K-Nearest Neighbors

Jiahui Chen
Department of Mathematical Sciences
University of Arkansas

Introduction

- K-Nearest Neighbors (K-NNs) have been used for statistical estimation and pattern recognition since the 1970s
- K-NN is a non-parametric technique (non-assumption on data distribution)
- It is still one of the top 10 data mining algorithms.
- It can be used for both classification and regression

Iris Dataset

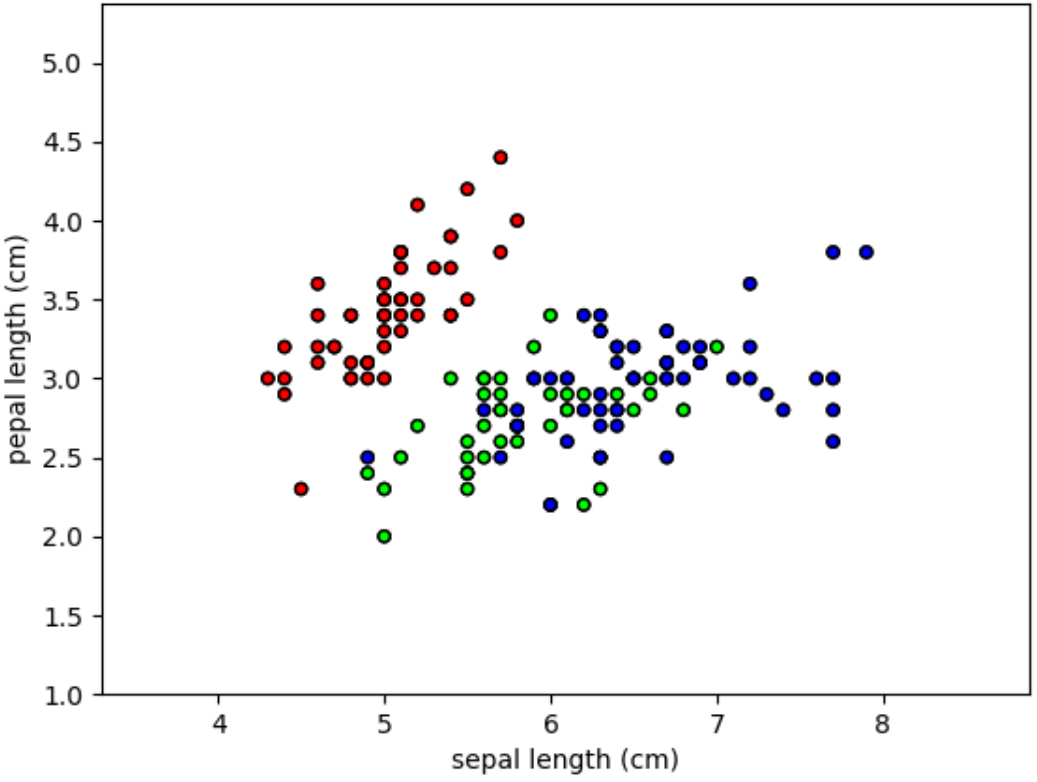
Classification

We consider the **iris** dataset

- Include three types of iris plant:
  - iris setosa,
  - iris versicolour
  - iris virginica
- 4 features:
  - sepal length in cm
  - sepal width in cm
  - petal length in cm
  - petal width in cm
- 150 samples (50 in each of three classes)

UNIVERSITY OF ARKANSAS

Distribution



Predictor Construction

- Construct a predictor:

$$p_{\mathbf{c}}(\mathbf{x}) = ?$$

- No explicit formulation for $p_{\mathbf{c}}(\mathbf{x})$ and no parameters $\mathbf{c}$

***k-NN* algorithm:**

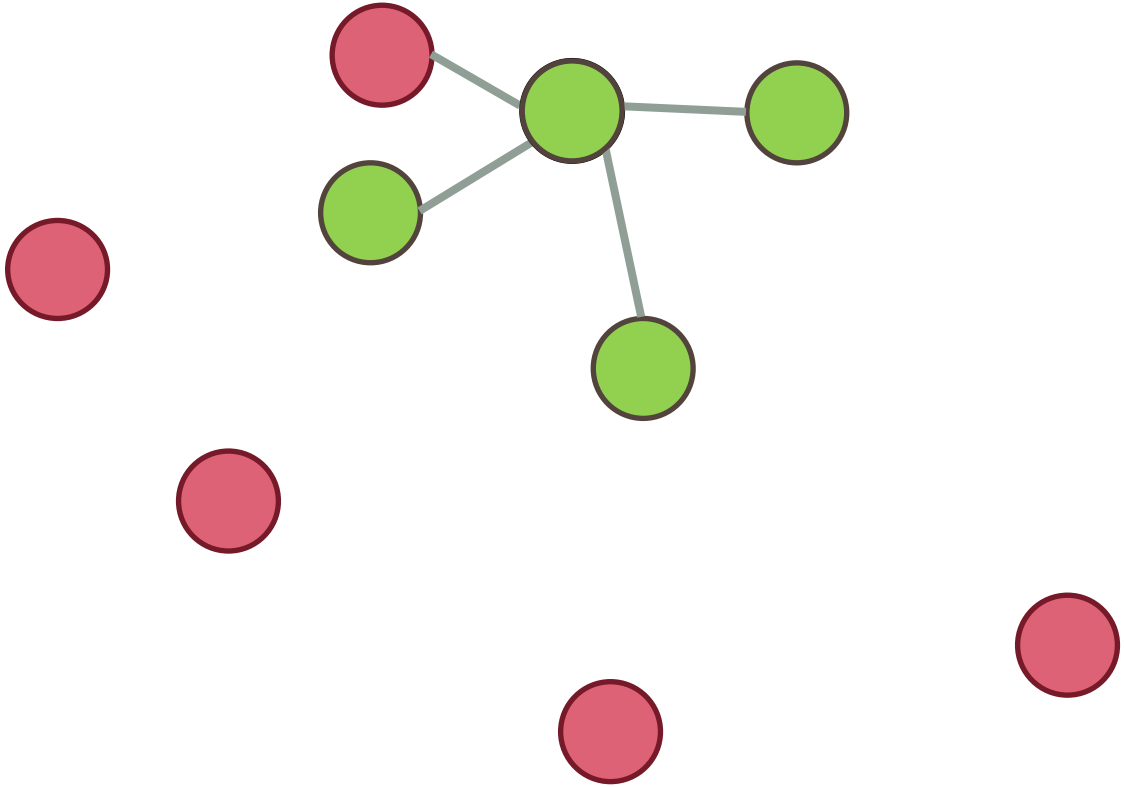
- k is a given positive number
- $\mathbf{x}$ is the feature vector of new sample associated with unknown label y
- find k entries in our dataset that are closest to the new sample $\mathbf{x}$
- label of $\mathbf{x}$ decided by those k entries

UNIVERSITY OF ARKANSAS

K-NN Predictions

- In classification, the K-NN prediction is based on the majority rule of the *k* nearest neighbors
- In regression, the K-NN prediction is the average of the *k* nearest neighbor labels (values)

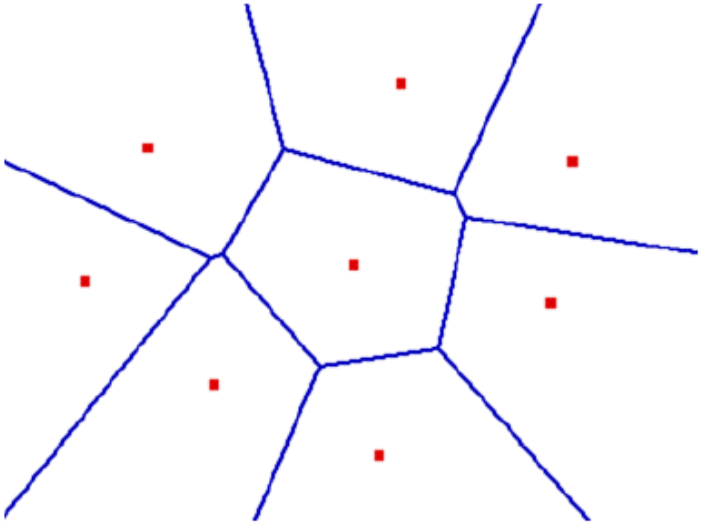
Intuitive algorithm Illustration



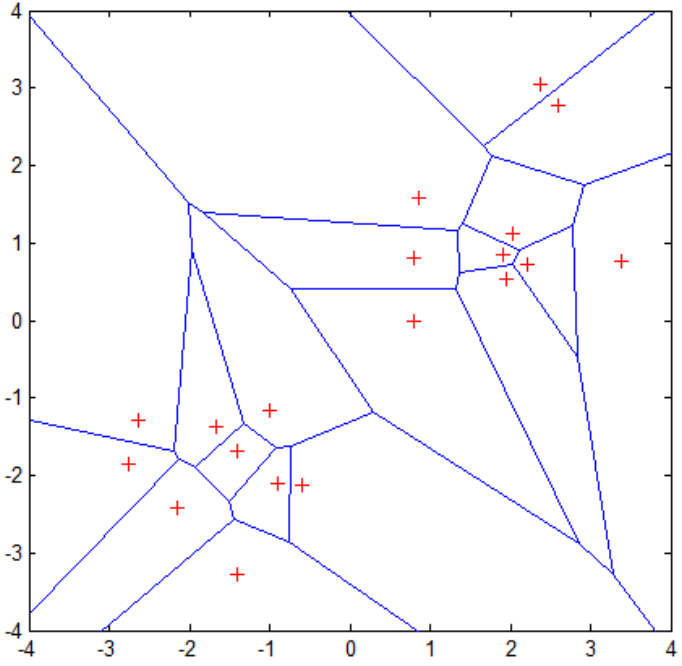
4-NN

Decision Boundary

- Given a set of points, a **Voronoi diagram** describes the areas that are nearest to any given point.
- These areas can be viewed as zones of control.



Graphic Depiction

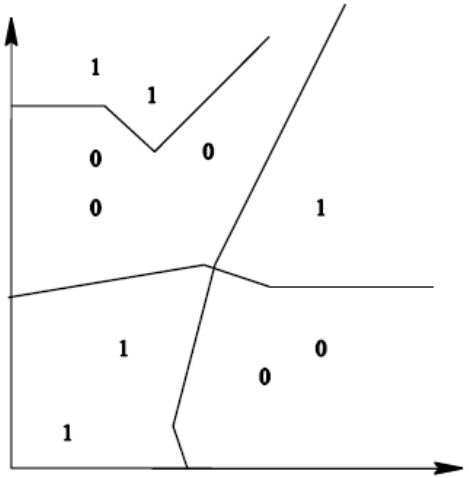


- Properties:

1) All possible points within a sample's Voronoi cell are the nearest neighboring points for that sample
2) For any sample, the nearest sample is determined by the closest Voronoi cell edge

Graphic Depiction

- Decision boundaries are formed by a **subset** of the Voronoi diagram of the training data
- Each line segment is equidistant between two points **of opposite class**.
- The more examples that are stored, the more fragmented and complex the decision boundaries can become.



How to Define Closest Entries

- Distance metrics
  - Euclidean distance (L_2)

$$d(\mathbf{x}, \mathbf{z}) = \left(\sum_{i=1}^{n} |x_i - z_i|^2 \right)^{1/2}$$

  - Manhattan distance (L_1)

$$d(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} |x_i - z_i|$$

  - Minkowski distance (L_p)

$$d(\mathbf{x}, \mathbf{z}) = \left(\sum_{i=1}^{n} |x_i - z_i|^p \right)^{1/p}$$

How to Define Closest Entries

- Distance metric
  - Chebyshev distance

$$d(\mathbf{x}, \mathbf{z}) = \max_{\boldsymbol{i}} |x_i - z_i|$$

  - Natural log distance

$$d(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{n} \ln(1 + |x_i - z_i|)$$

  - Generalized exponential distance

$$d(\mathbf{x}, \mathbf{z}) = e^{-\left(\frac{\|\mathbf{x}-\mathbf{z}\|}{\eta}\right)^{\kappa}}$$

  - Generalized Lorentzian distance

$$d(\mathbf{x}, \mathbf{z}) = \frac{1}{1+\left(\frac{\|\mathbf{x}-\mathbf{z}\|}{\eta}\right)^{\kappa}} \qquad (\kappa = 1,2, \ldots)$$

How to Define Closest Entries

- Camberra:

$$d(\mathbf{x},\mathbf{z}) = \sum_{i=1}^{n} \frac{|x_i - z_i|}{|x_i + z_i|}$$

- Quadratic: (with a problem specific $\boldsymbol{Q}$ matrix)

$$d^2(\mathbf{x},\mathbf{z}) = (\mathbf{x}-\mathbf{z})^T \boldsymbol{Q} (\mathbf{x}-\mathbf{z}) = \sum_{j=1}^{n} \left(\sum_{i=1}^{n} (x_i - z_i) q_{ji} \right) (x_j - z_j)$$

- Mahalanobis:

$$d^2(\mathbf{x},\mathbf{z}) = [\det \boldsymbol{V}]^{1/n} (\mathbf{x}-\mathbf{z})^T \boldsymbol{V}^{-1} (\mathbf{x}-\mathbf{z})$$

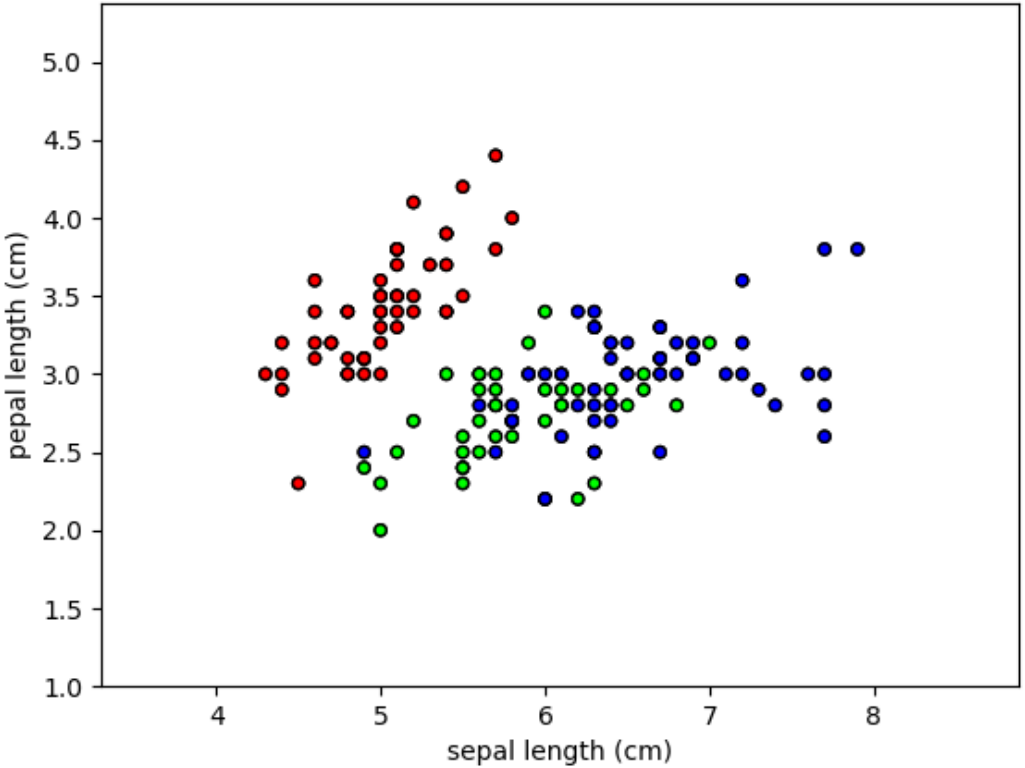
$\boldsymbol{V}$ is the covariance matrix of $\boldsymbol{A_1}, \ldots, \boldsymbol{A_n}$, and $\boldsymbol{A_j}$ is the vector of values for attribute j occurring in the training set instances $1, \ldots, m$

UNIVERSITY OF ARKANSAS

Issues with Distance Metrics

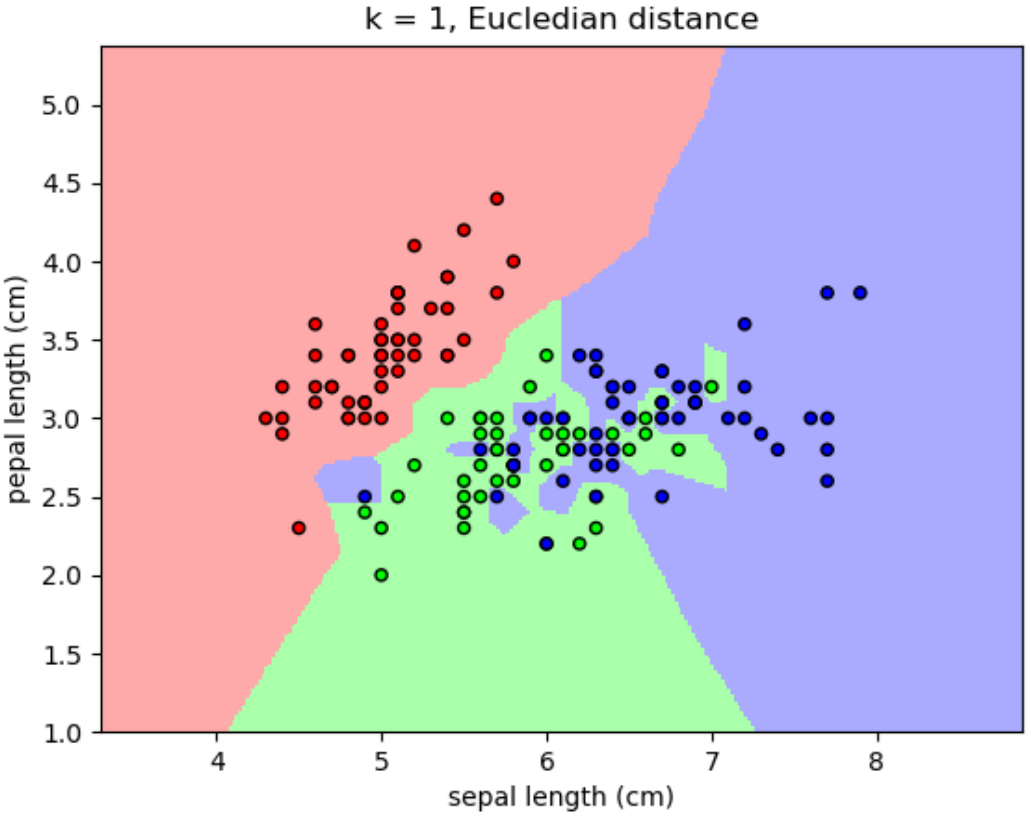
- Most distance measures were designed for linear/real-valued attributes
- Two important questions in the context of machine learning:
  - How to best handle nominal attributes
  - What to do when attribute types are mixed (which ones carry heavier weights)

Use k-NN for Iris Dataset



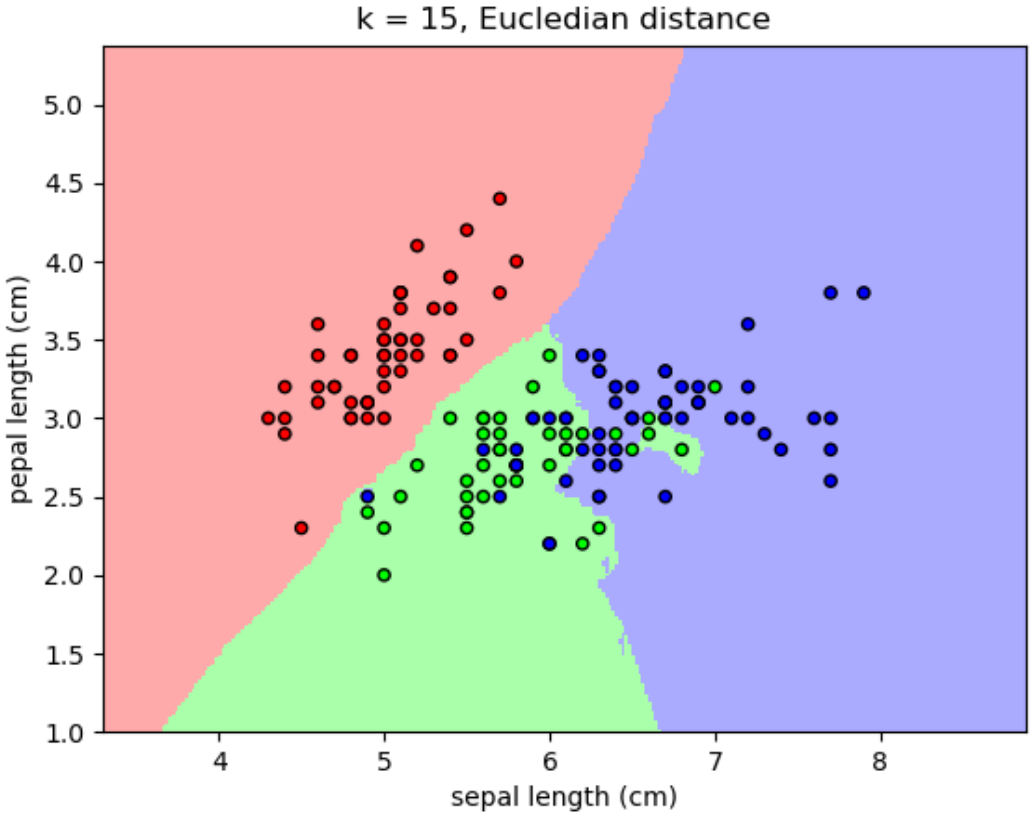
UNIVERSITY OF ARKANSAS

Use k-NN for Iris Dataset



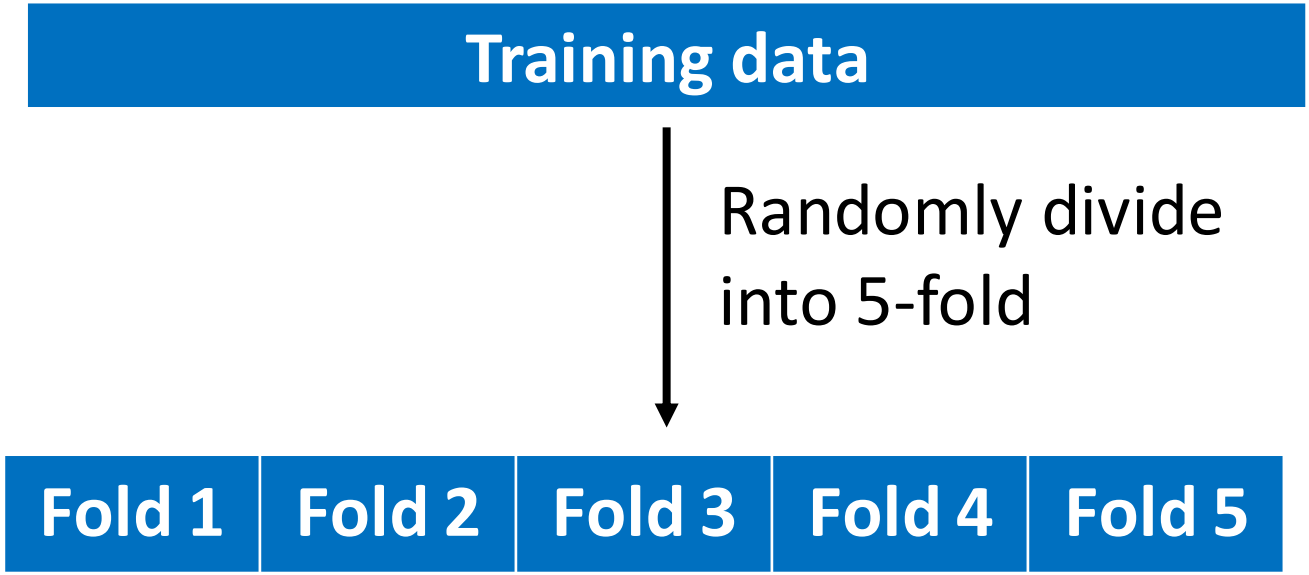
UNIVERSITY OF ARKANSAS

Use k-NN for Iris Dataset



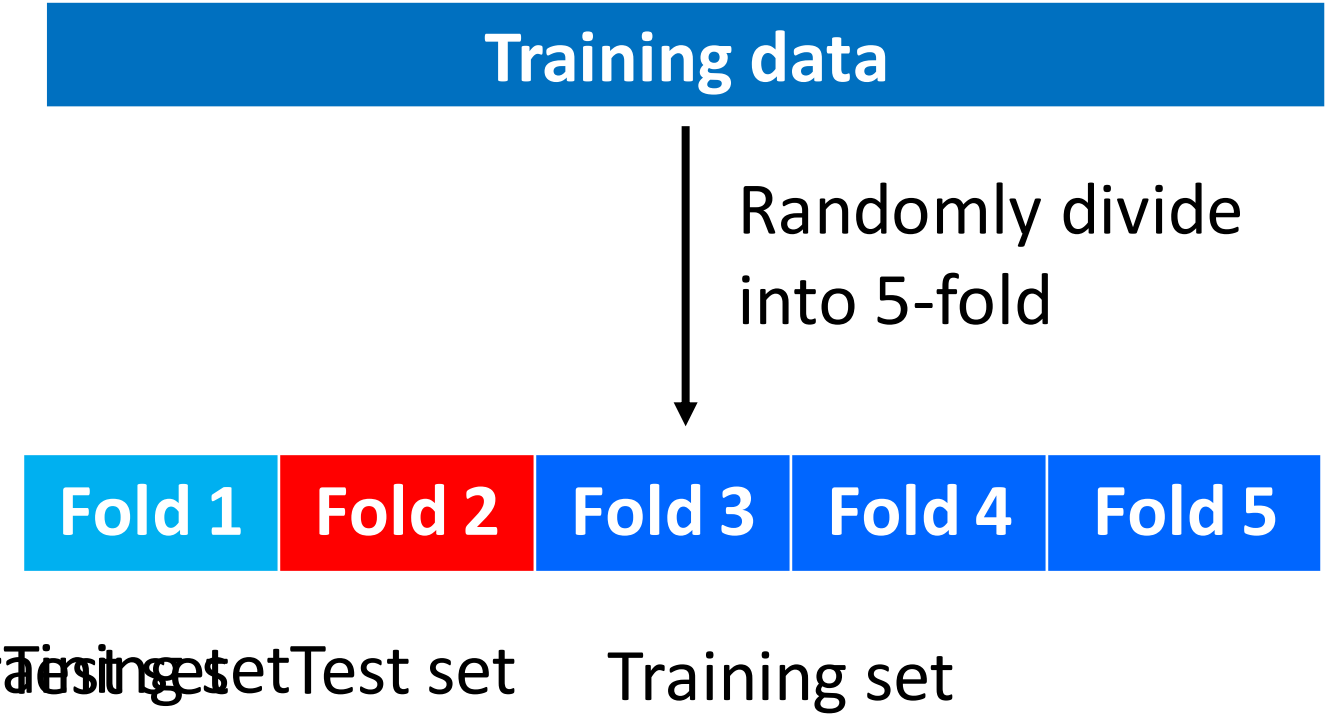
How to Choose k?

- Do cross-validation



How to Choose k?

- Do cross-validation



UNIVERSITY OF ARKANSAS

Pros and Cons

- Pros
  - Simple to understand and easy to implement
  - Zero to little training time (lazy method)
  - No parameters, no need to optimize loss function
  - Quite good accuracy (but other supervised methods are better)
- Cons
  - Computationally expensive
  - Not effective for high-dimension data (use PCA for dimension reduction first)
  - Prediction procedure might be slow
  - Sensitive to the noise (irrelevant data)
  - Memory requirement can be a problem too (Use data structure, like kd-tree)