Machine Learning
2nd-Term Semester 2016-2017
Assignment 3

SYM

March 30, 2017

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

- Write your answers of **theory section** and your report of **programming section**, both 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 if you allocate your time properly to work on it.
- The deadlines are as follows,
  - exercise 1 and 2: Saturday 1.4.2017 at 23.59 UTC+7,
  - exercise 3: Wednesday 5.4.2017 at 23.59 UTC+7,
- Submit your work (PDF and all your codes) into one directory (as a **zip** file) through email to the **lecturer** and the **teaching assistant** before the deadline. Do not include any of the data files in your solution file.
- In all the exercises, do not just give your answer, but also the derivation of how you obtained it.
- **Cheating is strictly forbidden**.

We have 8 datasets for this assignment. (1) Aggregation, (2) Compound, (3) D31, (4) Flame, (5) Jain, (6) Pathbased, (7) R15, and (8)Spiral. We use each data set as both training set and test set.

On each file of data set, we find that it has 3 columns. 1st and 2nd columns are attributes while 3rd column is class label.

Use your student ID to select the data set as follows: let t is the result of modulo operation to the **last digit** of your student ID by 8. For exercise 2, use the data set $t + 1$, while for exercise 3, use the data set $t + 2$.

Programming problems

1. Create a function to calculate performance of classifier that implements F1-micro average, F1-macro average, and simple accuracy (i.e. number of correct prediction divided number of data). (Hints: give an input parameter of the function that inform the function when it should use either F1-micro average, F1-macro average, or simple accuracy). **[5 points]**
2. **(45 points)** In this problem we implement Naive Bayes for classification.

(a) Load the selected data set. Visualize all data points using scatter plot. From your scatter plot, one could easily distinguish each class. (hint: Use attribute 1 as x -axis, attribute 2 as y -axis. Use different color and/or different symbol for each class label.[**5 points**]

(b) Apply Naive Bayes classifier on the selected data set. Your codes have to clearly contain

i. Function for learning that implements the calculation of prior and likelihood probabilities. [**10 points**]

ii. Function for predicting/classifying data that implements the calculation of posterior probabilities. It had better for you to implement the logarithm (log) function.[**10 points**]

iii. Classify each data point using the trained Naive Bayes. Plot the results using scatter plot. From your scatter plot, one could easily distinguish each class. (hint: use different color and/or different symbol).[**5 points**]

iv. By visually comparing figures created from point 2(a) and 2(b)iii, what do you think of the classification results using Naive Bayes. [**5 points**]

(c) Use the function of exercise 1 to evaluate performance of Naive Bayes classifier. What is the F1-micro average of Naive Bayes classifier on the data set. [**5 points**]

(d) Plot the decision boundary resulted from Naive Bayes classifier on the figure that has been created by point 2(a). (Hints: generate data points using range of minimum and maximum value of each attribute, then classify each generated data points using trained Naive Bayes classifier. Use attribute 1 and attribute 2 as both x -axis and y -axis of decision boundary location, while the predicted class label for coloring).[**5 points**]

3. **(50 points)** In this problem we implement Multi-Layer Perceptron (MLP) for classification using backpropagation as a learning algorithm.

(a) Load the selected data set. Visualize all data points using scatter plot. From your scatter plot, one could easily distinguish each class. (hint: Use attribute 1 as x -axis, attribute 2 as y -axis. Use different color and/or different symbol for each class label.[**5 points**]

(b) Give a figure on your report illustrating the MLP architecture that you apply for this exercise. Describe also the activation function. [**5 points**]

(c) Apply Multi-Layer Perceptron as a classifier on the selected data set. Your codes have to clearly contain

i. Function for learning that implements the backpropagation algorithm. [**10 points**]

ii. Function for predicting/classifying data. [**10 points**]

iii. Classify each data point using the trained MLP. Plot the results using scatter plot. From your scatter plot, one could easily distinguish each class. (hint: use different color and/or different symbol).[**5 points**]

iv. By visually comparing figures created from point 3(a) and 3(b)iii, what do you think of the classification results using MLP. [**5 points**]

(d) Use the function of exercise 1 to evaluate performance of MLP as a classifier. What is the F1-micro average of MLP as a classifier on the data set. [**5 points**]

(e) Plot the decision boundary resulted from MLP as a classifier on the figure that has been created by point 3(a). (Hints: generate data points using range of minimum and maximum value of each attribute, then classify each generated data points using trained MLP. Use attribute 1 and attribute 2 as both x -axis and y -axis of decision boundary location, while the predicted class label for coloring).[**5 points**]