COMP-4730/5740 – Advanced AI - Machine Learning and Pattern Recognition

Fall 2019 – Project Guidelines

Deadlines:

Project proposal: Nov 28, 2019

Final Submission: Dec 16, 2019

The project is one of the most important component of the course. It consists of developing or deploying a machine learning system, using any (or a combination of) of the techniques learned in class and apply it to a real-life classification problem. Students are encouraged to, additionally, propose new methods and approaches and add their own new ideas and test them in real-life classification.

The project can be done on an individual basis or in groups; no more than 3 students. If 2 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's work will be evaluated individually and a mark will be assigned to each student, also individually.

In the implementation, 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 must follow the format of a conference paper like IEEE, ACM or NeurIPS, for example. 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 COMP-8740 can work on Projects 3-7, while students taking COMP-4730 can work on any project. The topic/problem you choose should be addressed/solved using the tools discussed in class. 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 and implementing the machine learning methods. 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. This requirement will be more stringent for graduate students. For students registered in COMP-8740, an A-range grade will be given to a project that includes at least 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.

Ethical, Legal and Societal Aspects of Machine Learning

The project must involve addressing these issues, which must be reflected in the final report. It is expected that at least one page of the report is devoted to discussion of these aspects. 10% of the final project mark will correspond to this item.

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 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 cBioPortal, which contains gene expressions for a few dozen thousand genes, is to be used. 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.

Project 3: Download any dataset from cBioPortal (http://www.cbioportal.org/data_sets.jsp). You are free to work on one or more problems as discussed in class: dimensionality reduction, classification, clustering or regression, or a combination of these, by using one or more clinical variables (e.g., clinical stage of progression, primary site, Gleason score, age, type of therapy, recurrence, survivability, etc.). The dataset should contain at least 100 samples and a few thousand features.

Project 4: Consider a machine learning task for the CIFAR-10 and CIFAR-100 datasets available at https://www.cs.toronto.edu/~kriz/cifar.html. The task could involve representation learning, like dimensionality reduction, classification, clustering or regression, or a combination of these. Develop a machine learning method that perform these tasks. The model should contain innovative features.

Project 5: Involves any new method or variant of a method in machine learning and pattern recognition. Note that novelty must be demonstrated by means of providing: description of the problem, literature review, method, experimental results and comparison with the previous model.

Project 6: A network/graph like Citeseer or PPI. Involves finding patterns, graph classification, node classification, vertex classification or community/cluster detection.

Project 7: Take a paper from the following conferences: ICPR 2018, ICLR 2019, CVPR 2019, SPR/SSPR 2018, RECOMB 2019, ISMB 2019, IEEE PAMI 2018-2019 (journal), NeurIPS 2019, Workshop on Representation Learning on Graphs and Manifolds (ICLR 2019; https://rlgm.github.io). Implement the methods presented in the paper. The methods must involve

one or more techniques discussed in class. A paper from another conference or journal can be considered only by permission of the instructor.

Submission

1. **Project proposal: 4% of the course grade**

   A short report or a short PwP presentation (2-3 pages) that gives a summary of the topic and describes the problem to be dealt with. You should also briefly summarize how you will solve the machine learning problem. Ethical aspects of machine learning may not be discussed at this stage.

   Each student in the group must submit the project proposal via Blackboard Submit by November 28, 2019 at 10:00am.

   COMP-8740 students (only) will present the proposal in class on Nov 28; a presentation of 8-10 minutes per group is expected, depending on the number of groups.

2. **Product and Report: 28% 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 16, 2019.

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