CS 456 Data Mining

Central Washington University

Today …

- K Nearest Neighbor Classification (KNN)

Outline

- What is KNN?
- How to calculate KNN?
- KNN Examples

What is KNN?

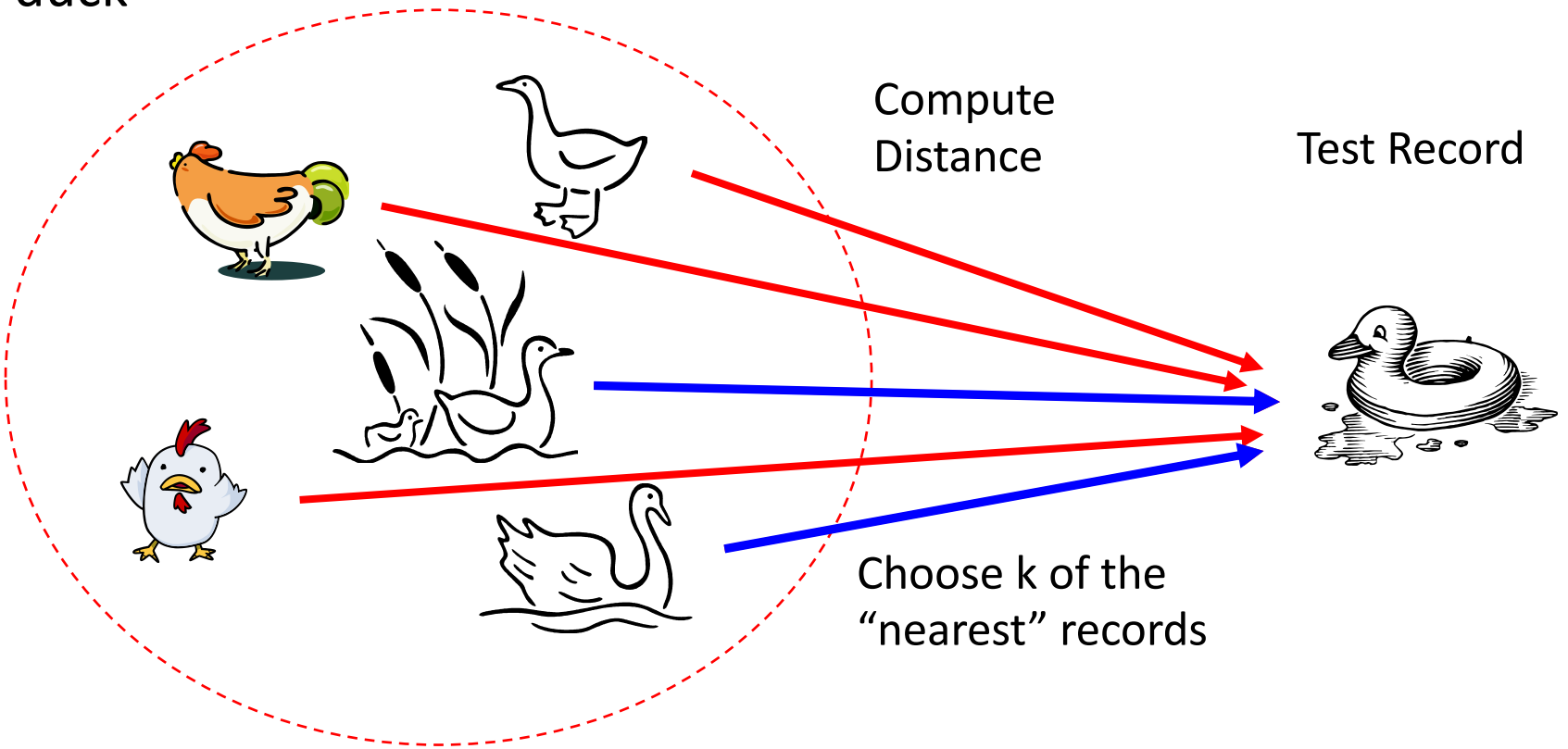
- K-Nearest Neighbors is a simple, versatile, and widely used algorithm in machine learning for both classification and regression tasks.
- KNN operates on the principle that similar data points are likely to have similar outcomes or belong to the same class.
- In KNN, the 'K' represents the number of nearest neighbors to consider when making predictions.

What is KNN?

- For classification, KNN assigns a class to a new data point based on the majority class among its 'K' nearest neighbors.
- For regression, it predicts a value based on the average of the values of its 'K' nearest neighbors.
- The algorithm's simplicity lies in its direct approach of learning from the training data without building an explicit model – it simply stores the data and makes predictions using a distance metric, typically Euclidean distance.

Nearest Neighbor Classifiers

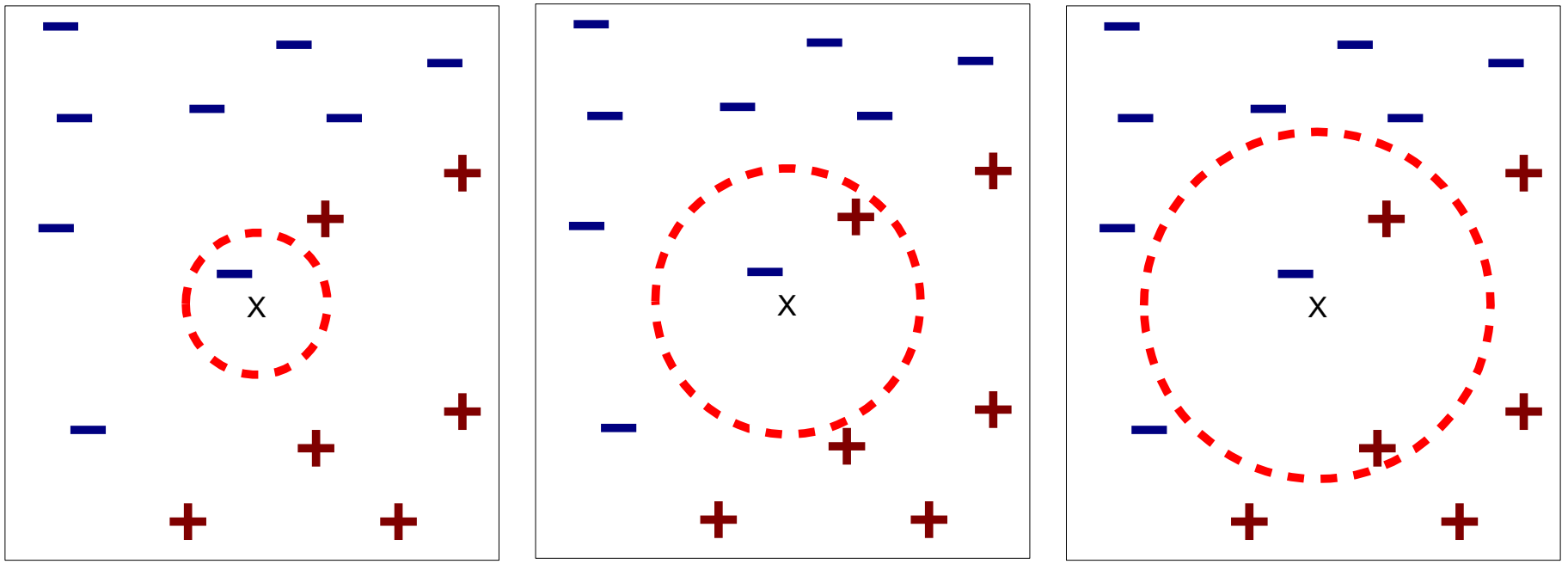
- Basic idea:
  - If it walks like a duck, quacks like a duck, then it's probably a duck



Basic Idea

- *k*-NN classification rule is to assign to a test sample the majority category label of its *k* nearest training samples
- In practice, *k* is usually chosen to be odd, so as to avoid ties
- The *k* = 1 rule is generally called the nearest-neighbor classification rule

Definition of Nearest Neighbor



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x are data points that have the k smallest distance to x

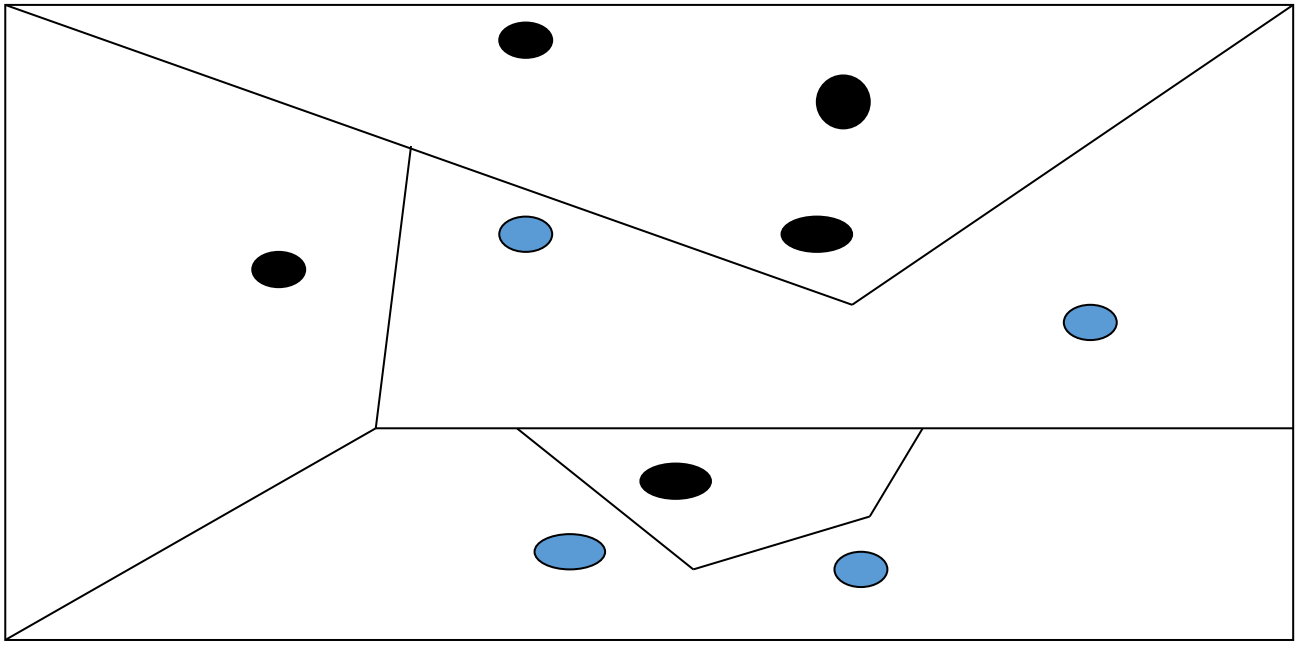
K-Nearest Neighbor

- An arbitrary instance is represented by$(a_1(x), a_2(x), a_3(x),.., a_n(x))$
  - $a_i(x)$ denotes features
- Euclidean distance between two instances

  $d(x_i, x_j)=\text{sqrt}\ (\text{sum for r=1 to n}\ (a_r(x_i) - a_r(x_j))^2)$
- Continuous valued target function
  - mean value of the k nearest training examples

Voronoi Diagram

- Decision surface formed by the training examples



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

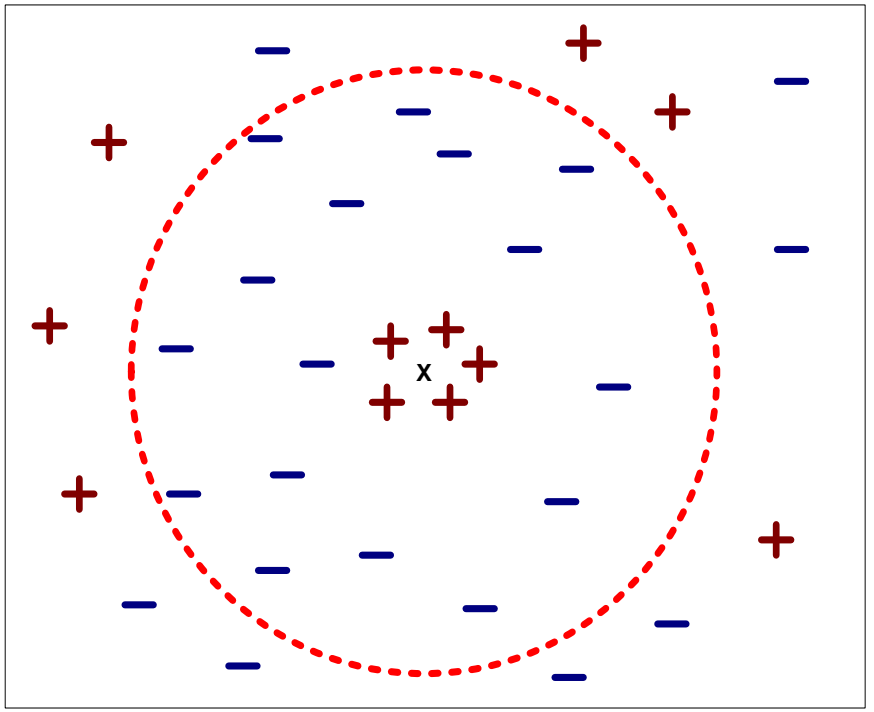
Nearest-Neighbor Classifiers: Issues

- The value of *k*, the number of nearest neighbors to retrieve
- Choice of Distance Metric to compute distance between records
- Computational complexity
  - Size of training set
  - Dimension of data

Value of K

- Choosing the value of k:
  - If k is too small, sensitive to noise points
  - If k is too large, neighborhood may include points from other classes

Rule of thumb:
K = sqrt(N)
N: number of training points



Distance Measure: Scale Effects

- Different features may have different measurement scales
  - E.g., patient weight in kg (range [50,200]) vs. blood protein values in ng/dL (range [-3,3])
- Consequences
  - Patient weight will have a much greater influence on the distance between samples
  - May bias the performance of the classifier

Standardization

- Transform raw feature values into z-scores

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

  - x_{ij} is the value for the i^{th} sample and j^{th} feature
  - μ_j is the average of all x_{ij} for feature j
  - σ_j is the standard deviation of all x_{ij} over all input samples

- Range and scale of z-scores should be similar (providing distributions of raw feature values are alike)

Summary of KNN

Advantages of KNN

1.Simplicity and Intuitiveness: KNN is easy to understand and implement, making it a good starting point for classification and regression problems.
2.No Model Training Required: KNN is a non-parametric method, meaning it doesn't assume anything about the underlying data distribution and doesn't require explicit training phase, which can be beneficial in scenarios with real-time data.
3.Versatility: It can be used for both classification and regression tasks and works well with multi-modal classes.
4.Adaptability: KNN can handle any number of classes and is naturally suited for multi-class classification.

Summary of KNN

Disadvantages of KNN

1.Scalability and Efficiency: KNN can be computationally expensive, especially with large datasets, as it requires storing all training data and calculating distances for each query.
2.Performance with High Dimensionality: Its performance degrades with high-dimensional data (curse of dimensionality) due to the difficulty in calculating distance in many dimensions.
3.Sensitivity to Imbalanced Data: KNN can be biased towards the more frequent classes in imbalanced datasets.
4.Choosing 'K': Selecting the right value of 'K' is crucial and can be challenging, as a small value of 'K' can be noisy and subject to the effects of outliers, while a large 'K' makes boundaries between classes less distinct.