جامعة الطائف
TAIF UNIVERSITY

DATA MINING

Assoc. Prof. Dr. Salha Alzahrani

College of Computers and Information Technology
Taif University
Saudi Arabia

s.zahrani@tu.edu.sa

Classification | Classification using Naïve Bayes Algorithm

Recap of Previous Lecture

Content of This Lecture

Summary & Checklist

Recap of Lecture 3

- What is Classification?
- Naïve Bayes Classifiers
- Probability of an event
- The train example
- The prior probability
- The conditional (or posterior) probability
- Naïve Bayes Algorithm
- Naïve Bayes Algorithm: The train example
- Naïve Bayes Algorithm: classification of unseen instance
- Naïve Bayes Algorithm: summary of steps
- Self-assessment Exercise.

Recap of Previous Lecture
Content of This Lecture
Summary & Checklist

Content of Lecture 4

- Introduction
- Nearest instance
- k-Nearest Neighbour Classification
- Example of classification using Nearest Neighbour Algorithm
- Distance Measures
- Distance Measures: Euclidean
- Distance Measures: Manhattan
- Distance Measures: Maximum Dimension
- Nearest Neighbour Algorithm: Step-by-Step
- Self-Assessment Exercise
- Summary & Checklist

Classification | Introduction

- Nearest Neighbour classification is mainly used when all attribute values are **continuous**.
- The idea is to estimate the classification of an unseen instance using the classification of the instance or instances that are ***closest*** to it,

Classification | Nearest instance

- Supposing we have a training set with just two instances:

a	b	c	d	e	f	Class
yes	no	no	6.4	8.3	low	negative
yes	yes	yes	18.2	4.7	high	positive

- **What is the classification for the following instance?**

yes	no	no	6.6	8.0	low	???

- It seems that the unseen instance is ***nearer*** **to** the first instance than to the second.
- In the absence of any other information, we could predict its classification as 'negative'.

Classification | k-Nearest Neighbour Classification

- It is usual to base the classification on those of the *k* nearest neighbours (where *k* is a small integer such as 3 or 5), not just the nearest one. The method is then known as ***k-Nearest Neighbour*** or just ***k-NN classification***

***k*-Nearest Neighbour Classification** A method of classifying an **unseen instance** using the **classification** of the **instance** or instances closest to it.

Basic *k*-Nearest Neighbour Classification Algorithm

– Find the k training instances that are closest to the unseen instance.

– Take the most commonly occurring classification for these k instances.

Figure 2.4 The Basic k-Nearest Neighbour Classification Algorithm

Classification | Example of classification using Nearest Neighbour Algorithm

- Given a training set with 20 instances, each giving the values of **two attributes** and an **associated classification**.
- How can we estimate the classification for an 'unseen' instance where the first and second attributes are 9.1 and 11.0, respectively**?**

Attribute 1	Attribute 2	Class
0.8	6.3	−
1.4	8.1	−
2.1	7.4	−
2.6	14.3	+
6.8	12.6	−
8.8	9.8	+
9.2	11.6	−
10.8	9.6	+
11.8	9.9	+
12.4	6.5	+
12.8	1.1	−
14.0	19.9	−
14.2	18.5	−
15.6	17.4	−
15.8	12.2	−
16.6	6.7	+
17.4	4.5	+
18.2	6.9	+
19.0	3.4	−
19.6	11.1	+

Figure 2.5 Training Set for k-Nearest Neighbour Example

Classification | Example of classification using Nearest Neighbour Algorithm

- For this small number of attributes, we can represent the training set as **20 points** on a **two-dimensional graph**.
- Each point is labelled with **+** or **–** symbol to indicate its classification.
- The five nearest neighbours are labelled with three + signs and two – signs, so a basic *5-NN* classifier would classify the unseen instance as 'positive' **by majority voting**.

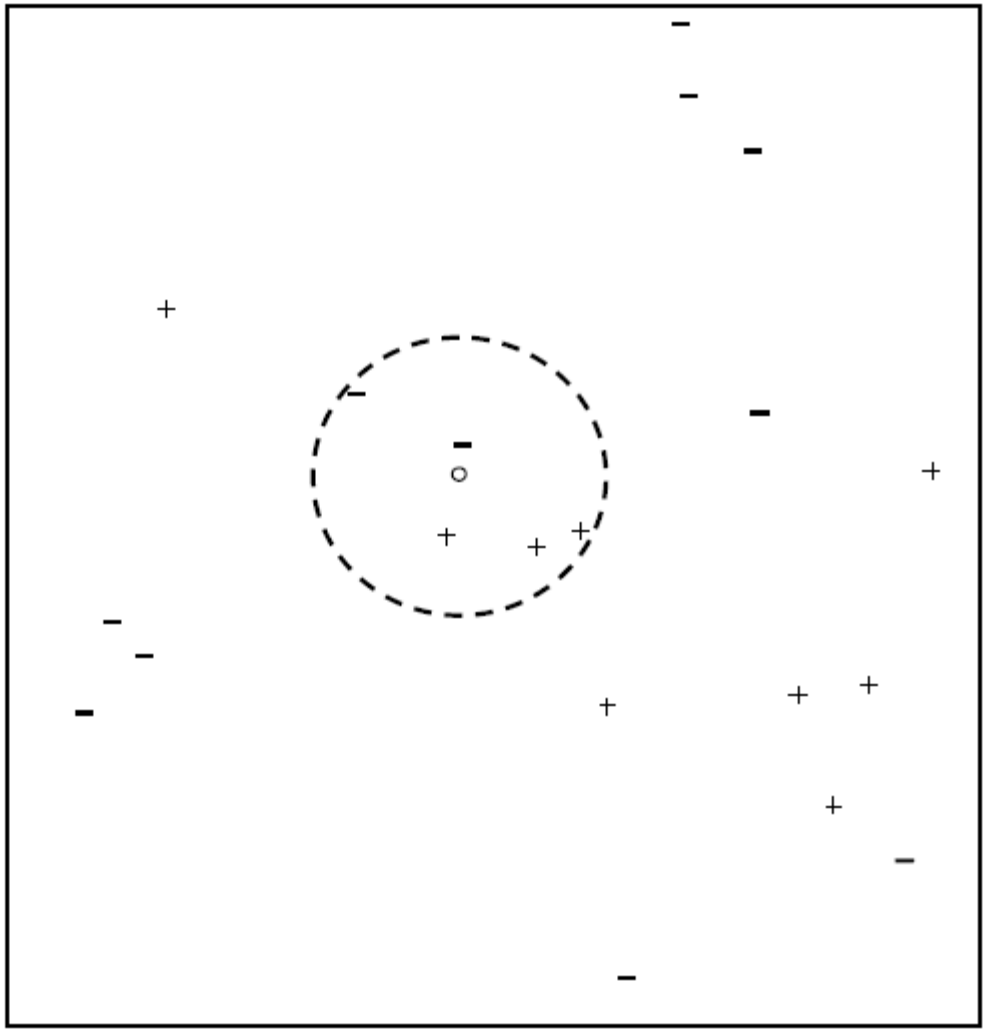

Figure 2.6 Two-dimensional Representation of Training Data

Classification | Distance Measures

- As the number of dimensions (attributes) increases, it becomes impossible to visualise them on 2D graph.

- So, what should we use?

- **Distance Measures**
- There are many possible ways of measuring the distance between two instances with ***n* attribute** values

Distance Measure A means of measuring the similarity between two **instances**. The smaller the value, the greater the similarity

Classification | Distance Measures: Euclidean Distance

- The most popular distance measure is the ***Euclidean Distance***
- Euclidean distance formula in two dimensions:

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

- If we denote an instance in the training set by (a_1, a_2) and the unseen instance by (b_1, b_2), the length of the straight line joining the two points.

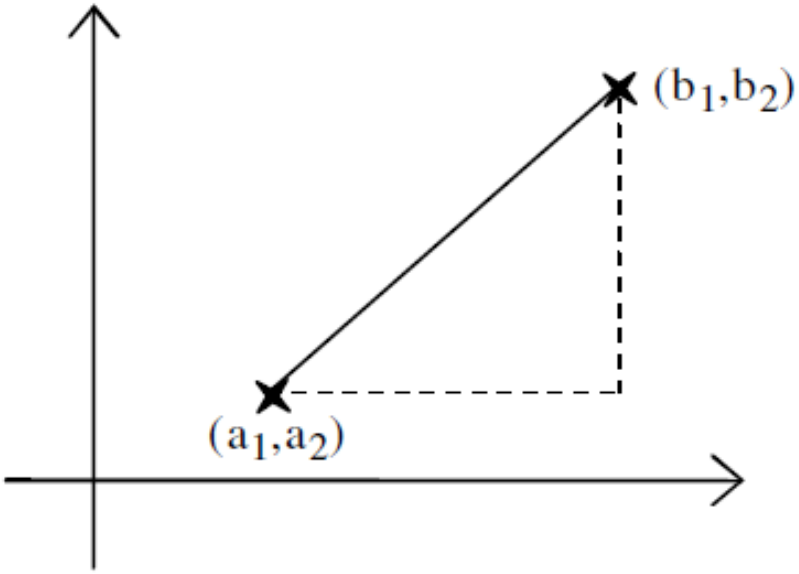

Figure 2.8 Example of Euclidean Distance

Classification | Distance Measures: Euclidean Distance

- If there are two points (a_1, a_2, a_3) and (b_1, b_2, b_3) in a three-dimensional space, Euclidean distance formula is

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2}$$

- In general, the formula for Euclidean distance between points (a_1, a_2, . . . , a_n) and (b_1, b_2, . . . , b_n) in n-dimensional space is:

$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_n - b_n)^2}$$

Classification | Distance Measures: Manhattan Distance or City Block Distance

- Another measure is called ***Manhattan Distance*** or ***City Block Distance***.

- For example, if you are travelling around a city such as Manhattan, you cannot (usually) go straight from one place to another but only by moving along streets aligned horizontally and vertically.

- Example: Manhattan distance between the points (4, 2) and (12, 9) is
  $(12 - 4) + (9 - 2) = 8 + 7 = 15$.

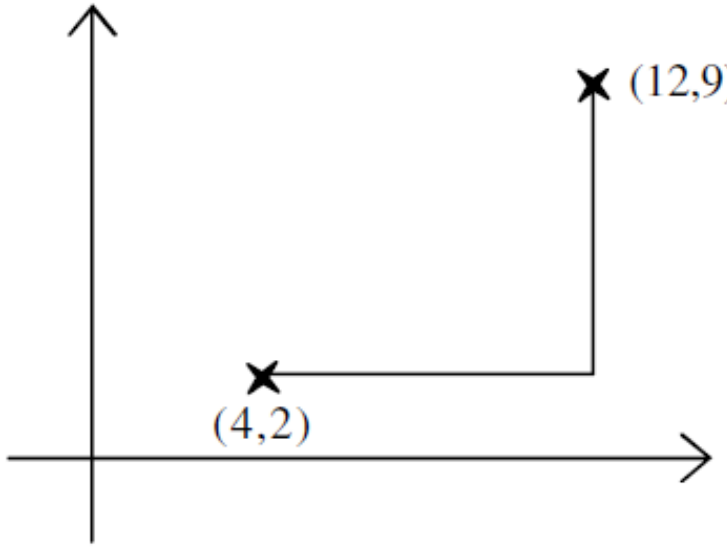

Figure 2.9 Example of City Block Distance

Classification | Distance Measures: Maximum Dimension Distance

- A third possibility is the ***maximum dimension distance***. This is the largest absolute difference between any pair of corresponding attribute values.
- Note: absolute difference is the difference converted to a positive number if it is negative.
- For example, the maximum dimension distance between the instances below is 12.4 − (−7.1) = 19.5.

6.2	−7.1	−5.0	18.3	−3.1	8.9

8.3	12.4	−4.1	19.7	−6.2	12.4

Classification | Nearest Neighbour Algorithm: Step-by-Step

Nearest Neighbour Algorithm

Step 1: Define the value of k, where k can be any value 3, 5, 7 etc.

Step 2: Calculate the similarity between the unseen/unclassified instance and each instance in the training using one of the distance measures:

- Euclidean distance
- Manhattan distance
- Maximum dimension distance

Step 3: Find the most k nearest instances to the unseen instance.

Step 4: Use the classification that is used by the majority of nearest instances as a classification for the unseen instance.

Classification | Self-Assessment Exercise

Using the training set shown in Figure 2.5 and the Euclidean distance measure, calculate the 5-nearest neighbours of the instance with first and second attributes 9.1 and 11.0, respectively ?

Recap of Previous Lecture
Content of This Lecture
Summary & Checklist

Summary & Checklist

- Introduction
- Nearest instance
- k-Nearest Neighbour Classification
- Example of classification using Nearest Neighbour Algorithm
- Distance Measures
- Distance Measures: Euclidean Distance
- Distance Measures: Manhattan Distance or City Block Distance
- Distance Measures: Maximum Dimension Distance
- Nearest Neighbour Algorithm: Step-by-Step
- Self-Assessment Exercise

Reminder | Next Lecture!

Next Lecture...

Classification using Decision Trees (Ch. 3)

- *Be ready!*
- *Download & print the lecture notes before your class.*

Thank You !

✉ s.zahrani@tu.edu.sa