60-473/574 – Machine Learning/Pattern Recognition – Fall 2017

Project Guidelines

Deadlines:

Topic and Description: Dec 1, 2017

Final Submission: Dec 13, 2017

The project is the most important component of the course. It consists of developing or deploying a pattern recognition system, using any (or a combination of) programming tool, and applying one or more of the techniques studied in class to a real-life classification problem. Students are encouraged to, additionally, add their own new ideas and test them in real-life classification.

The project can be done on an individual basis or in groups, typically 3-4 students – group work is encouraged. If two or more students would like to work together, they must clearly state which part of the project was developed by each student and clearly specify it in the final report. Each student will be evaluated individually and a mark will be assigned to each student, also individually.

You can use your favorite programming and/or machine learning tool, including Matlab, Octave, R, Weka, Scikit or the like. A report (10-15 pages) must be submitted along with all the programs developed for the project (in case of Weka/Scikit you must also explain how you use it via screenshots and sample outputs). Note that the report will be evaluated and a mark will be assigned towards the total mark for the project.

The topics are listed below. Students taking 60-574 can only work on Project 3, while students taking 60-473 can choose project can work on Project 1, 2 or 3. The topic/problem you choose should be addressed/solved using the tools discussed in class. For example, using neural networks or hidden Markov models is not acceptable (unless you run the perceptron algorithm). If you use tools like Weka or Scikit, you must choose techniques that were discussed in class.

It is expected that students use their creativity in devising the pattern recognition system. This implies going beyond the basic solutions derived in assignments. The baseline will be a B-range grade. That is, a simple classification system on a standard classification problem will imply a B-range grade. An A-range grade will be given to a project that includes a significant challenge, such as a classification problem not solved before, a large dataset, a large number of features, a new variant of a classification method, and/or a combination of these.

Project Topics:

Project 1: The problem consists of classifying breast cancer patients using gene expression data. The patients are to be categorized into one of five classes: Normal, LumA, LumB, Her2, or Basal. The features are the gene expression ratios of 13,582 genes. This is a multi-class classification problem, and the dataset contains 158 samples in total. The file called “breast cancer subtypes.zip” available at the Resources tab contains a paper with a full description of the problem, the data, and other details, and the dataset. You are free to work on one or more problems as discussed in class: classification, solving the multi-class problem, feature selection, other aspects, or a combination of these.

Project 2: The problem consists of finding meaningful biomarkers for different bladder cancer stages. This can be done via classification and feature selection for selecting genes that contribute to one or more different classifications between stages or among all stages. The dataset contains several samples that belong to different stages. The stage (class) is given in the first column. All other columns contain expressions (features) for many different genes. You are free to work on one or more stages as follows: classification, solving the multi-class problem, feature selection, other aspects, or a combination of these. For the multiclass problem, you can consider 3 classes: T1, T2 and Ta. The dataset and a short description of it are available at the Resources tab in a file called "Bladder cancer dataset.zip".

Project 3: The problem consists of finding meaningful biomarkers in prostate cancer. This can be done via classification and feature selection for selecting genes that contribute to one or more different classifications. A dataset of 494 samples downloaded from the Genomic Data Commons (formerly cBioPortal) contains gene expressions for a few dozen thousand genes. You are free to work on one or more problems as discussed in class: classification, solving the multi-class problem, feature selection, other aspects, or a combination of these, by using one or more clinical variables (e.g., clinical stage of progression, primary site, Gleason score, etc.). The dataset and a short description of it along with the clinical information table are in file "prostate cancer dataset.zip" available at the Resources tab.

Submission

1. **Topic and Description: 3% of the course grade**

   A short description (maximum one page) that gives a summary of the topic and describes the classification problem. You should also briefly summarize how you will solve the classification problem.

   Each student in the group should submit the project via Blackboard.

   Submit this part by December 1, 2017.

2. **Product and Report: 40% of the course grade**

   A report (10-15 pages in PDF) along with the source code must be submitted via Blackboard by the corresponding deadline. The source code or Weka/Scikit files used must also be attached to the submission. The report should have a paper format (conference or journal paper), and it should contain a summary, an introduction, description of the problem, description of the solution, results, discussion of results, conclusion, and references. All these items will contribute to the final mark.

   Each student in the group should submit the project via Blackboard.

   Submit this part by December 13, 2017.

Marks will be deducted for late submissions (10% per day for up to 3 days). After 3 days the mark will be zero.