Data Mining: Practical Assignment #3

Due on Thu & Fri, May 04-05 2017,

Task 1

Get familiar with Python and numpy:

i.) Which expression generates the column vector $A = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$?

ii.) Given a matrix $B = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$.

Which of the following expressions would give the matrix $C = \begin{bmatrix} 2 & 3 \\ 5 & 6 \\ 8 & 9 \end{bmatrix}$?

1. `B[:,2]`
2. `B[:,1:3]`
3. `B[:,2:3]`
4. `B[2:3,:]`

iii.) Suppose you wish to generate a 3x1 vector D that contains the number 5 in every position.
Which of the following expressions will accomplish this task?

1. `numpy.eye(3)*5`
2. `numpy.identity(3)*5`
3. `numpy.ones(3)*5`
4. `numpy.ones(3,1)*5`
5. `numpy.ones((3,1))*5`

iv.) Which expression allows to create a new matrix E by appending the column vector D to the matrix B?

v.) Suppose you wish to generate a 2x3 matrix F that contains only zeros.
Which of the following expressions will achieve this goal?

1. `numpy.zeros(3)`
2. `numpy.zeros((2,3))`
3. `numpy.zeros(2,3)`
4. `numpy.zeros(3,2)`
5. [`0 0 0; 0 0 0`]
6. `[[0, 0, 0], [0, 0, 0]]`

Task 2

Perceptron and Multi-Layer Perceptron.

a) Design a perceptron at hand (pen & paper should suffice) with inputs x_A and x_B, which implements the boolean function $f(x_A, x_B)$, representing the propositional formula $A \wedge \neg B$.
Draw the network, indicating all relevant parameters.

Hint: Use the value 0 *for false and* 1 *for true, and the activation function* $\varphi(x) = \max(sign(x), 0)$.

b) Explain the term *linear separable* and contrast the boolean functions AND with XOR.

c) Design a perceptron, which implements the function representing A XOR B. *(same hint as in first task)*

d) Let's dive into the code examples that are given in the `data3.zip` – you can choose between Matlab **or** Python. The given code considers a Multi Layer Perceptron solving the XOR problem, including the backpropagation algorithm for learning, as discussed in the lecture. Go to through the code and comment on the steps involved in perceptron learning (forward- and backpropagation steps, error computation, ect).

Task 3

Finally, we want go get practical with the MLP in classification. For this task you have the choice between working on an exercise designed for Matlab and an exercise designed for Python. For both options the procedure is a bit different, but we want to work on the same data here.

Matlab:
Matlab provides a huge Neural Network Toolbox for pattern recognition. In this task we want to rapid prototype a neural network for classification of wine data.
For the description of the data set type *help wine_dataset* in your Matlab command window and start the Neural Networks toolbox (nprtool) by clicking on the provided link. Load the data into your Matlab path and follow the steps. Make some experiments by configuring the parameters, e.g. number of hidden neurons and check your results using the confusion matrix and the ROC curve.

Python:
Python is a nice tool for quickly implementing neural networks and for reasonably quick working on large data if efficient libraries (e.g. numpy) are used. In this task we want to see our neural network in action for a larger real data set.
For the description of the data have a look in the *README_task3.txt* within `data3.zip` file.
In the data folder you also find a Python script for this task. Call the script and make some experiments by configuring the parameters, e.g. number of hidden neurons, and learning parameters. Check your results using the confusion matrix, and comparing development of the mean training error of different runs.

For **both options** discuss the following:

- What is *overfitting* and *generalization*?
- What is the purpose of a *training*, *validation* and *test set*?
- How are these sets created from the whole data set?
- Is it enough to test your parameter only on the test set?

@home Task

To prepare the next tutorial, your homework will be to learn about the following topics:

- Data representation for classification, especially have a look at confusion matrices. In addition, you should be familiar with the quantities you can derive from them, e.g. False Positive Rates (FPR) and their computation
- Association rules, especially the apriori-algorithm
- Decision Trees: Entropy, Information Gain and their computation

Recommended literature:

1. Han J. & Kamber, M. Data mining: Concepts and techniques. Elsevier/Morgan Kaufmann, Amsterdam, 2006.
2. Kantardzic, M.: Data mining : concepts, models, methods, and algorithms. Wiley, NY, 2011.