Data Mining

Homework 3

Important Notes:

1. Submit in electronic form before 11:59pm on Monday, May 1, 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 [10 pts]

We have collected a dataset of patients information and we wish to predict whether any heart disease is absent (1) or present (2).

Patient	Heart disease	Age	Resting Blood Pressure
1	present	70	130
2	absent	67	115
3	present	57	124
4	absent	64	128
5	absent	74	120
6	absent	65	120

1. Why using linear regression is not a good idea when the problem is a classification problem?

Assume we are using linear regression for a classification problem. We got the hypothesis h(x), and now we also need to use a "threshold". E.g. when h(x) > 0.5, the heart disease is present, otherwise it is absent. However this threshold needs to change when new data-points are added.

...

2. Consider a logistic regression model that predicts heart disease absence, as

$$absent = g(\theta_0 + \theta_1\, age + \theta_2\, resting\_blood\_pressure)$$

where g is the logistic function. For θ_0=-4.2, θ_1=0.04, and θ_2= 0.012, what is the chance that a patient 50 y/o with resting blood pressure equal to 140 has a heart disease? What is your conclusion if instead of a logistic regression you use a perceptron with a step function g?

Using the sigmoid function for g:
Absent = 1/(1+e^(-4.2+(0.04*50)+(0.012*140))) = 0.627
Therefore the chance of this person having a heart disease is 1-0.627=0.373

Conclusion using perceptron:
z = -4.2+(0.04*50)+(0.012*140) = -0.52
Since z < 0, the result will be: absent = g(z) = 0. So in this case the heart-disease is NOT absent

3. Consider the table above, and assume that the prediction of a logistic regression algorithm is shown in the table below. What is the **accuracy** of the algorithm, i.e. what is the proportion of correct predictions? Show your calculation.

Patient	Predicted Heart disease
1	present
2	present
3	absent
4	absent
5	present
6	present

In the prediction table there are 4 wrong predictions (out of 6 in total): 2, 3, 5 and 6.

This means (6-4)/6 = 1/3 = 0.333

So the accuracy is 33,3%