Machine Learning
1st Term Semester 2017-2018
Assignment 1

SYM

September 7, 2017

General instructions: All course participants are requested to handle their exercise solutions as follows:

- Write your answers as PDF using one of the following text processing tools: MS-Word, LibreOffice, or Latex.
- Always mention your **name** and **student ID** in the PDF file.
- For programming section, write the source code using programming language that you prefer or familiar with.
- The assignment is designed to be solved in a week. However you could accomplish it less than a week when you allocate your time properly to work on it.
- The deadlines are as follows,
  - CLO1: Saturday 9.9.2017 at 21.00 UTC+7,
  - CLO2: Monday 11.9.2017 at 21.00 UTC+7,
  - CLO3: Friday 15.9.2017 at 21.00 UTC+7,
- Submit your work through email to the lecturer (**sym.milo.at.gmail.com**) before the deadlines.
- For CLO3, submit your work (PDF and all your codes) into one directory (as a **zip** file) Do not include any of the data files in your solution file.
- **All forms of cheating are strictly prohibited.**

Section 1: CLO1 (Totally 100 points)

1. **(5 points)** What is machine learning?
2. **(5 points)** Give an example of machine learning implementations? (Give other example than those mentioned on the slides)
3. **(12 points)** There are 3 models of machine learning that can be distinguished according to their main intuition. Explain the 3 models, and give an example of algorithms/methods that belongs to each model!
4. **(16 points)** What are supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning?
5. **(12 points)** What are dataset, a data object and an attribute?
6. **(5 points)** What is data preprocessing?
7. **(10 points)** Many machine learning algorithms use measures of similarity or dissimilarity between data objects. Explain differences between similarity and dissimilarity measures!

8. **(15 points)** There are many similarity measures, explain 3 of them!
9. **(15 points)** There are many dissimilarity measures, explain 3 of them!
10. **(5 points)** What is the goal of data visualization?

Section 2: CLO2 (Totally 100 points)

Instructions: In all the exercises, do not just give your answer, but also the derivation of how you obtained it.

11. Given three vectors a = (1 0 0 1 1 0 1), b = (1 1 0 1 0 1 0) and c = (1 0 0 0 0 1 1).
    (a) **(25 points)** Using Jaccard coefficient, what pair of vectors that have high similarity?
    (b) **(25 points)** Using Simple Matching Coefficient, what pair of vectors that have high similarity?
12. Given three vectors p = (0.1 0.8 -0.2), q = (0.1 -0.3 0.6) and r = (-0.1 0.5 0.3).
    (a) **(25 points)** Using Cosine similarity, what pair of vectors that have high similarity?
    (b) **(25 points)** Using Euclidean distance, what pair of vectors that are close to each other?

Section 3: CLO3 (Totally 100 points)

Instructions:

- Write a report (as PDF) of this section.
- We use the report as the main basis for grading: All your results should be in the report. We also look at the code, but we won't however go fishing for results in your code.
- The code needs to be submitted as a runnable file or set of files (command to run it given in the report).
- In your report, the results will be mostly either in the form of a figure or program output. In both cases, add some sentences which explain what you are showing and why the results are the answer to the question.
- If we ask you to test whether an algorithm works, always give a few examples showing that the algorithm indeed works

13. In this problem we will consider similarity measures for movies on the Movielens dataset.

    Download the Movielens data that we sent to you through email. In addition to the data, the file also contains some functions for easily loading the data into Matlab/Octave/R and some example code that you can use if you wish. See the README files for details.

    (a) We will now construct a similarity measure over the movies. For simplicity, let us first consider a simple measure that does not use the explicit (numerical) ratings given by the users, nor the time stamps of the ratings, but only whether or not a given movie was rated by a given user.

    i. **(20 points)** Create a function that, given two different movie IDs as input, outputs the Jaccard coefficient: the number of users who rated both movies divided by the number of users who rated at least one of the movies. For example, for the movies 'Toy Story' and 'GoldenEye' the coefficient should be 0.217.

    ii. **(5 points)** What is the Jaccard coefficient between 'Three Colors: Red' and 'Three Colors: Blue'?

    iii. **(10 points)** What are the 5 movies with highest Jaccard coefficient to 'Taxi Driver'?

    iv. **(10 points)** Select a movie of your own choosing (which you are familiar with), what are the 5 movies with highest Jaccard coefficient to that movie? Do they make sense?

    (b) Now let's try a similarity measure that uses the explicit ratings.

i. **(20 points)** Create a second function that, given two different movie IDs as input, outputs the correlation coefficient of the ratings given to those two movies by all users which have rated both movies. (Note, the function may need to return 0 when the number of users who have rated both is so low that one cannot compute a correlation coefficient.)

ii. **(5 points)** What is now the similarity between 'Toy Story' and 'GoldenEye'?

iii. **(5 points)** How about 'Three Colors: Red' and 'Three Colors: Blue'?

iv. **(10 points)** What are the 5 movies with highest similarity to 'Taxi Driver'?

v. **(10 points)** Again, select a movie of your own choosing and list the 5 movies with highest similarity.

(c) **(5 points)** Provide some brief thoughts on which similarity measure seems to work 'better', in the sense that the computed similarity matches your intuitive sense of similarity. Why do you think this is? Explain.