Machine Learning
Exercise Sheet 8

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

December 12th, 2016
Submission until December 19th, 13.00 via learnweb!

Exercise 15: Distance Metrics (10 Points)

Given are 5 cities with following coordinates x_1 and x_2:

City i	x_1	x_2
1	11	5
2	6	4
3	4	10
4	4	2
5	2	4

The distance between the city x with coordinates (x_1, x_2) and the city $\tilde{x}$ with coordinates $(\tilde{x}_1, \tilde{x}_2)$ is defined by:

$$d(x, \tilde{x}) = \sqrt{(x_1 - \tilde{x}_1)^2 + (x_2 - \tilde{x}_2)^2}$$

a) Estimate the distance matrix $\mathbf{D}$ for the 5 cities.

b) A common task in bio informatics is to compare DNA sequences, which are sequences consisting of four different bases: thymine (T), adenine (A), cytosine (C) and guanine (G). 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 16: Nearest-Neighbor and Kernel Regression (10 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 $x = 0$, $x = 2.5$ and $x = 5.75$ using 2-nearest-neighbor regression using the L_2 metric.

b) Make a scetch of the final prediction for $x \in [0, 10]$ of the resulting 2-nearest-neighbor regression. What is noticeable?

c) The nearest-neighbor regression considers instances in its neighborhood, but neglects the actual distance. Kernel regression is similar to nearest-neighbor regression where the neighborhood **does not have a fixed size**. Instead, all instances contribute to the final prediction weighted by their similarity to the instance for which we want to predict. Precisely, the prediction is

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

where K is a similarity measure. So the prediction is an average of the targets seen in the training data, however weighted by the similarities. Use the similarity function

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

where $D(t)$ is defined as

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

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