F20/21DL. Data Mining and Machine Learning

Lab 3. Feature Selection

Covering Practical work to be done by students in Week 3

The purpose of this lab sheet is:

1. to give you practice with more advanced data processing in Python, and in particular, with a common in DM&ML technique of *feature selection*
2. to help you to progress with your **DM & ML portfolio**, in groups.
3. You will learn to
   - Find most correlated attributes to output label
   - *(optional for BSc but recommended for higher marks, mandatory for MSc)* Experiment with various feature selection (feature reduction) methods
   - *(optional, possibly for higher marks)* Gain more confidence with using Computer Vision or high-dimensional real world data sets

1 Lectures and Weekly Tests

At this stage it is becoming impossible to progress with programming without knowing a bit of theory. By the end of week 3, you should be have completed the study materials for lectures Week 1-3.

Of particular relevance to this week exercises are the lectures on attribute selection:

Feature Selection-1: `https://web.microsoftstream.com/video/a7ff6769-8ab1-402f-8dc7-32e58aadf976`

and

Feature Selection-2: `https://web.microsoftstream.com/video/9e1e116b-9708-482e-ae80-879f4941d4d6`

To test your understanding of the material, please complete the following tests /exercises

- Week 2 exercises (Binning and Normalization)
- Week 3 exercises (NN, ZeroR and OneR classifiers)
- Test 1.1 Data Types and Simple Methods (week 1-2). `https://canvas.hw.ac.uk/courses/20761/quizzes/42653`
- Test 1.2 Data Preparation (week 3) `https://canvas.hw.ac.uk/courses/20761/quizzes/42658`

2 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:

- Start studying the `SciKit-Learn` library of Python. This is the subject of the Python tutorial: *Tutorial P2.2 Getting started with sklearn.*

  Although in principle, this week's main goal – feature selection – does not assume you know how to run classifiers and use `SciKit-Learn`, in practice the best way to test whether your feature selection is good is to see whether it improves your classification results. So, you need to pick up this week only as much as it takes to run a simple classifier on your data. Therefore it is helpful to go through Part II of Tutorial P2.2. Next week will be devoted specifically to `SciKit-Learn` and classification: so don't worry, you will be able to study the topic in detail soon.

- (optional) Complete Python Tutorial *P2.1 Image Processing in Python.* This code is in plain Python. You can find a similar code in *P2.2 Part I* in a notebook format. Run the code in Jupyter notebook.

- in both or either cases – Run your chosen data set with the same code. Make conclusions.

3 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.

3.1 What to Do:

3.2 Before the lab:

- (compulsory) Using the methods explained in lectures and tutorials (or additional sources), analyse most correlating features/attributes of the data set, generally and per class. Form 3 data sets, that contain progressively fewer features/attributes.

  Example 1. Suppose my data set has 300 input features and 3 classes. I can find 2, 5, 10 features that best correlate with class 1, 2 and 3 respectively. E.g., for class 1, features 40 and 50 are most correlating. For class 2, features 1 and 5 are most correlating, and for class 3, features 200 and 300 are most correlating. As a result, I will get 3 data sets:

  Data set 1: contains 2 top features for each of 3 classes: 2*3 = 6 features

  Data set 2: contains 5 top features for each of 3 classes: 5*3 = 15 features

  Data set 3: contains 10 top features for each of 3 classes: 10*3 = 30 features

- *(compulsory)* Make sure you add these three data sets to your portfolio. They will be needed on several occasions during the course.

- (*compulsory)* The most straightforward method to choose best correlated features is to use *standard correlation coefficient (Pearson's r) method* that was already introduced in Python code last week. So, the beginner groups should at least be able to use the information about the best correlating features that were computed last week.

- *(optional for BSc, compulsory for MSc)* Study what other feature selection algorithms are available? Do they work well with your data?

- *(optional for BSc, compulsory for MSc)* Hint: Python library `Scikit-Learn` has a package called **PCA**, for Principal Component Analysis, that could be used for feature transformation /reduction. This is one of the most famous methods of feature selection.

  Also see: Aurelien Geron *"Hands-on Machine Learning"*, Chapter 8 *"Dimensionality Reduction"* with accompanying code on GitHub.

- *(compulsory)* To evaluate whether the smaller data sets (with selected features) work better, run a classifier of your choice on the original and smaller data sets, and record their accuracy on these data sets. Note: this will require you to run some basic classifier from `Scikit-Learn`.

Other resources that you might find helpful:

- Data Mining and Machine Learning book, Section 8.1 Attribute Selection
- Feature selection in Python `https://towardsdatascience.com/feature-selection-techniques-in-machine-learning`
- Getting Started with Feature Selection `https://www.kdnuggets.com/2020/08/getting-started-feature-selection.html`

3.3 During the lab:

- Share your proposed solutions, discuss which of the feature selection methods worked best.
- Make conclusions. You may want to think about the following questions: what kind of information about this data set did you learn, as a result of the above experiments? Which features are more important/reliable for which class? Which are less reliable? You will get more marks for more interesting and "out of the box" questions and answers.
- **The tutors will mark:** quality of your code for feature selection and the quality of the summary your provide as a group at the end of the lab.

3.4 After the lab:

- *Group rep:* Make sure all group members have tasks for the week
- *Everyone:* Incorporate the discussion during the lab into your Python code
- *Everyone:* Incorporate all data sets used in the lab into your Portfolio repository.

Further recommendations:

- 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)