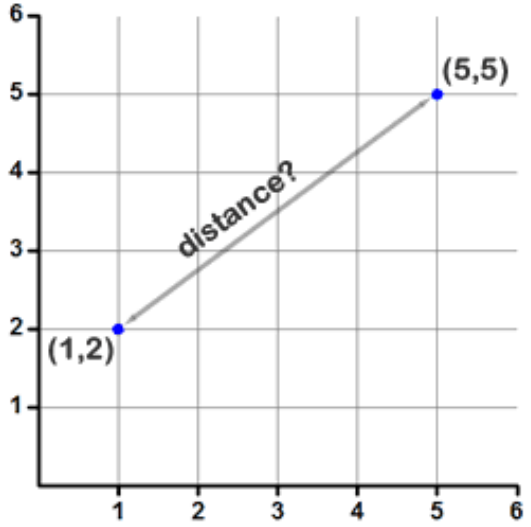
DISTANCES IN CLASSIFICATION

CAFÉ SCIENTIFIQUE - 07/01/2016

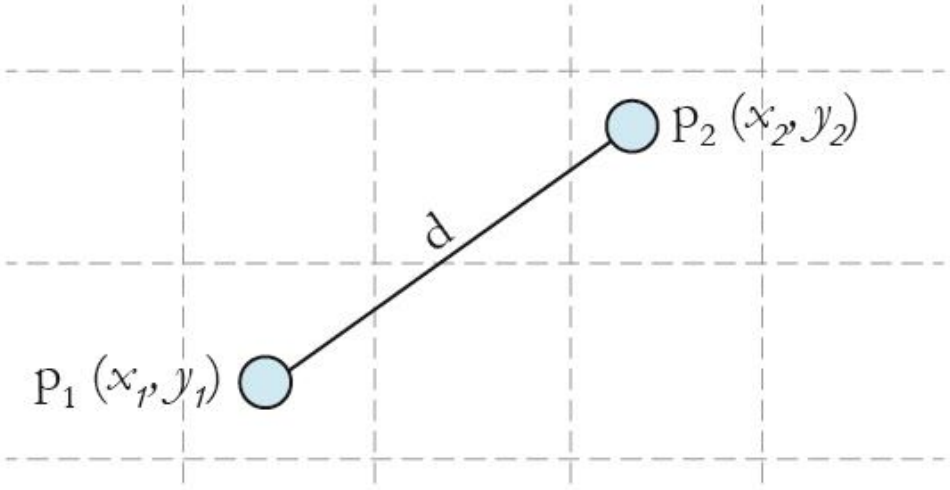
INTRODUCTION

- The notion of distance is the most important basis for classification.
- Standard distances often do not lead to appropriate results.
  - For every individual problem the adequate distance is to be decided upon.
  - The right choice of the distance measure is one of the most decisive steps for the determination of cluster properties.



EUCLIDEAN DISTANCE

- The **Euclidean distance** or **Euclidean metric** is the "ordinary" (i.e. straight-line) distance between two points in Euclidean space.
- The **Euclidean distance** between points **p** and **q** is the length of the line segment connecting them ($\overline{\mathbf{pq}}$).



$$\text{Euclidean distance (d)} = \sqrt{(x_2\text{-}x_1)^2 + (y_2\text{-}y_1)^2}$$

$$\mathrm{d}(\mathbf{p}, \mathbf{q}) = \mathrm{d}(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

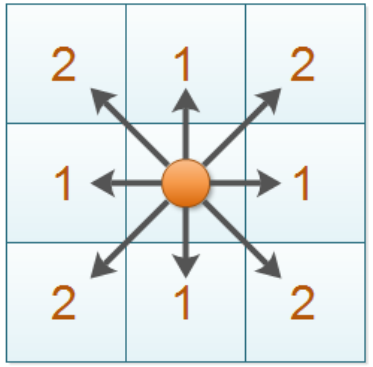
$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

MANHATTAN DISTANCE

- **Taxicab geometry** is a form of geometry in which the usual metric of Euclidean geometry is replaced by a new metric in which the distance between two points is the sum of the (absolute) differences of their coordinates.
- The **Manhattan distance**, also known as rectilinear distance, city block distance, taxicab metric is defined as the sum of the lengths of the projections of the line segment between the points onto the coordinate axes.

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

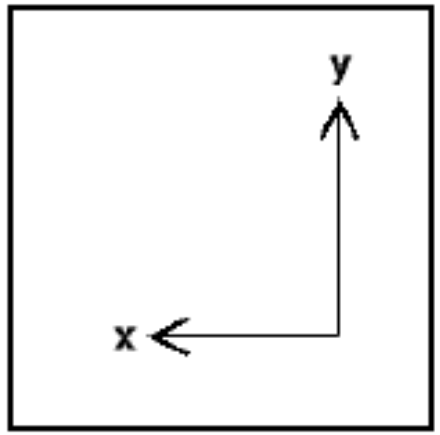
Manhattan Distance



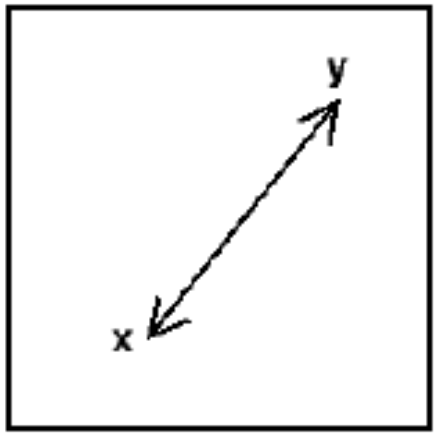
$$|x_1 - x_2| + |y_1 - y_2|$$

- In chess, the distance between squares on the chessboard for **rooks** is measured in Manhattan distance.

EUCLIDEAN VS. MANHATTAN DISTANCE



Manhattan



Euclidean

CHEBYSHEV DISTANCE

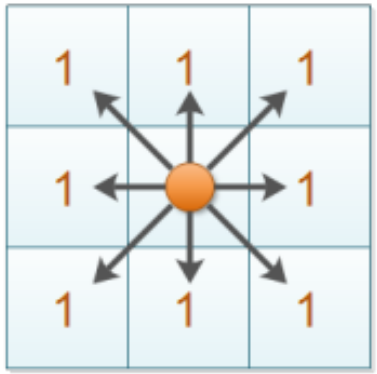
- The Chebyshev distance between two vectors or points p and q, with standard coordinates p_i and q_i respectively, is :

$$D_{\text{Chebyshev}}(p,q) := \max_i(|p_i - q_i|).$$

- It is also known as **chessboard distance**, since in the game of chess the minimum number of moves needed by a king to go from one square on a chessboard to another equals the Chebyshev distance between the centers of the squares

	a	b	c	d	e	f	g	h	
8	5	4	3	2	2	2	2	2	8
7	5	4	3	2	1	1	1	2	7
6	5	4	3	2	1	♔	1	2	6
5	5	4	3	2	1	1	1	2	5
4	5	4	3	2	2	2	2	2	4
3	5	4	3	3	3	3	3	3	3
2	5	4	4	4	4	4	4	4	2
1	5	5	5	5	5	5	5	5	1
	a	b	c	d	e	f	g	h	

Chebyshev Distance



$$\max(|x_1 - x_2|, |y_1 - y_2|)$$

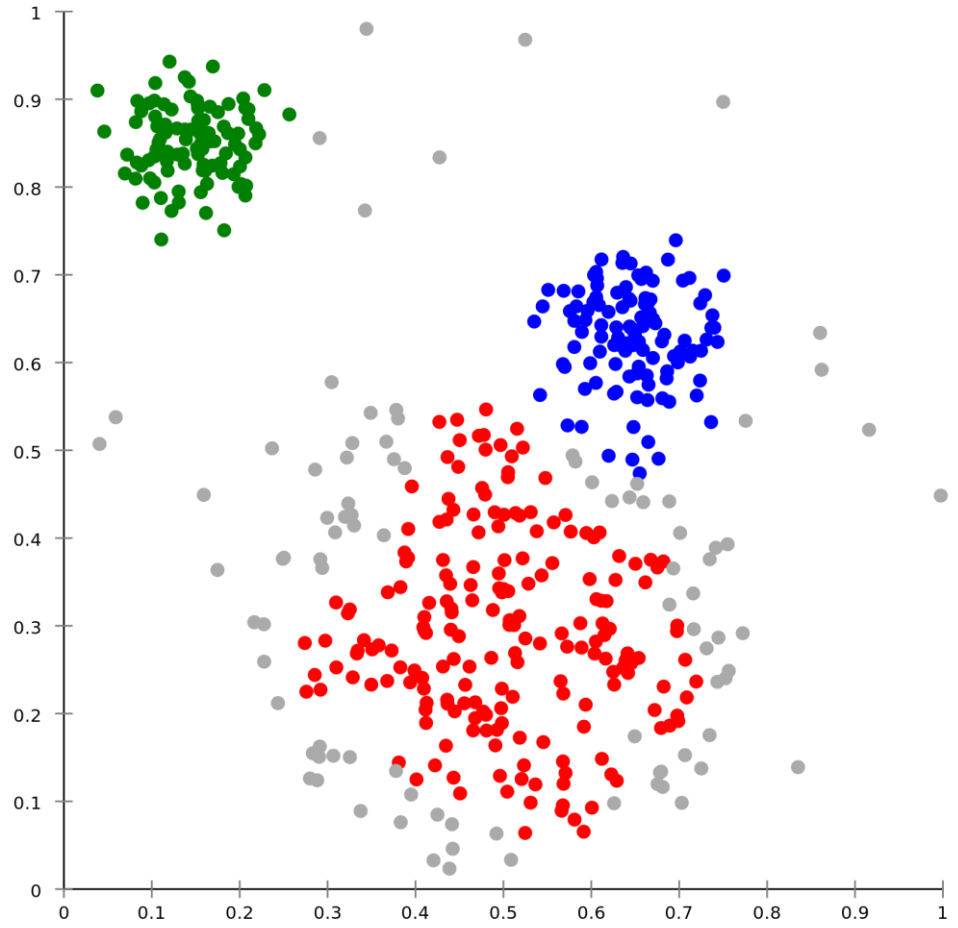
HAMMING DISTANCE

- The **Hamming distance** between two strings of equal length is the number of positions at which the corresponding symbols are different.
  - In another way, it measures the minimum number of *substitutions* required to change one string into the other.

- Example : The Hamming distance between:
  - "karolin" and "kathrin" is 3.
  - "karolin" and "kerstin" is 3.
  - **1011101** and **1001001** is 2.
  - **2173896** and **2233796** is 3.

- It is used in telecommunication to count the number of flipped bits in a fixed-length binary word as an estimate of error, and therefore is sometimes called the **signal distance**.

DISTANCE CALCULATION IN CLUSTERS



LOCAL DISTANCE FUNCTIONS, GLOBAL DISTANCE FUNCTIONS AND WEIGHTS

- A **global distance function**, *dist*, can be defined by combining in some way a number of local distance functions, *dist*$_A$, one per attribute.
  - The easiest way of combining them is to *sum* them: $\mathrm{dist}(x,q) =_{\mathrm{def}} \sum_{i=1}^{n} \mathrm{dist}_{A_i}(x.A_i, q.A_i)$

- More generally, the global distance can be defined as a *weighted sum* of the local distances :

$$\mathrm{dist}(x,q) =_{\mathrm{def}} \sum_{i=1}^{n} w_i \times \mathrm{dist}_{A_i}(x.A_i, q.A_i)$$

- A *weighted average* is also common :

$$\mathrm{dist}(x,q) =_{\mathrm{def}} \frac{\sum_{i=1}^{n} w_i \times \mathrm{dist}_{A_i}(x.A_i, q.A_i)}{\sum_{i=1}^{n} w_i}$$

HETEROGENEOUS LOCAL DISTANCE FUNCTIONS

- **Hamming distance :** The easiest local distance function, known as the *overlap function*, returns 0 if the two values are equal and 1 otherwise:

$$\mathrm{dist}_A(x.A, q.A) =_{\mathrm{def}} \begin{cases} 0 & \text{if } x.A = q.A \\ 1 & \text{otherwise.} \end{cases}$$

- **Manhattan distance for numeric attributes :** If an attribute is numeric, then the local distance function can be defined as the *absolute difference* of the values, local distances are often *normalised* so that they lie in the range 0 . . . 1 :

$$\mathrm{dist}_A(x.A, q.A) =_{\mathrm{def}} \frac{|x.A - q.A|}{A_{max} - A_{min}}$$

- We can combine absolute distances and the overlaps in order to handle both numeric and symbolic attributes:

$$\mathrm{dist}_A(x.A, q.A) =_{\mathrm{def}} \begin{cases} \frac{|x.A - q.A|}{A_{max} - A_{min}} & \text{if } A \text{ is numeric} \\ 0 & \text{if } A \text{ is symbolic and } x.A = q.A \\ 1 & \text{otherwise.} \end{cases}$$

KNOWLEDGE-INTENSIVE DISTANCE FUNCTIONS

- Human experts can sometimes define domain-specific local distance functions, especially for symbolic-valued attributes.
- For example, the last meal a person ate has values ***none***, ***snack*** and ***full***. These can be thought of as totally ordered by the amount of food consumed:
  - **None < Snack < Full**
- We can assign integers to the values in a way that respects the ordering:
  - ***none*** **= 0**
  - ***snack*** **= 1**
  - ***full*** **= 2**

EXAMPLE

- Sex (male/female)
- weight (between 50 and 150 inclusive)
- amount of alcohol consumed in units (1-16 inc.)
- Last meal consumed today (none, snack or full meal)
- Duration of drinking session (20-320 minutes inc.)
- The classes are : **over** or **under** the drink driving limit.

x_1

sex	*female*
weight	60
amount	4
meal	*full*
duration	90
class	*over*

x_2

sex	*male*
weight	75
amount	2
meal	*full*
duration	60
class	under

q

sex	*male*
weight	70
amount	1
meal	*snack*
duration	30
class	?

What is the distance between x_1 and q, and x_2 and q ?

QUESTIONS