Exercise 1 k-NN

The Health Office wants to automatize health warnings for certain weather condidtions. The following data have been collected:

ID	Temperature	Humidity	Age	Health Risk
1	High	High	Young	Yes
2	Low	High	Old	Yes
3	Medium	Normal	Old	No
4	High	Normal	Old	Yes
5	High	High	Old	Yes
6	Medium	High	Young	No
7	Medium	Normal	Young	No
8	Medium	High	Average	No

In the following, use k-nearest neighbor (k-NN) with majority voting, Manhattan distance, and 0/1 distance between attribute values.

1–a   Use the provided data as training data for a 1-NN classifier and compute the predicted classes for the following test instances. Also provide the one or all nearest neighbors and the computed distance(s).

Temperature	Humidity	Age	Health Risk	Prediction	next neighbors
Low	Normal	Young	Yes		
Medium	High	Old	No		
High	High	Average	Yes		

1–b   In the course we have introduced the terms "bias" and "variance". Does, in general, a k-NN classifier with $k = 1$ or with $k = 3$ have a higher bias? Justify your answer.

Exercise 2 Clustering

Given are the following 12 instances in a two-dimensional vector room.

id	x	y		id	x	y		id	x	y		id	x	y
$\vec{d_1}$	1	1		$\vec{d_4}$	2	4		$\vec{d_7}$	4	3		$\vec{d_{10}}$	5	5
$\vec{d_2}$	1	3		$\vec{d_5}$	3	1		$\vec{d_8}$	4	4		$\vec{d_{11}}$	6	5
$\vec{d_3}$	2	2		$\vec{d_6}$	3	5		$\vec{d_9}$	4	5		$\vec{d_{12}}$	5	6

The distance between two vectors $\vec{d_1}$ und $\vec{d_2}$ is computed with the Manhatten distance, i.e.

$$d(\vec{d_i}, \vec{d_j}) = |x_i - x_j| + |y_i - y_j|$$

So the distance between vectors $(1, 1)$ and $(4, 6)$ is $3 + 5 = 8$

2–a   Perform one iteration of the k-means clustering with $k = 2$, i.e.:

1. Begin with the randomly chosen cluster centers $(1, 1)$ and $(3, 5)$
2. Compute the clusters which results, and show them, i.e., show the list of instances which constitute each cluster
3. Compute the new cluster centers.

2–b   If you would continue with more iterations, when would the algorithm stop?

2–c   (Homework:) Assume that the instances $\{\vec{d_4}, \vec{d_5}, \vec{d_6}, \vec{d_7}, \vec{d_8}, \vec{d_9}\}$ belong to the class $\oplus$, all the other to the class $\ominus$.

Compute the Rocchio classifier. Which problem do you encounter?