Maschinelles Lernen
Übungsblatt 6

Dr. Steffen Rendle, Christoph Freudenthaler
Wirtschaftsinformatik und Maschinelles Lernen (ISMLL)
Universität Hildesheim

1. Dezember 2010
Abgabe bis 10. Dezember

Aufgabe 1: Nächste-Nachbar-Verfahren

Gegeben seien 12 Städte mit den folgenden Koordinaten:

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

Die Distanz zwischen der Stadt a mit Koordinaten (a_1, a_2) und der Stadt b mit Koordinaten (b_1, b_2) ist durch folgende Formel definiert:

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

a) Legen Sie ein Koordinatensystem an, in welches Sie die 12 Städte einzeichnen.

b) Bestimmen Sie zu den gegebenen 12 Städten die zugehörige Distanzmatrix $\mathbf{D} = (d(a,b))$.

c) Lösen Sie das TSP (*travelling salesman problem*, das Finden der optimalen Reiseroute) mit der 1-Nearest-Neighbor-Heuristik, wobei Sie bei der Stadt 1 beginnen. Zeichnen Sie die Wegstrecke in das Koordinatensystem ein. Wie lang ist die zurückgelegte Strecke? Ist der gefundene Weg optimal, d.h. hat er die kleinstmögliche Wegstrecke?
Wählen Sie eine beliebige andere Stadt als Startpunkt und bestimmen Sie von dort aus eine Lösung mit der 1-Nearest-Neighbor-Heuristik. Zeichnen Sie die Wegstrecke in das Koordinatensystem ein.

Aufgabe 2: Distanzmaße

a) Wiederholen Sie die Definition der Minkowski-Metrik. Welche Metrik wurde in Aufgabe 1 verwendet. Ist die verwendete Metrik eine Minkowski-Metrik?

b) In der Vorlesung wurde behauptet, dass die 0-1-Distanz

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

eine Minkowski-Metrix L_p mit $p = \infty$ ist. Beweisen Sie diese Aussage.

c) Eine Aufgabe der Bioinformatik ist es mehrere DNA-Sequenzen miteinander zu vergleichen. Ein Standardarbeitsschritt ist es dabei, 2 DNA-Sequenzen hinsichtlich ihrer Edit-Distanz auf (Un-)ähnlichkeit hin zu überprüfen. Führen Sie diesen ersten Analyseschritt durch und erechnen Sie die Edit-Distanz der folgenden beiden DNA-Sequenzen:

AGTCTGTA
GTTCTA