# Data Mining

## Homework 4

Important Notes:

1. Submit in electronic form before 11:59pm on Wednesday, May 10, 2017
2. No late homework will be accepted.
3. The homework should be completed and submitted by each individual.
4. The homework should be submitted through Gradescope. Entry Code: **9BW66M**
5. The homework should be written in English.
6. The HW is worth it 10 points.
7. The [Research] questions require from you to do some research on the Web and get to understand things that were not covered during the lecture.
8. For questions, please use Piazza (English only!)

## Exercise 1: Logistic Regression [3 pts]

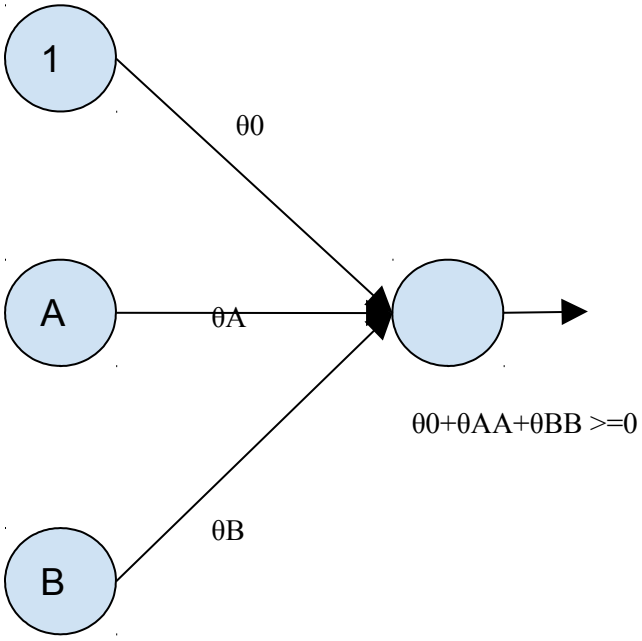
Consider two input variables A and B. They can take one of two values 0, or 1. Now, consider the following logic operators: AND, OR, and XOR

| A | B | A AND B |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 0 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |

| A | B | A OR B |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 1 |

| A | B | A XOR B |
|---|---|---|
| 0 | 0 | 0 |
| 0 | 1 | 1 |
| 1 | 0 | 1 |
| 1 | 1 | 0 |

Further consider the following perceptron:



1. Can this perceptron learn to predict the AND operator? How about the OR? How about the XOR?

AND and OR can be learned, in contrast to the XOR port.

2. Now consider only the logic operators which the perceptron can learn to predict, according to you answer above. For each operator, what should $\theta_A$, what should $\theta_B$be, and what should $\theta_0$ be? Demonstrate why the values you choose are the right ones. (This is a slightly hard question.)

AND:
ThetaA = 1
ThetaB = 1
Theta0 = -0.5

(-0.5 * 1) + (1 * 0) + (1 * 0) = -0.5
(-0.5 * 1) + (1 * 0) + (1 * 1) = 0.5
(-0.5 * 1) + (1 * 1) + (1 * 0) = 0.5
(-0.5 * 1) + (1 * 1) + (1 * 1) = 1.5

OR:
ThetaA = 0.5
ThetaB = 0.5
Theta0 = -0.5

(-0.5 * 1) + (0.5 * 0) + (0.5 * 0) = -0.5
(-0.5 * 1) + (0.5 * 0) + (0.5 * 1) = 0
(-0.5 * 1) + (0.5 * 1) + (0.5 * 0) = 0
(-0.5 * 1) + (0.5 * 1) + (0.5 * 1) = 0.5

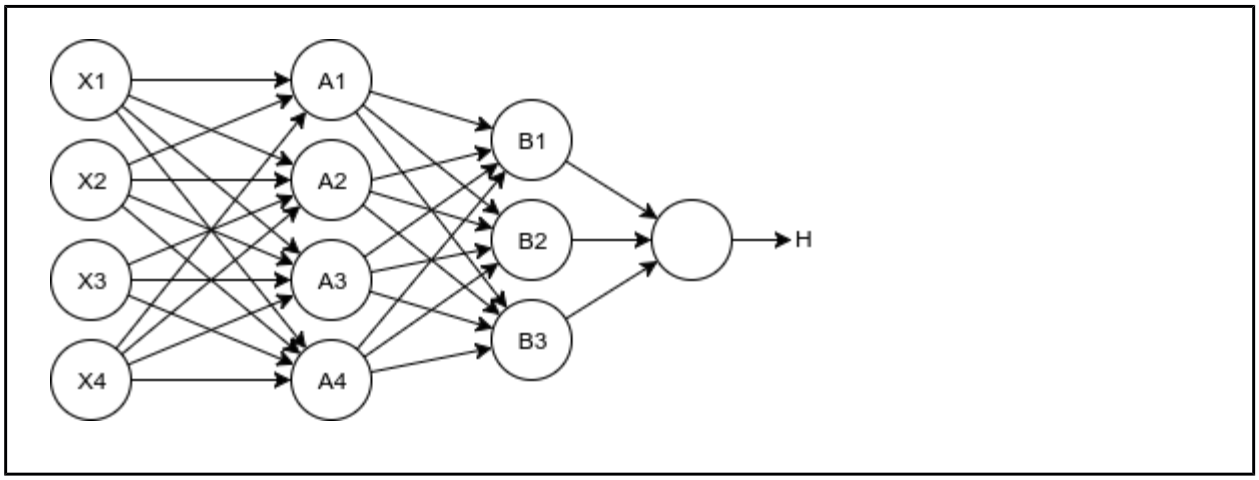
## Exercise 2: Neural Network [4 pts]

Assume that you are working for Booking.com, the online travel agent, used by people to book hotels. When a customer books a hotel, Booking.com keep the following record of their booking,

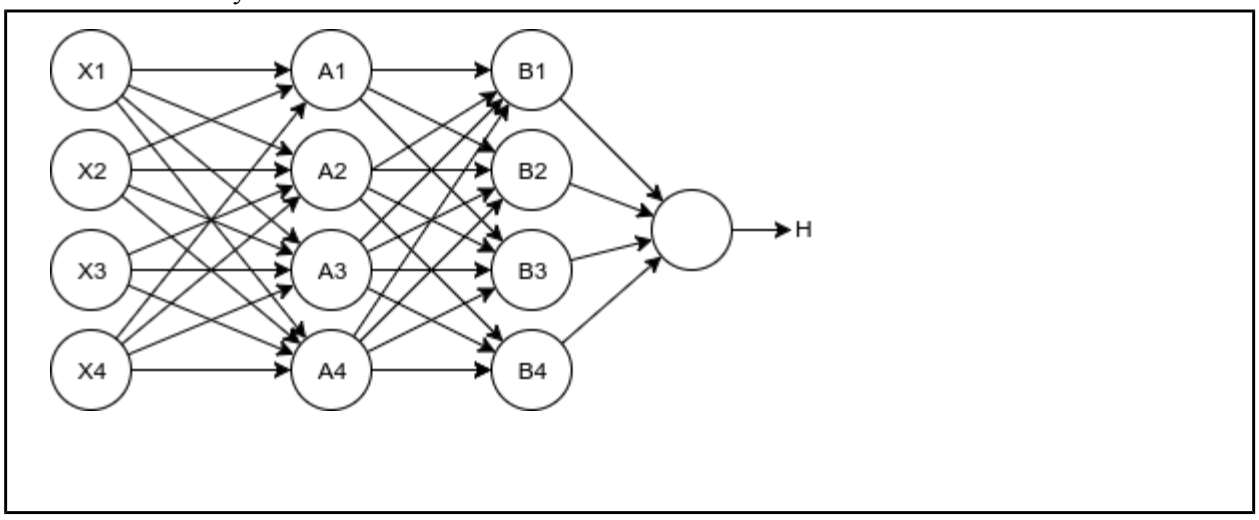
- Hotel price per night ($x_1$)
- Number of nights booked ($x_2$)
- Hotel distance from the center ($x_3$)
- Hotel distance from the airport ($x_4$)

Consider the following classification problem. Given a booking record with the aforementioned information, predict whether the customer is going the destination for *business, leisure,* or due to an *overnight stopover.*

1. Construct (draw) a neural network with 4 layers. For the 2nd layer use 4 activation units and for the 3nd layer use 3 activation units.




2. Construct (draw) a neural network with 4 layers. For the 2nd layer use 4 activation units and for the 3nd layer use 4 activation units.



## Exercise 3: Evaluation [3 pts]

Assume that you are given a set of 1000 emails, {email 1, email 2, …, email 1000}, and you have built a neural network and you want to decide whether it has a good performance.

1. Describe on what set of records will you train and test the network.

I would pick a training set (and a test set) with the same amount of non-spam mails as spam mails in order to train it more fairly.

2. Consider the following confusion matrices:

| | Spam | Not Spam |
|---|---|---|
| Predicted Spam | 5 | 10 |
| Predicted Not Spam | 15 | 170 |

What is the accuracy of the neural network? What is the precision, what is the recall, and what is the $F_1$ measure?

Accuracy: (5 + 170) / 200 = 7/8
Precision: 5 / (5 + 10) = 1/3
Recall: 5 / (5 + 15) = 1/4
F1 = 2 * ((1/3 * ¼) / (1/3 + ¼)) = 2/7

3. Consider a naive classifier that for any email given to it as an input it always predicts Not Spam. Construct the confusion table for this classifier, and compare it to the neural network above. What are your conclusions?

| | Spam | Not Spam |
|---|---|---|
| Predicted Spam | 0 | 0 |
| Predicted Not Spam | 20 | 180 |

Accuracy: 180 / 200 = 9/10
Precision: -
Recall: 0

Accuracy is higher, therefore the model can be classified as “better”. However the recall is lower, which shows the importance of multiple measurement-methods to show the correctness of your model.

This model would be worse than the other model in real life because you would rather have a spam-filter classifying non-spam as spam (instead of the other way around) in order to make it work properly. Otherwise the filter would be useless.

A better way to measure the performance would be to even out the set, so there would be as many spam as non spam (50/50).