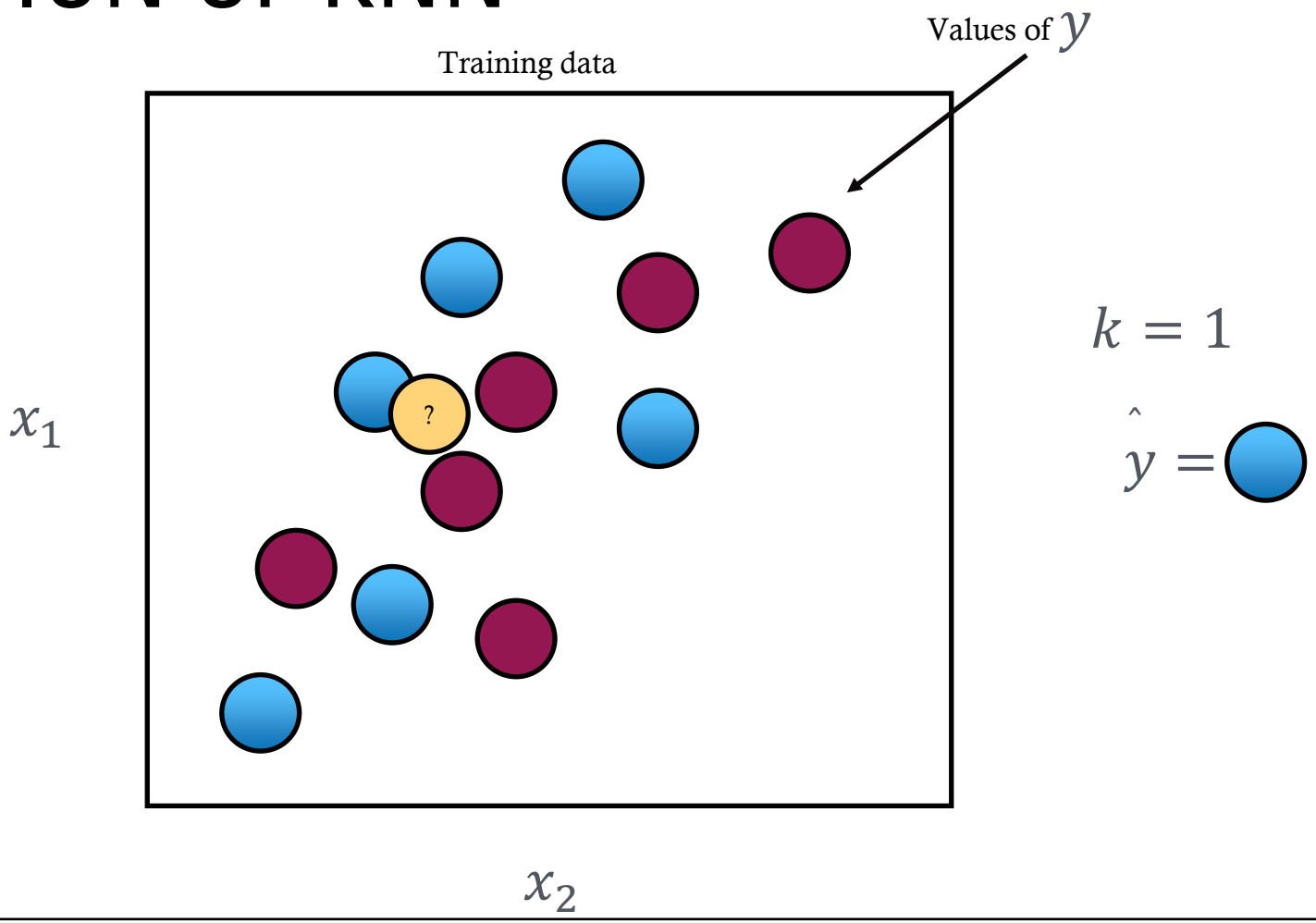
VISUALIZATION OF KNN



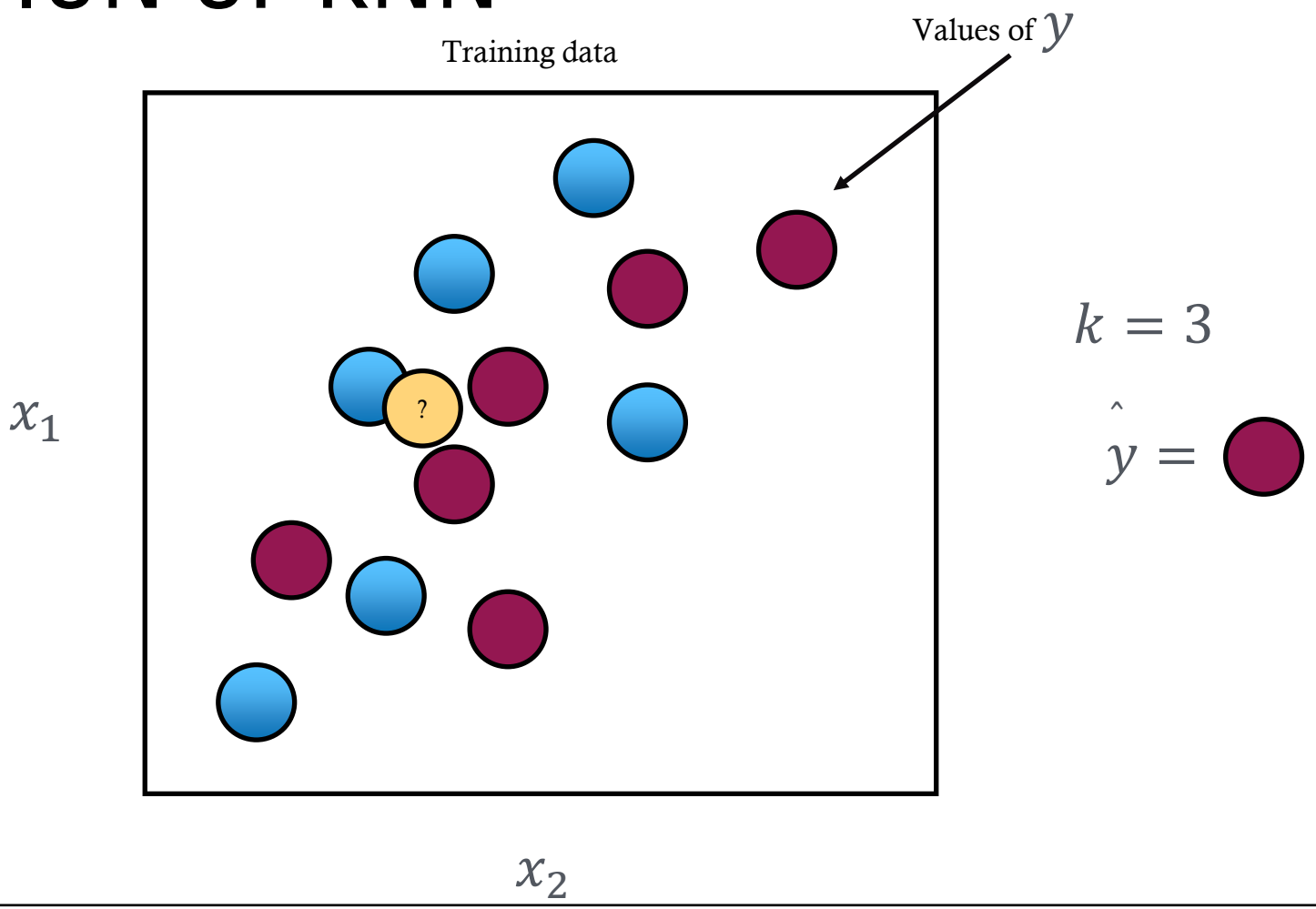
VISUALIZATION OF KNN



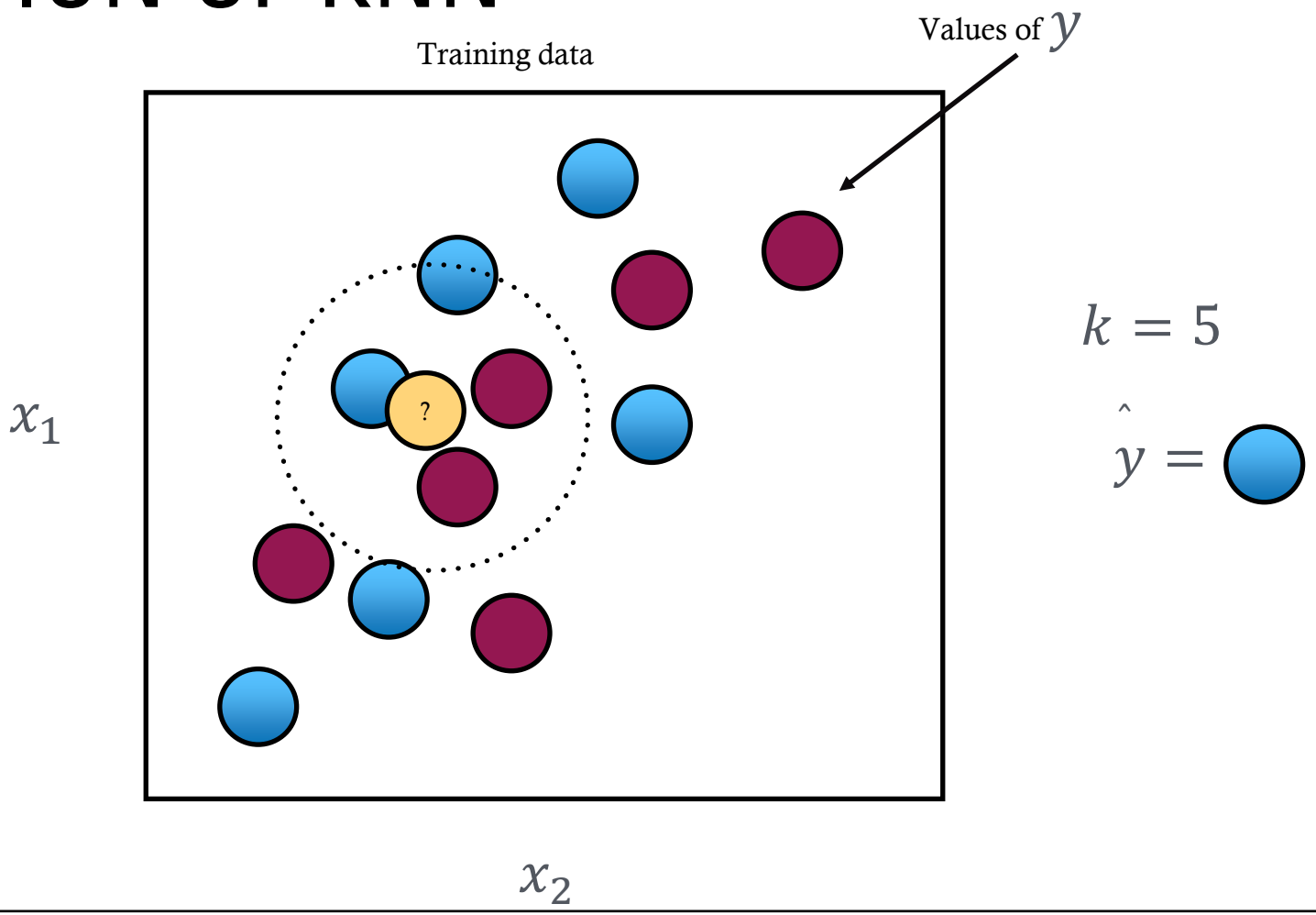
VISUALIZATION OF KNN



VISUALIZATION OF KNN

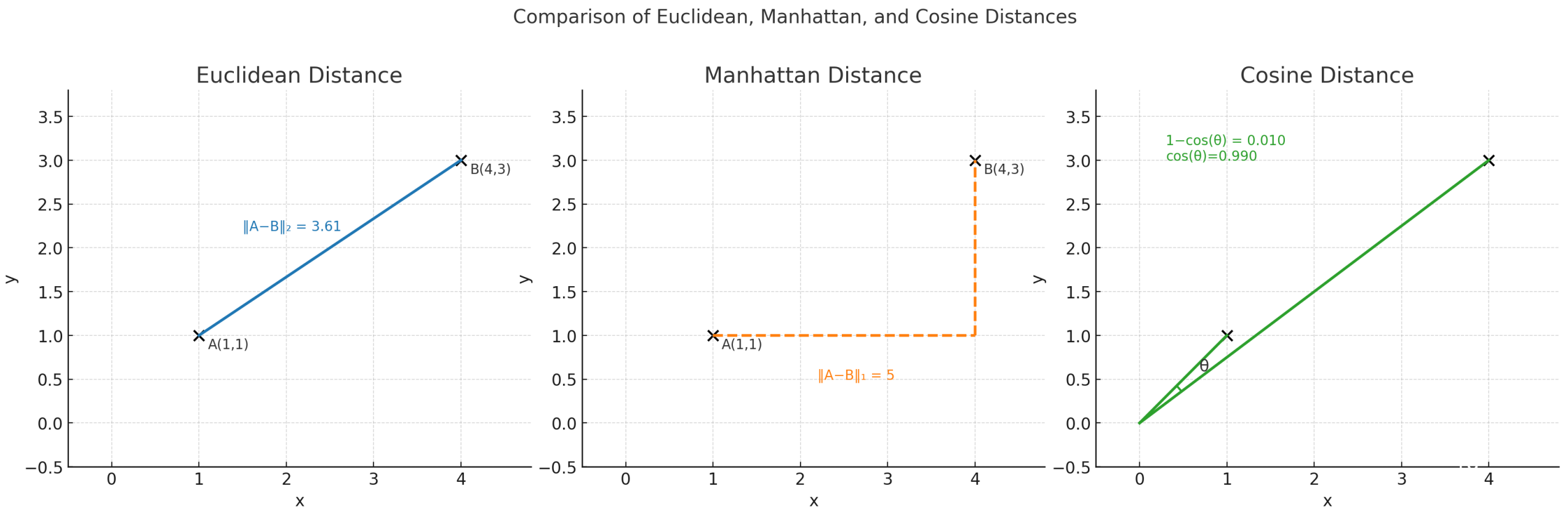


VISUALIZATION OF KNN



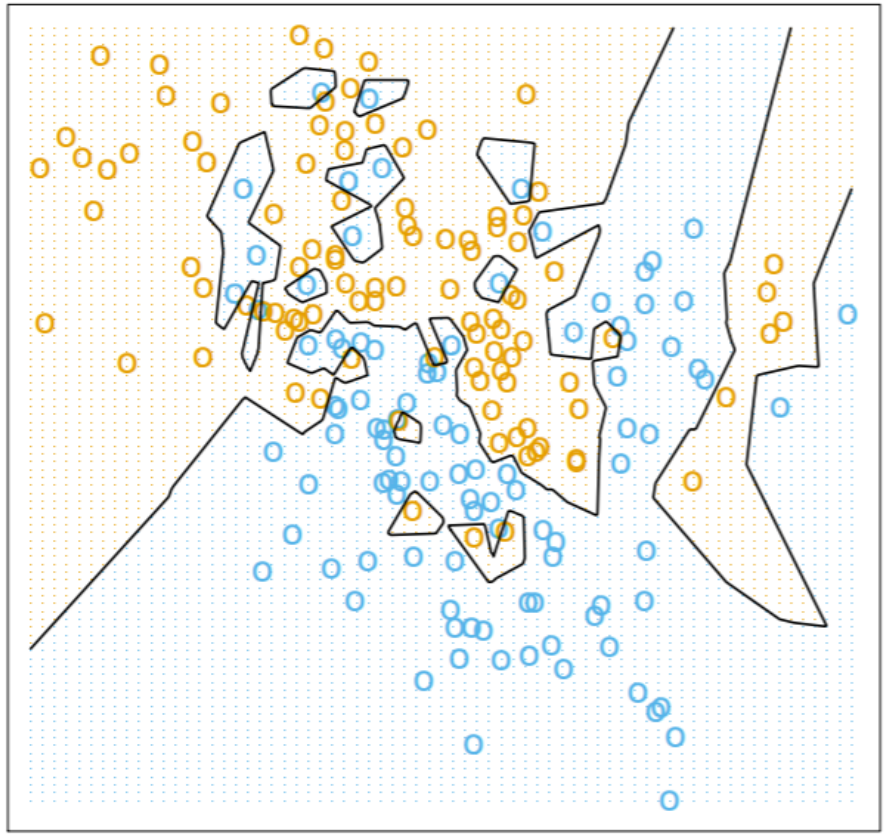
CHOICE OF DISTANCE

- Euclidean most common
- Cosine is often used for high-dimensional data

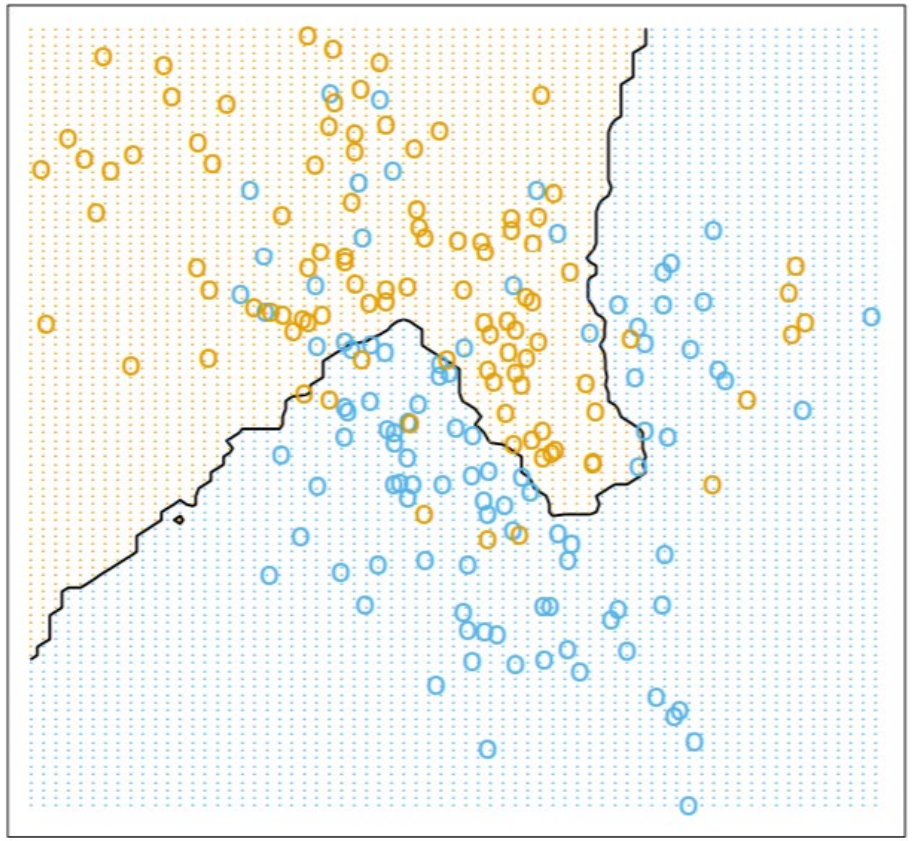


EFFECT OF K ON DECISION BOUNDARY

1 - NN



15 - NN



https://fderyckel.github.io/machinelearningwithr/knnchapter.html

KNN CONSIDERATIONS

- For what types of data is kNN most suitable?
- - Continuous, well-scaled numerical data with moderate dimensionality.
- Why does kNN struggle in high-dimensional spaces?
- - Distances lose meaning ("curse of dimensionality"), making neighbors unreliable.
- What happens when the dataset is very large?
- - kNN must compute distances to all points → slow predictions.
- How does noise or irrelevant features affect kNN?
- - They distort distance calculations and harm neighbor selection.
- Why do we standardize or normalize features before using kNN?
- - To ensure features contribute fairly; prevents scale dominance.
- How does class imbalance impact kNN?
- - Majority-class neighbors dominate, biasing predictions.
- How does the choice of k affect outcomes?
- - Low k: overfit, high k: underfit
- Is kNN discriminative or generative?
- - Discriminative
- Is kNN parametric or non-parametric?
- - Non-parametric

K-NN SUMMARY

Choosing k :

- Small k : Sensitive to noise, can overfit.
- Large k : Smooths boundaries, but may underfit.

Distance Metrics: Common ones are:

- Euclidean Distance (default in Scikit-learn)
- Manhattan Distance
- Minkowski Distance (generalization)
- Cosine Distance (based on angles)

IMPORTANT Scaling Features: Since KNN is distance-based, ensure features are on a similar scale (e.g., use Min-Max Scaling or Standardization).

Advantages:

- Simple to understand and implement.
- No explicit training phase; training data is directly used.

Disadvantages:

- Computationally expensive for large datasets.
- Sensitive to irrelevant features and feature scaling.