Exercise: 5

Meeting on February 2nd / 4th

Aufgabe 1: k-Nearest Neighbor Classifier

Consider the following table of symptoms and corresponding disease classifications by a medical expert:

Case	Fever	Vomit	Diarrhea	Shivering	Classification
c_1	no	no	no	no	healthy (G)
c_2	average	no	no	no	Influenza (I)
c_3	high	no	no	yes	Influenza (I)
c_4	high	yes	yes	nein	Salmonella poisoning (S)
c_5	average	no	yes	no	Salmonella poisoning (S)
c_6	no	yes	yes	no	Bowel Inflammation (D)
c_7	average	yes	yes	no	Bowel Inflammation (D)

Table 1: Case base for disease classification using k-nearest neighbor

For computation using the Hamming-distance we need a similarity measure for new cases. Using their experience, the expert thought of a similarity measure for a given case c from the case base and a query q as well as corresponding weights w_a:

sim_F

q/c	no	average	high
no	1.0	0.7	0.2
average	0.5	1.0	0.8
high	0.0	0.3	1.0

$\text{sim}_E = \text{sim}_D = \text{sim}_Z$

q/c	yes	no
yes	1.0	0.0
no	0.2	1.0

Weights
$w_F = 0.3$
$w_E = 0.2$
$w_D = 0.2$
$w_Z = 0.3$

(a) Compute the similarity between all cases c from the case base with the query q = (high, no, no, no). Which classification do you get for $k = 1$? Which one for $k = 3$?

(b) Compute the similarity between all cases c from the case base with the query q = (high, no, yes, yes). Which classification do you get for $k = 1$? Which one for $k = 3$?

Aufgabe 2: Expectation Maximization Algorithm

A friend of your's wants to play a game and explains: He has two coins A and$B$, for which you are supposed to determine wether the coin is biased. This means that instead of the usual 50/50 chance the probabilities for "heads" are given by parameters θ_A and θ_B (and the probability for "tails" is given by $1 - \theta_{A/B}$.

Your friend allows you to test the coins 5 times, however he will only give you one coin at a time randomly (50/50 chance) and you don't know, which one it is. During each experiment you toss the coin 10 times and write down the results. This leads to Tab. 2. The used coin is only known for task (a).

# Experiment	Heads	Tais	Coin
1	5	5	B
2	9	1	A
3	8	2	A
4	4	6	B
5	7	3	A

Table 2: Series of 5 experiments with 10 tosses each. The coin is given here as additional information, but is unkown for the EM-algorithm task.

The probability of observing k heads for a coin $i \in A, B$ with parameter θ is given by a binomial distribution:

$$p_i(k, \theta) = \binom{10}{k} \theta_i^k (1 - \theta_i)^{10-k} \tag{3}$$

(a) Use the maximum likelihood method to compute θ_A und θ_B with the information given in the table.

(b) The log-likelihood is often used instead of the simple likelihood. Compute this for the binomial distribution.

(c) You now want to use the expectation maximization algorithm to approximate $\hat{\theta}_A$ and $\hat{\theta}_B$ from the given observations. You use $\hat{\theta}_A = 0.6$ and $\hat{\theta}_B = 0.5$ as start values. Compute the first iteration of the EM-algorithm by hand. *Hint:* Since the probability for both coins is the same ($P(A) = P(B) = 0.5$) you can leave them out for simplicity.

(d) Now program the EM-algorithm until convergence. The convergence criterium is $\Delta < 0.001$ with $\Delta = \max(|\hat{\theta}_A^{t+1} - \hat{\theta}_A^t|, |\hat{\theta}_B^{t+1} - \hat{\theta}_B^t|$.