

Homework Assignment #1

Dmytro FISHMAN, Anna LEONTJEVA and Jaak VILO

February 14, 2014

Task 1

Use your own experience and knowledge you have gained during the lecture to come up with the problem that can be solved by the means of Data Mining. Describe it. Make an example of the data. What are the features (variables)? How many observations (size of the data) needed to solve it? Explain what is the type of the data (temporal, spatial, stream, etc.), type of your features? What are the techniques and approaches that can be used to solve it?

Task 2

Read the first chapter on Probability theory from MathWiki web-site We recommend you to solve all the given exercises for training purposes. Play with the simulation of a dice. Explain how the increase of rolling times changes probability distribution of the dice.

Task 3

Now read about probability of two events. Solve the following exercise from the page:

A company makes computer discs. It tested a random sample of discs from a large batch and found that the probability of any disc being defective is 0,025.

Bob buys two discs. Calculate the probability that

- both discs are defective;
- that only one disc is defective.

The company found 4 defective discs in the sample they tested. How many discs were likely tested?

Task 4

Read chapter on conditional probability in order to answer the following question:

At the exam there is 0.8 probability that student has prepared and 0.2 that he has not prepared. Those who are prepared have 0.7 probability of success, those who have not prepared have 0.4 probability of success. What is the probability that randomly selected student will succeed?

Task 5

Consider an example of a recommendation engine e.g. Amazon online shop, make an educated guess on how this kind of system is built. Find several weak points in Amazon recommendation engine that you would like to improve, explain.

Task 6 (2pt)

What is the probability to get 9 or 10 heads when you throw a fair coin 10 times? What is the probability to get 70 or more heads when you throw a fair coin 100 times? Conduct a computational experiment by generating 10,000 times such sequences of 10 coin tosses or 100 coin tosses.