F20/21DL. Data Mining and Machine Learning

Lab 4. Classifier Testing and Evaluation

Covering Practical work to be done by students in Week 4

The purpose of this lab handout is:

1. to give you practice with using Python functions for *classifier testing and evaluation* often used in DM&ML projects
2. to help you to progress with your **DM & ML portfolio**, in groups.
3. You will learn to
   - experiment with different evaluation metrics like accuracy, TP rate, FP rate, precision, recall, F measure, confusion matrix, ROC curve
   - *(optional for BSc but recommended for higher marks, mandatory for MSc)* Compare between train/test splits and cross validation for testing classifier performance
   - *(optional, possibly for higher marks)* Use pipelines to test several data transformations

1 Lectures and Weekly Tests

By the end of week 4, you should have completed the study materials for lectures Week 1-4.
Of particular relevance to this week exercises are the lectures on Model evaluation:
Evaluation Metrics-1 `https://web.microsoftstream.com/video/e343255a-2233-4925-954a-67195adb9e23`
and
Evaluation Metrics-2: `https://web.microsoftstream.com/video/04f0950e-3d12-411e-be0c-cc17b910e0dd`
To test your understanding of the material, please complete the following tests /exercises

- Week 3 exercises (NN, ZeroR and OneR classifiers)
- Week 4 exercises (Confusion Matrix - Recall -Precision)
- Test 1.3 Model Testing - Credibility Measures and Testing Methods (Week 4). `https://canvas.hw.ac.uk/courses/20761/quizzes/42654`

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:

- Study *Tutorial P2.2* thoroughly. This tutorial will walk you through steps to train and test your model and use evaluation metrics like : accuracy, precision , recall, F1 score, confusion matrix.
- While studying *Tutorial P2.2* and the introductory slides on sklearn, look up the `SciKit-Learn` documentations of the following functions:

  - train_test_split()
  - cross_val_score(sgd_clf, X_train, y_train, cv=3, scoring="accuracy")
  - knn.fit(X_train, y_train)
  - knn.predict(X_test)
  - knn.score(X_test, y_test))

- (optional for BSc, Mandatory for MSc) Complete Python Tutorial *P3 End to End Machine Learning* and supporting lecture material on canvas `https://canvas.hw.ac.uk/courses/20761/pages/week-4-machine-learning-pro` `module_item_id=1514200`. Note that this is a regression case study, however much of the data processing and transformation steps can be useful for your ML portfolio.

- (optional) Study the *Pipeline Class* in `SciKit-Learn` documentations. Pipelining can help you test different variations of preprocessing - features selection - and classification. Think how this can help you with your ML portfolio.

- (**Important Note**) The evaluation metrics and the tasks in Lab4 mainly deal with classification problems. If you have chosen a data set with a numerical output (i.e. a regression problem) you will need to discretize your output variable into categories/classes. In this case, please explain and justify your approach. We recommend you to use a classification problem for your entire portfolio.

- Use the code from your tutorials and run it with your data. 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) Use the methods explained in lectures and tutorials (or any additional sources) on classifier testing and evaluation. Run any classifier model of your choice on your data set, and record the major metrics: accuracy, TP rate, FP rate, precision, recall, F measure, the ROC area etc (as explained in the lectures). Make conclusions. You may want to continue your experiments to choose the best data processing steps or features for your data; but this time based on further evaluation metrics.

- (compulsory) Plot your confusion matrix and make comparisons between different evaluation metrics. Which ones seem more suitable for your data set?

- *(optional for BSc, compulsory for MSc)* Compare between using train/test split on your data set and using cross validation. Report your results and summarize findings.

- *(optional for BSc and MSc)* `Scikit-learn` provides a Pipeline class for sequence of transformations. This can help you automate some experiments and test various settings. You could define your transformation pipeline to deal with missing values, attribute combination, normalization,....etc. Hint: *P3 Tutorial End to End Machine Learning* presents some useful pipelines. Adapt and use for your data set.

  Also see: Aurelien Geron *"Hands-on Machine Learning"*, Chapter 2 *"End-to-End Machine Learning Project,"* with accompanying code on GitHub.

3.3 During the lab:

- Share your proposed solutions, discuss which evaluation techniques worked best.
- Make conclusions. You may want to think about the following questions: Now that you have expanded your experiments to include more evaluation metrics/techniques, did you learn something new about your data set? For example could you reveal more information regarding data pre-processing or feature selection? Which evaluation metrics are more important/reliable for your application? Which are less reliable evaluation metrics? How do they work to indicate the performance related to different classes? 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 model testing and evaluation and the quality of the summary you 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)