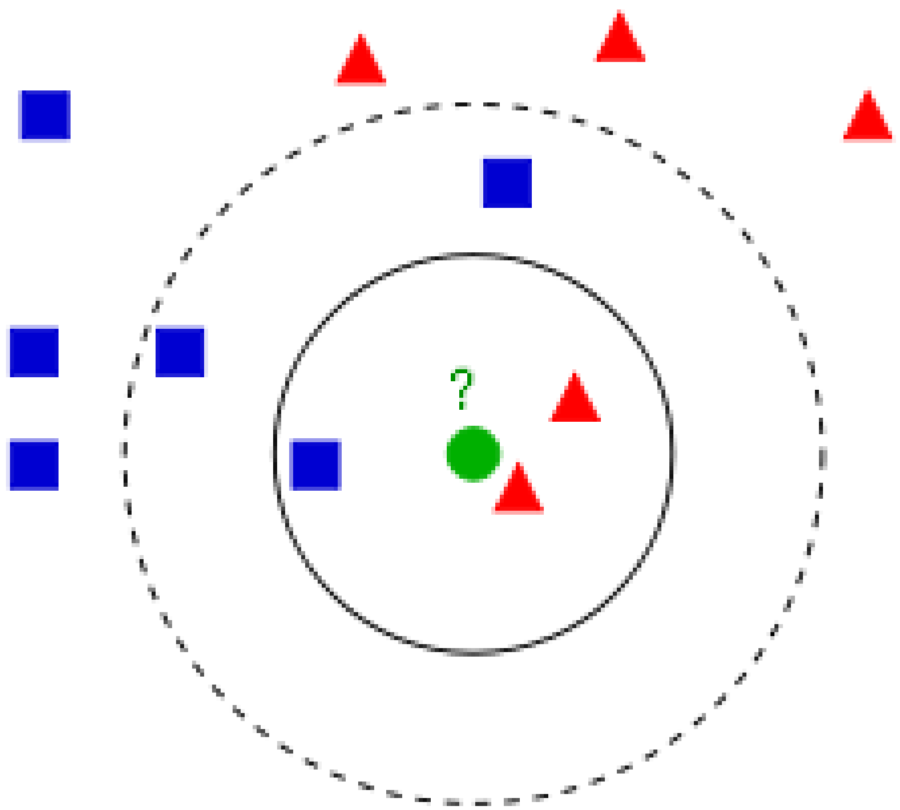
Week 3: K-Nearest Neighbor (KNN)

Outline

- KNN definition
- How it works
- Decision boundaries
- Distance measures
- Effect of K
- Advantages and disadvantages

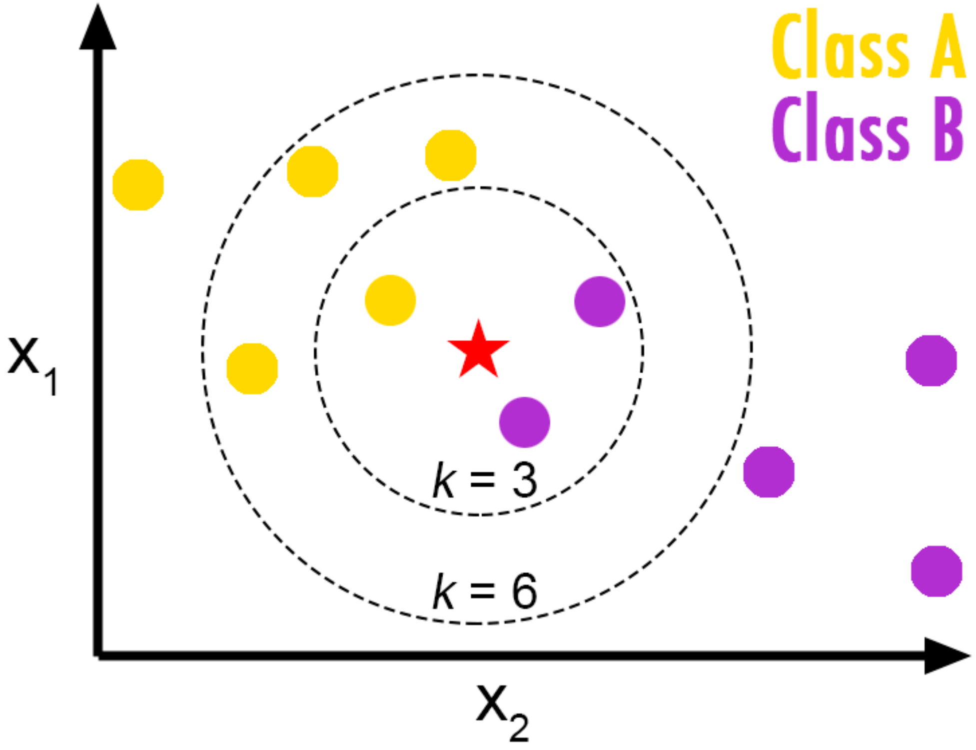
KNN: Definition

- Classifies data based on their similarity with neighbors
- Given a new example **x**, find its closest training example $<x_i, y_i>$ and predict the class label, y



KNN: Definition

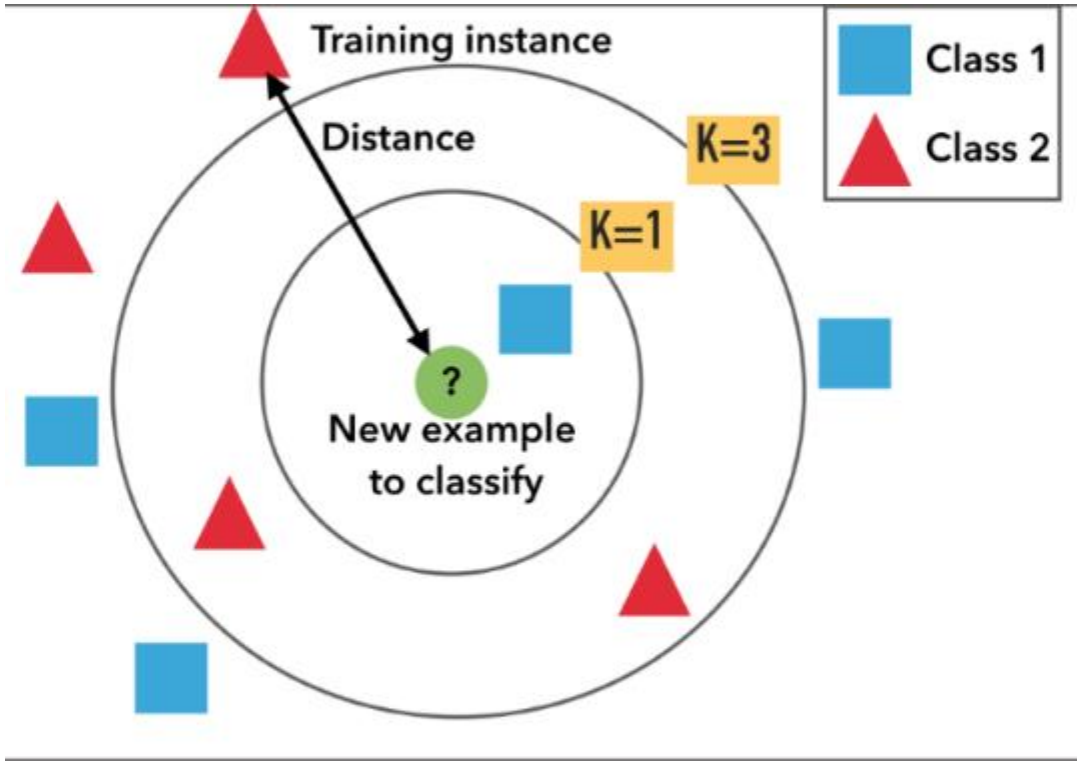
- "K" stands for number of data points that are considered for the classification
- To classify a new input vector x, examine the k-closest training data points to x and assign the object to the most frequently occurring class (predicted class based on majority voting)



Common values for k: 3, 5, 7, …

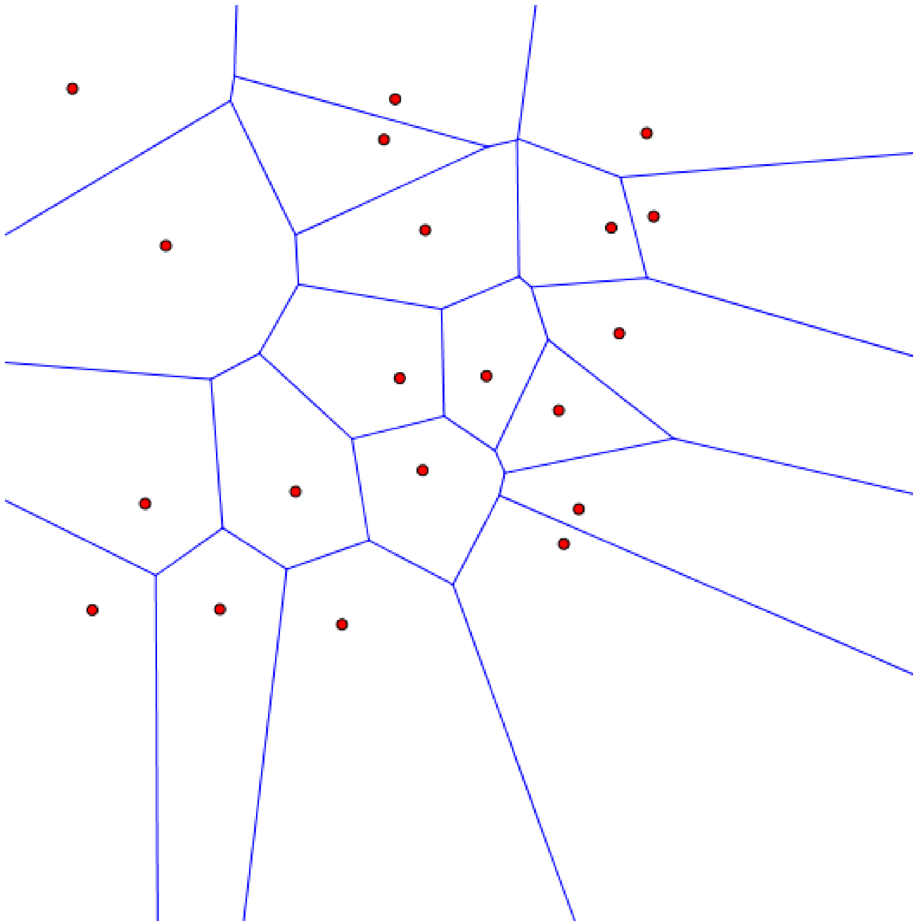
How It Works

- Step 1: Calculate **distance** between the new data point and all the training data points
- Step 2: Pick k training data points closest to the new data point
- Step 3: Calculate average or majority voting to guess label of new data



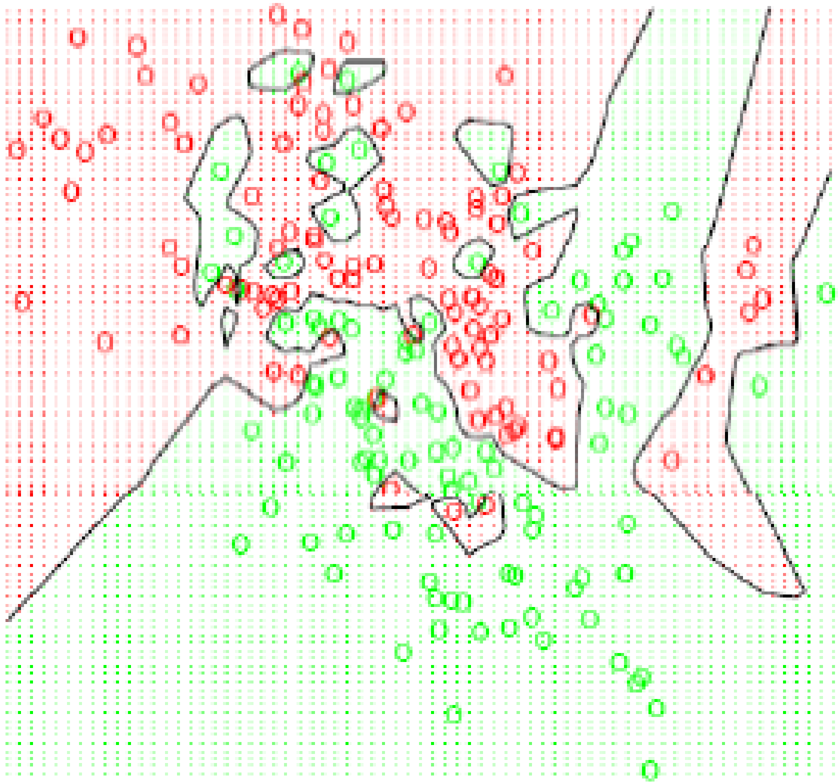
Decision Boundaries

- Given a set of points, a **Voronoi diagram** describes the areas that are nearest to any given point
- Areas can be viewed as **zones of control**
- The more examples stored, the more fragmented and complex the decision boundaries can become



Decision Boundaries

- With large number of examples and possible noise in the labels, the decision boundary can become nasty!
- May end up overfitting the data



Distance Measures

- Euclidean distance (most common)

Given two points **P** = (p_1, p_2,..., p_n) and **Q** = (q_1, q_2,..., q_n)

dist(P, Q) = $\sqrt{(\boldsymbol{p_1} - \boldsymbol{q_1})^2 + (\boldsymbol{p_2} - \boldsymbol{q_2})^2 + \ \ldots + (\boldsymbol{p_n} - \boldsymbol{q_n})^2}$

Example: Given P = (-2, 2) and Q = (2, 5)
Euclidean Distance = dist(P, Q)
= $\sqrt{(-2-2)^2 + (2-5)^2}$
= $\sqrt{(-4)^2 + (-3)^2}$
= $\sqrt{16 + 9}$
= $\sqrt{25}$
= 5

Distance Measures

- Manhattan distance

Given two points **P** = (p_1, p_2,..., p_n) and **Q** = (q_1, q_2,..., q_n)

dist(P, Q) = $|(\boldsymbol{p_1} - \boldsymbol{q_1})| + |\boldsymbol{p_2} - \boldsymbol{q_2}| + \ldots + |\boldsymbol{p_n} - \boldsymbol{q_n}|$

Example: Given P = (1, 2) and Q = (2, 5)
Manhattan Distance = dist(P, Q)
= |1 – 2| + |2 – 5|
= |-1| + |-3|
= 1 + 3
= 4

Class Activity

Lisa has lost gender information of one of her customers, and does not know whether to make a skirt or trousers. She is planning to throw a coin. Can you help her to make a better decision using a KNN classifier?

The customer who is missing gender information: **Gender ?, Waist 28, Hip 34**

Let us use **K = 3** nearest neighbors.

Class Activity

Fill in the table to calculate KNN.

Gender	Waist (cm)	Hip (cm)	Euclidean Distance	Rank minimum distance	Belongs to the neighborhood?
Male	28	32			
Male	33	35			
Female	27	33			
Female	31	36			

Count of male neighborhood members = _____
Count of female neighborhood members = _____
Class based on the majority vote, gender that gets the most votes = _____

Class Activity

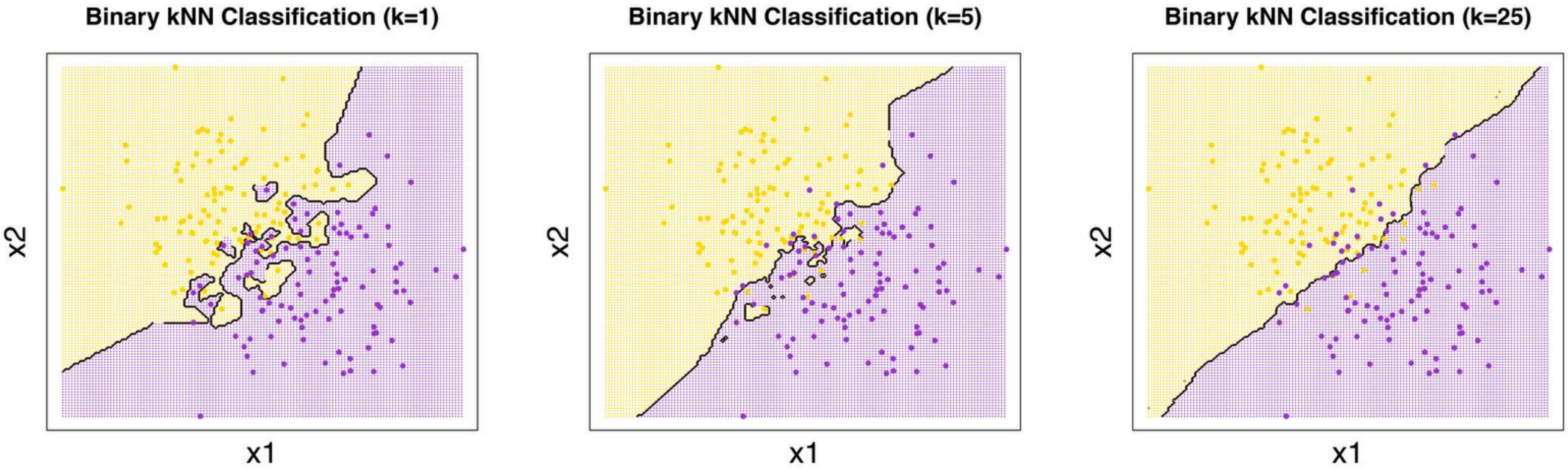
Fill in the table to calculate KNN.

Gender	Waist (cm)	Hip (cm)	Euclidean Distance	Rank minimum distance	Belongs to the neighborhood? (Yes/No)
Male	28	32	2.0	2	Yes
Male	33	35	5.1	4	No
Female	27	33	1.4	1	Yes
Female	31	36	3.6	3	Yes

Count of male neighborhood members = 1
Count of female neighborhood members = 2
Class based on the majority vote, gender that gets the most votes = Female

Effect of K

- What is the impact of K on classification?



Effect of K

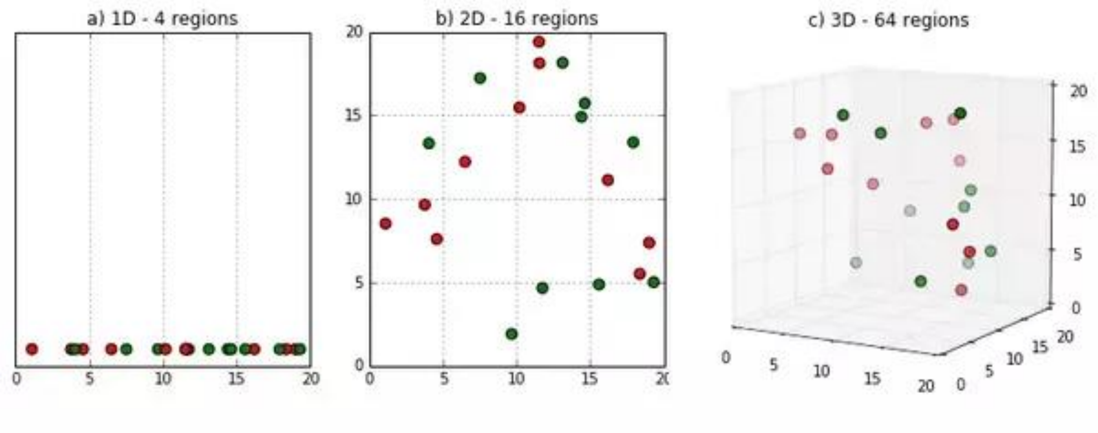
- Larger number of neighbors (K)
- Larger regions
- Smoother class boundaries (reduce impact of noise)

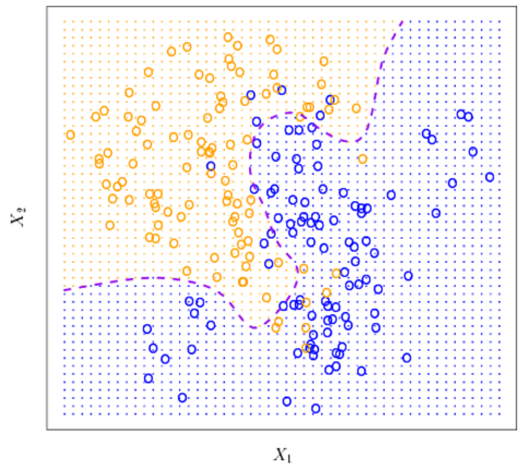
What happens when $K = N$ (all training data points)?

Always predict the majority class!

Problems with KNN

- **Curse of dimensionality**
  - Break down in high-dimensional space (neighborhood becomes very large)
- **Curse of noise**
  - Nearest neighbor is easily misled by noisy/irrelevant features





Advantages

- Can be applied to the data from any distribution
  - Data does not have to be separable with a linear boundary
- Very simple and intuitive
- Good classification if the number of samples is large enough

Disadvantages

- Dependent on K value
- Irrelevant or correlated features have high impact and must be eliminated
- Typically cannot handle high dimensionality
- Computational costs: memory and classification-time computation
  - Test stage is computationally expensive
  - No training stage (all the work is done during the test stage)
- Need large number of samples for accuracy