Machine Learning
Exercise Sheet 5

Prof. Dr. Dr. Lars Schmidt-Thieme, Martin Wistuba
Information Systems and Machine Learning Lab
University of Hildesheim

November 24th, 2015
Submission until December 1st, 13.00 to wistuba@ismll.de

Exercise 9: Distance Metrics (5 Points)

Given are 5 cities with following coordinates:

City i	x_i	y_i
1	11	5
2	6	4
3	4	10
4	4	2
5	2	4

The distance between the city a with coordinates (a_1, a_2) and the city b with coordinates (b_1, b_2) is defined by:

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

a) Estimate the distance matrix $\mathbf{D} = (d(a, b))$ for the 5 cities.

b) Is the distance metric a Minkowski Metric?

c) A task in bio informatics is to compare DNA sequences. A usual task is to compare two sequences with respect to its edit distance to check if they are similar. Execute the algorithm introduced in the lecture to estimate the edit distance of following DNA sequences:

AGTCTGTA
GTTCTA

Exercise 10: Nearest-Neighbor and Kernel Regression (5 Points)

Given is following data set:

x	y	x	y
1	2	6	12
2	4	7	14
3	6	8	16
4	8	9	18
5	10	10	20

a) Predict the target for 0, 2.5 and 5.75 using 2-nearest-neighbor regression

b) The nearest-neighbor regression considers only instances in its neighborhood but does not consider that they might be very far away. Kernel regression is similar to nearest-neighbor regression but the neighborhood does not have a fixed size. Instead, all instances that are close enough contribute to the prediction with a weight defined by its distances. A common prediction function for the kernel regression is

$$\hat{y}\left(x_0\right) = \frac{\sum_{(x,y)\in\mathcal{D}_{train}} K\left(x, x_0\right) y}{\sum_{(x,y)\in\mathcal{D}_{train}} K\left(x, x_0\right)}$$

where K is a similarity measure.
Use

$$K\left(x, x_0\right) = D\left(\frac{|x - x_0|}{\lambda}\right)$$

with

$$D\left(t\right) = \begin{cases} \frac{3}{4}\left(1 - t^2\right) & t < 1 \\ 0 & \text{otherwise} \end{cases}$$

and $\lambda = 2$ to predict the target for 0, 2.5 and 5.75.

c) Plot the data and the models of parts a) and b). Compare the models. Where do you see advantages or disadvantages?