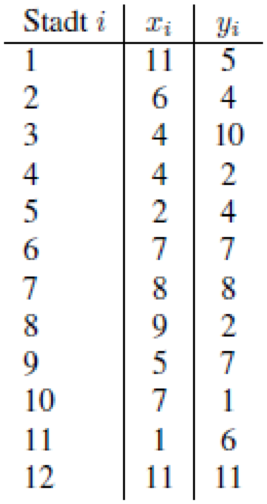
Machine Learning – WS’12
Exercise-6

Prof. Dr. Dr. Lars Schmidt-Thieme, Umer Khan
Information Systems and Machine Learning Lab (ISMLL),
University of Hildesheim

k – Nearest Neighbors

Problem-1:

Given the data of 12 cities with following co-ordinates:



Stadt i	x_i	y_i
1	11	5
2	6	4
3	4	10
4	4	2
5	2	4
6	7	7
7	8	8
8	9	2
9	5	7
10	7	1
11	1	6
12	11	11

The distance between city ‘a’ with co-ordinates (a_1, a_2) and city ‘b’ with (b_1,b_2) is defined by the following formula:

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

a) Plot the cities in coordinate system.

b) For 12 cities, define the corresponding distance matrix.

c) Solve for Travelling Salesman Problem, starting with city 1, by using 1NN heuristic. Draw the path in coordinate system. What is the total length of Travel Path. Is this path an optimal one ? Does there exists a shorter one?

d) Freely choose another city, as starting point and use 1NN to find an optimal path from it. Draw in coordinate system.

Problem 2:

a) Given the description of Minkowski’s Metric, which metric has been used in Problem-1. Is it a Minkowski’s metric ?

b) As mentioned in lecture slides,

$$d(x,y) := 1 - I(x=y) \quad \text{mit } I(x=y) := \begin{cases} 1 & \text{falls } x=y \\ 0 & \text{sonst} \end{cases}$$

is a Minkowski metric L_p with $p=\infty$. Prove it.

c) One common task in Bioinformatic is comparison of DNA sequence. It is a common task to compare two sequences based on their edit distances for finding similarity. Calculate edit distance for following DNA sequence.

AGTCTGTA
GTTCTA

Problem-3:

Given a set of 5 data points: $x_1=2$, $x_2=2.5$, $x_3=3$, $x_4=1$, and $x_5=6$. Find Parzen probability density function estimates at $x=3$, using the Gaussian function with $\sigma=1$ as window function.