CDS503: Machine Learning

Topic 3

Classification Non-Parametric - KNN

Assoc. Prof. Dr UMI KALSOM YUSOF
SCHOOL OF COMPUTER SCIENCES
UNIVERSITI SAINS MALAYSIA (USM)

Contents

- Background of K-Nearest Neighbors
- Classifying with K-Nearest Neighbors
- Definition
- Distance Functions
- Choosing K

K-Nearest Neighbour

Some slides are based on K-NEAREST NEIGHBOR CLASSIFIER by Ajay Krishna Teja Kavuri Of West Virginia University and **Algorithms: K Nearest Neighbors by** Tilani Gunawardena

Simple Analogy

- Tell me about your friends(*who your neighbors are*) and *I will tell you who you are*.

KNN – Different names

- K-Nearest Neighbors
- Memory-Based Reasoning
- Example-Based Reasoning
- Instance-Based Learning
- Lazy Learning

What is KNN?

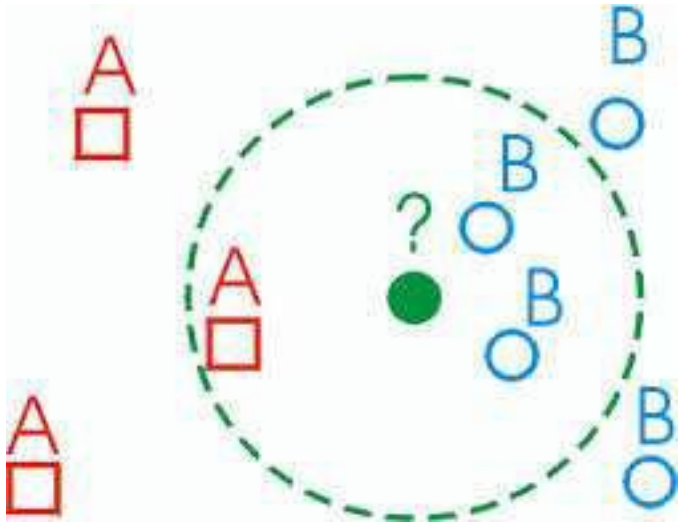
- A powerful classification algorithm used in pattern recognition.
- K nearest neighbors stores all available cases and classifies new cases based on a *similarity measure* (e.g distance function)
- One of the top data mining algorithms used today.
- A non-parametric lazy learning algorithm (An Instance-based Learning method).

Why is the KNN algorithm lazy?

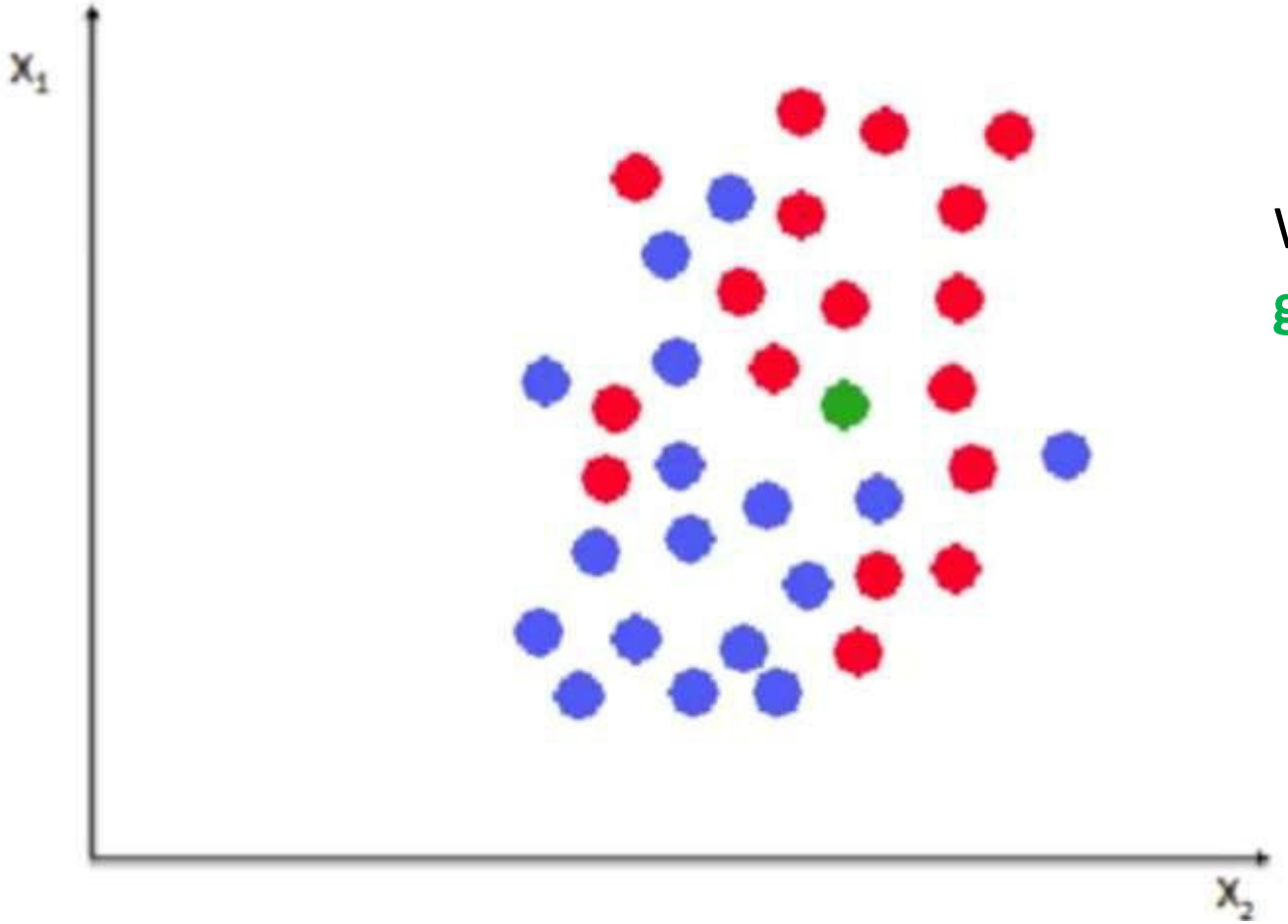
- KNN is considered lazy because no abstraction occurs.
  - The abstraction and generalization processes are not part of it.
    - Using a strict definition of learning, in which the learner summarizes raw input into a model (equations, decision trees, clustering, if then rules), a lazy learner is not really learning anything.
  - Instead, it is only storing the training data, which takes very little time. Classification, however, is very slow.
    - This is unlike most classifiers in which training takes a long time, but classification is very fast.
- Lazy learning is known as an instance-based learning.
- An instance-based learners do not build a model, the method is said to be in a class of non-parametric learning methods- in that no parameters are learnt about the data.

KNN: Classification Approach

- An object (a new instance) is classified by a majority votes for its neighbor classes.
- The object is assigned to the most common class amongst its K nearest neighbors.(*measured by a distant function*)



KNN: Example



What do you think the **green** ball belongs to?

ORIGIN OF K-NN

- Nearest Neighbors have been used in statistical estimation and pattern recognition already in the beginning of 1970's (non-parametric techniques).
- The method prevailed in several disciplines and still it is one of the top 10 Data Mining algorithm.

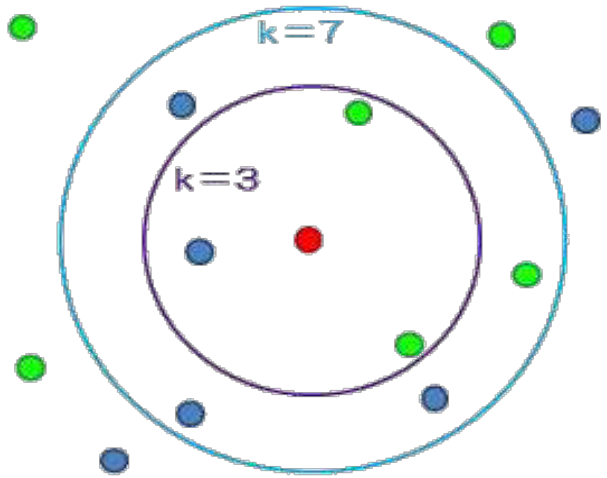
IN A SENTENCE K-NN IS …

- It's how people judge by observing **our peers**.

- We tend to move with people of similar attributes so does data.

DEFINITION

- K-Nearest Neighbor is considered a lazy learning algorithm that classifies data sets based on their similarity with neighbors.



"K" stands for number of data set items that are considered for the classification.

Ex: Image shows classification for different k-values.

TECHNICALLY…..

- For the given attributes $A=\{X_1, X_2 \ldots. X_D\}$ Where D is the dimension of the data, we need to predict the corresponding classification group $G=\{Y_1, Y_2 \ldots Y_n\}$ using the proximity metric over K items in D dimension that defines the closeness of association such that $X \in R^D$ and $Y_p \in G$.

THAT IS….

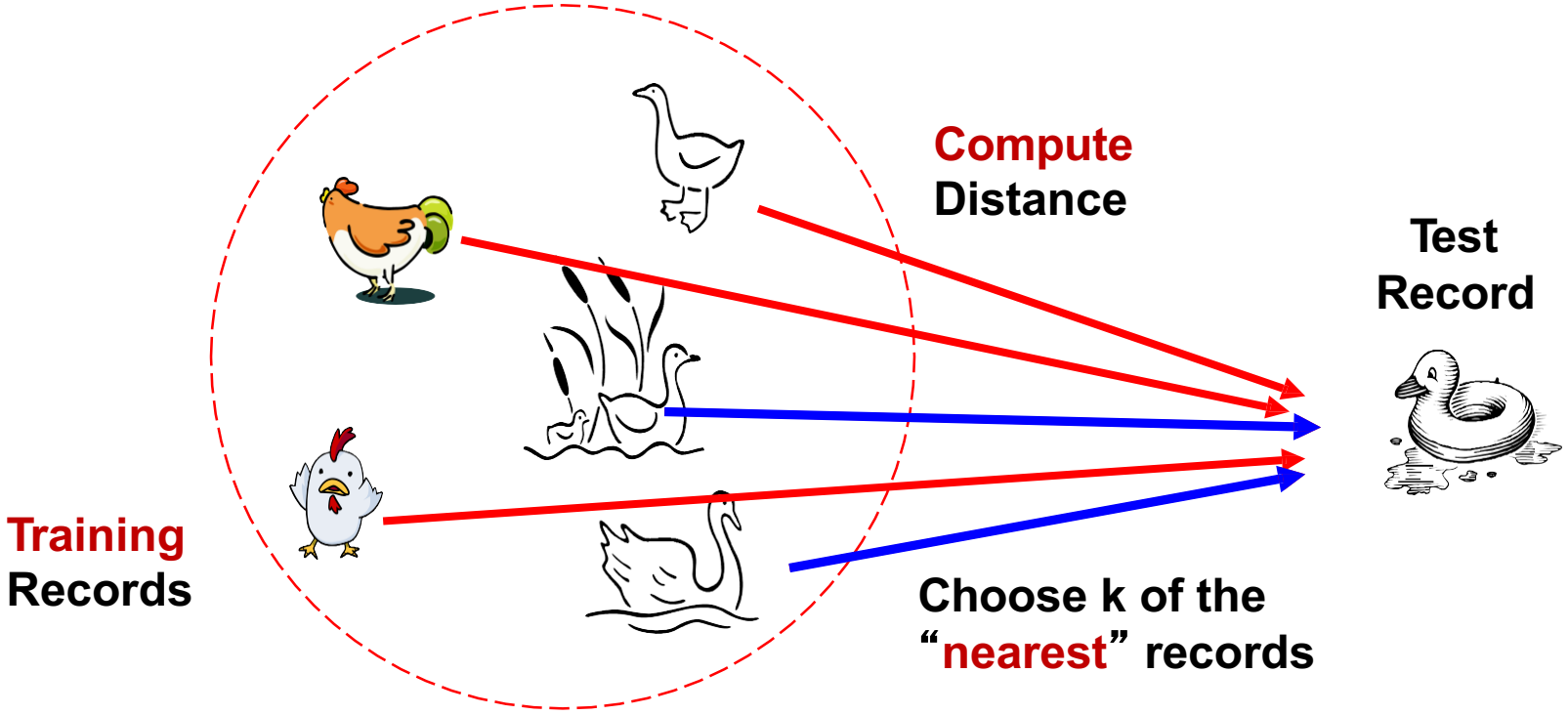
- Attribute A={Color, Outline, Dot}
- Classification Group, G={triangle, square}
- D=3, we are free to choose K value.

Attributes A

#	Attribute			Shape
	Color	Outline	Dot	
1	green	dashed	no	triange
2	green	dashed	yes	triange
3	yellow	dashed	no	square
4	red	dashed	no	square
5	red	solid	no	square
6	red	solid	yes	triange
7	green	solid	no	square
8	green	dashed	no	triange
9	yellow	solid	yes	square
10	red	solid	no	square
11	green	solid	yes	square
12	yellow	dashed	yes	square
13	yellow	solid	no	square
14	red	dashed	yes	triange

Classification Group

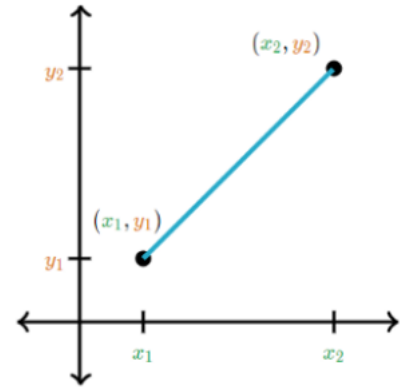
Distance Measure



Distance Between Neighbors

- Calculate the distance between new example (E) and all examples in the training set.

- *Euclidean* distance between two examples.
  - X = [x1,x2,x3,..,xn]
  - Y = [y1,y2,y3,...,yn]
  - The Euclidean distance between *X* and *Y* is defined as:



$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

K-Nearest Neighbor Algorithm

- All the instances correspond to points in an n-dimensional feature space.
- Each instance is represented with a set of numerical attributes.
- Each of the training data consists of a set of vectors and a class label associated with each vector.
- Classification is done by comparing feature vectors of different K nearest points.
- Select the K-nearest examples to E in the training set.
- Assign E to the most common class among its K-nearest neighbors.

KNN: Example

Customer	Age	Income	No Credit Cards	Class	Distance from John
George	35	35K	3	No	sqrt $[(35-37)^2+(35-50)^2+(3-2)^2]=15.16$
Rachel	22	50k	2	Yes	sqrt $[(22-37)^2+(50-50)^2+(2-2)^2]=15$
Steve	63	200k	1	No	sqrt $[(63-37)^2+(200-50)^2+(1-2)^2=152.23$
Tom	59	170k	1	No	sqrt $[(59-37)^2+(170-50)^2+(1-2)^2]=122$
Anne	25	40k	4	Yes	sqrt $[(25-37)^2+(40-50)^2+(4-2)^2]=15.74$
John	37	50k	2		

Determine a **Class of John**, given k = 3

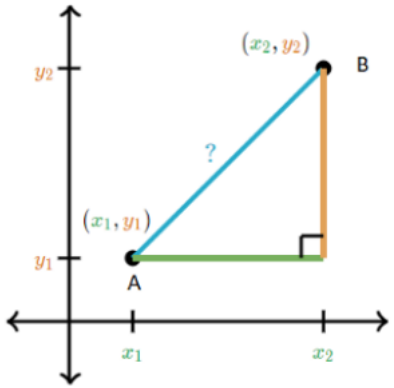
PROXIMITY METRIC

- Definition: Also termed as "Similarity Measure" quantifies the association among different items.
- Following is a table of measures for different data items:

Similarity Measure	Data Format
Contingency Table, Jaccard coefficient, Distance Measure	Binary
Z-Score, Min-Max Normalization, Distance Measures	Numeric
Cosine Similarity, Dot Product	Vectors

PROXIMITY METRIC

- For the numeric data let us consider some distance measures:
  - Manhattan Distance:



$$X = \langle x_1, x_2, \cdots, x_n \rangle \quad Y = \langle y_1, y_2, \cdots, y_n \rangle$$

$$dist(X,Y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|$$

Ex: Given X = {1,2} & Y = {2,5}

Manhattan Distance = dist(X,Y) = |1-2|+|2-5|

= 1+3

= 4

PROXIMITY METRIC

Another method for distance measures:

- Euclidean Distance:

$$X = \langle x_1, x_2, \cdots, x_n \rangle \qquad Y = \langle y_1, y_2, \cdots, y_n \rangle$$

$$dist((x, y), (a, b)) = \sqrt{(x - a)^2 + (y - b)^2}$$

Ex: Given X = {-2,2} & Y = {2,5}

Euclidean Distance = dist(X,Y) = $(-2-2)^2 + (2-5)^2)$

= dist(X,Y) = $(-4)^2 + (-3)^2$

= dist(X,Y) = **16 +9**

= dist(X,Y) = **25**

= dist(X,Y)= 5

Example of KNN

KNN IN ACTION

- Consider the following data: A={weight, color} G={Apple(A), Banana(B)}

- We need to predict the type of a fruit with: weight = 378, color = red

weight (g)	color	Type of fruit
303	3	Banana
370	1	Apple
298	3	Banana
277	3	Banana
377	4	Apple
299	3	Banana
382	1	Apple
374	4	Apple
303	4	Banana
309	3	Banana
359	1	Apple
366	1	Apple
311	3	Banana
302	3	Banana
373	4	Apple
305	3	Banana
371	3	Apple

Example of KNN

SOME PROCESSING….

- Assign color codes to convert into numerical data:

red	1
orange	2
yellow	3
green	4
blue	5
purple	6

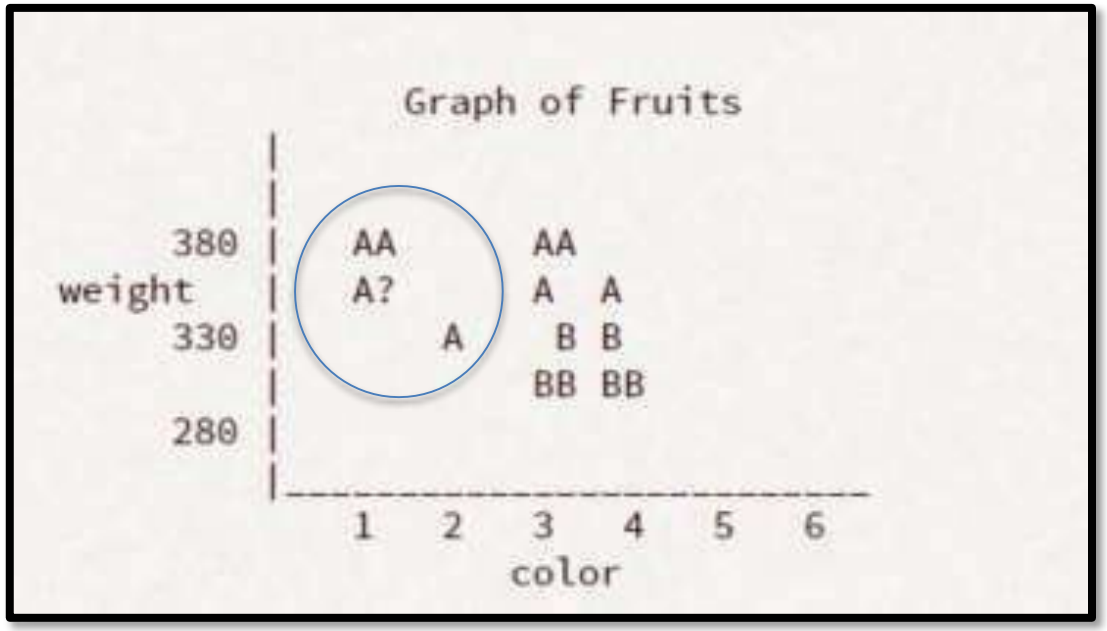
- Let's label Apple as "A" and Banana as "B"

Example of KNN

PLOTTING

- Using K=3, Our result will be,

- Using K=5, What will the result be?



Class Activity

Which group the fruit belongs to?

- Consider the following data: A={weight, size}
  G={Apple(A), Mangosteen (M)}

- A neighbour gives a fruit to me. However, the fruit is wrapped nicely in a white, soft wrapping paper. Please help me to predict the type of the fruit with:
  - Weight: 373 g,
  - Size = 4 cm
  - Let us use ***k* = 3** nearest neighbors.

Class Activity

Which group the fruit belongs to?

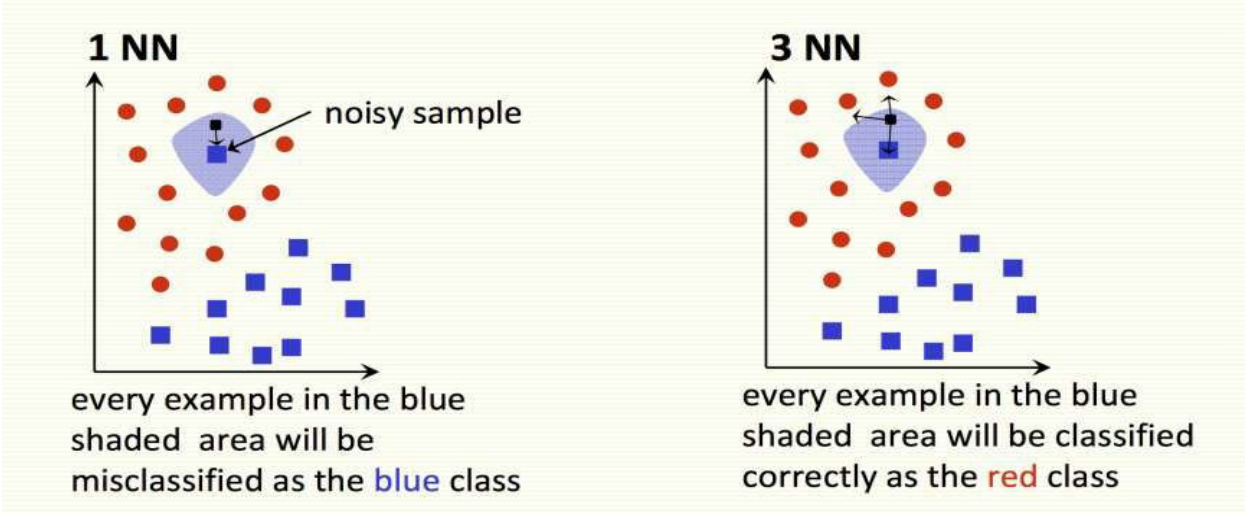
Fill in the table to calculate KNN.

Fruit Type	Weight (g)	Size (cm)	Euclidean Distance	Rank Minimum Distance	Belongs to the neighborhood?
Mangosteen	303	4			
Apple	378	5			
Mangosteen	298	3			
Mangosteen	277	4			
Apple	377	6			

Count of mangosteen neighborhood members = _____
Count of apple neighborhood members = _____
Class based on the majority vote, fruit that gets the most votes = _____

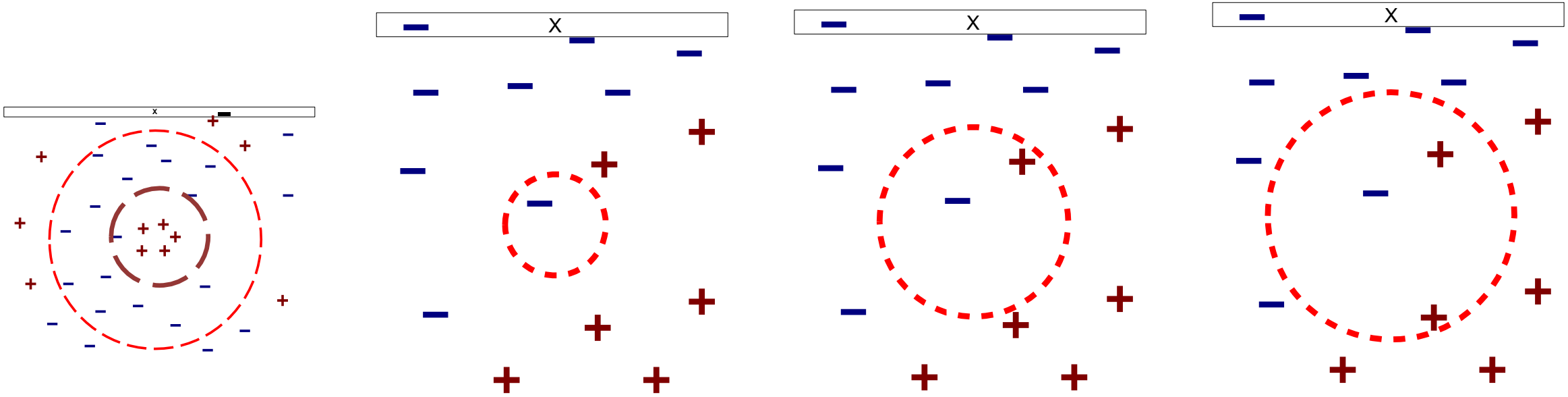
How to choose K?

- If K is too small it is **sensitive** to noise points.
- Larger K works well. But too large K may include majority points from other classes.



- Rule of thumb is K < sqrt(n), n is number of examples.

The affect of K



(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

How to Choose k?

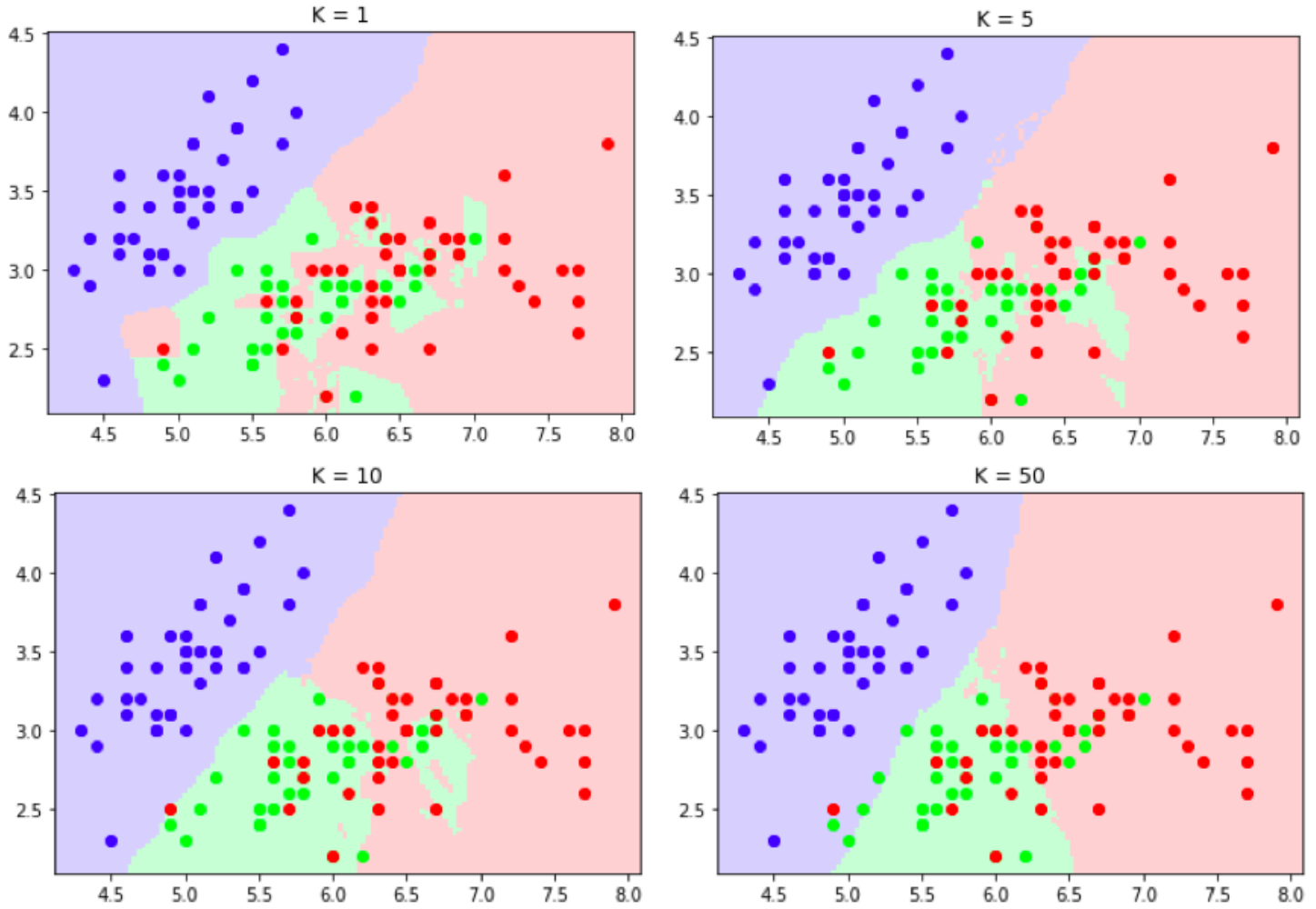
- When *k* is small, single instances matter;
  bias is small, variance is large (overfitting):
  High complexity
- As *k* increases, we average over more instances and variance decreases but bias increases (underfitting):
  Low complexity
- Cross-validation is used to fine tune *k*.

*The **bias** is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs.*
*The **variance** is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs.*

Choosing K

AS K VARIES ….

- Clearly, k has an impact on the classification



K-NN PROPERTIES

- K-NN is a lazy algorithm
- The processing defers with respect to k value.
- Result is generated after analysis of stored data.
- It neglects any intermediate values.

KNN: Advantages and Disadvantages

Advantages

- Can be applied to the data from any distribution
  - for example, 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 – maybe tricky
- Test stage is computationally expensive
- No training stage, all the work is done during the test stage
- This is actually the opposite of what we want.
  - Usually we can afford training step to take a long time, but we want fast test step
- Need large number of samples for accuracy

Thank you