F20/21DL. Data Mining and Machine Learning

Lab 2. Basics of Data Processing

Covering Practical work to be done by students in Week 2

The purpose of this lab sheet is:

1. to give you practice with basic data processing in Python
2. to help you to start with your **DM & ML portfolio**, in groups.
3. You will learn to
   - Load the data and find attribute information
   - Generate a statistical summary
   - Generate correlation matrix and scatter plot
   - Dealing with Categorical data
   - Dealing with Missing values

1 Python Tutorial and Programming Practice

This part is for your individual programming practice during the week.
Go to the **Python Tutorials** section on **Canvas**. Cover the following steps:

- Complete Python Tutorial P1: run the code in Jupyter notebook, see if you can modify it and answer provided questions
- Run your chosen data set with the same code: make sure you understand and analyise the properties of your data set revealed by this exercise.

2 Machine Learning and Data Mining Portfolio

This part is to be completed in groups, and will be assessed during the labs. Marking scheme: this lab will bring you up to 2 points. 1 point for completing the task, 1 additional point for any non-trivial analytical work with the material.

2.1 What to Do:

2.2 Before the lab:

- Make sure you share all code for your group in some common repository, prior to coming to the lab
- Make sure that you incorporated all tasks from the previous Lab 1 into this shared Jupyther notebook (i.e. you include initial discussion of data sets and reasons for choosing the data set you chose)

- Make sure you can run in your Jupyter Notebook the code given in Python Tutorial P1, using **your chosen data set**.
- In case you have any issues, contact your lab tutor and ask for help.

2.3 During the lab:

- Visualization and initial data exploration help to gain insights on the data attributes and guides in choosing suitable features and building appropriate ML models. Examine your data through visualization and analysis techniques (using the code provided to you on Canvas or adding your own code) and show how this helped you learn more about your data. Discuss how you fixed problems like missing values, errors or outliers - if applicable. Did you need to apply any pre-processing or normalization procedures? If so, why?

  Note: the recorded lectures for Week 1 and Week 2 list a few pre-processing and normalisation procedures. For full marks, you will need to show that you know and understand these methods.

- **The tutors will mark:** quality of your code and the quality of the summary your provide as a group at the end of the lab.

2.4 After the lab:

- *Group rep:* Make sure all group members are given tasks for the week ahead, i.e. assign who completes which incomplete parts of the notebook. Some group members should fill in the summary of the discussion that took place during the lab.
- *Everyone:* Incorporate all code that you started during the lab into this repository.

2.5 Preview of next week Lab tasks

- Next week, you will be asked to analyse the top correlating features in your data set, and create smaller data sets out of the bigger data set.
- Watch out for the next Lab handout with full instructions!

General recommendations (a reminder)

- Each student is advised to keep an individual copy of the python notebook for further experimentation
- Group members can communicate virtually to complete the lab tasks.
- Groups are advised to upload a summary of their findings on your group space on canvas (this could be reviewed by your instructor/tutor)