

Tutorial Business Analytics

Homework 9 - Solution

Exercise 9.3

Apply k-means clustering for the following items and initial cluster centers a and b.

p_i	x
1	1
2	2
3	7

p_i	x
 a	3
 b	0

Solution 9.3


1. Assign instances to nearest cluster centre

p_i	x
1	1
2	2
3	7

p_i	x
 a	3
 b	0


2. Update cluster centre

p_i	x
1	1
2	2
3	7

p_i	x
 a	4.5
 b	1

3. Assign instances to nearest cluster

p_i	x
1	1
2	2
3	7

p_i	x
 a	4.5
 b	1

4. Update cluster centre

p_i	x
1	1
2	2
3	7

p_i	x
 a	7
 b	1.5

5. Assign instances to nearest cluster centre – no reassignment: termination

Exercise 9.4

We want to automatize the Expectation Maximization algorithm using R. This will allow us to run a higher number of phases and to solve different instances. However, we assume that we will stick to two clusters only. As an example we solve the instance from Exercise 9.2.

- a) Write a function in R which, for a given vector x and parameters μ_A, σ_A , returns the solution of $f(x, \mu_A, \sigma_A) = \frac{1}{\sigma_A \sqrt{2\pi}} \cdot e^{-\frac{(x-\mu_A)^2}{2\sigma_A^2}}$. (Hint: Have a look at <https://www.statmethods.net/management/userfunctions.html> to see how to write a function in R.)

```
myf <- function(x, mu, sigma){  
  [...]  
  return(...)  
}
```

- b) Initialize your start values.

```
values <- c(.76, .86, 1.12, 3.05, 3.51, 3.75)  
mu_a <- 1.12  
sigma_a <- 1  
p_a <- .5  
mu_b <- 3.05  
sigma_b <- 1  
p_b <- .5
```

- c) Build a for-loop which repeats the expectation and the maximization step for two times.

```
for (i in 1:2){  
  #Calculate likelihoods  
  [...]  
  #Update parameters  
  [...]  
  mu_a <- ...  
  sigma_a <- ...  
  p_a <- ...  
  #same for b  
  [...]  
}
```

- d) Experiment with the following starting parameters and higher numbers of repetition (increase the counter in the for-loop). What do you observe?

Sigma_a = sigma_b = 1, p_a=p_b=0.5

mu_a	mu_b
.76	3.75
.86	1.12

Solution 9.4

See R Script Exercise 9.4.

Exercise 9.5

- a) Name benefits that an ensemble model (ideally) has in comparison to a single model.
- b) In terms of the training process, what is a major difference between bagging and boosting?

Solution 9.5

- a) Ensemble models tend to be more stable than single models. As the final prediction is the summary of a lot of different “expert opinions”, a small change in the input data does not necessarily change the final prediction. Moreover, the combining of models might reduce the predictor’s variance. Collectively can lead to better prediction performances.
- b) The bagging method draws samples and then trains several models at the same time, which is why it can be easily parallelized. In boosting on the other hand, the k^{th} model depends on the prediction of the $k-1^{\text{th}}$, the $k-1^{\text{th}}$ model depends on the prediction of the $k-2^{\text{th}}$ model and so on (the weights of the data change with each prediction model). In addition, boosting tends to overfit more easily than bagging.