MACHINE LEARNING - CS 577

Computer Science Department, University of Crete

Course Project

The project is a basic part of the course. The goal is to practice the techniques of machine learning on an actual problem of your interest. Also, to practice on publishing a research project. The projects are individual. The project could be very similar to the type of the analysis you did in the previous assignment, with a similar load in terms of analysis; the main difference is that you are selecting the exact topic. The deadline is to be decided when the exams schedule comes out and will be sometime after your final exam.

The procedure for the project includes the following:

1. Submission of the proposal of the project (20%), 1 page
2. Final report (60%), 6 pages max, format restrictions are described below
3. Oral presentation (20%), 15' presentation + 5' questions and answers.

The exact dates for each step are or will be announced.

Proposal

Discuss and negotiate your project and the idea with me and the TAs; ideally, before or after the class. Subsequently, submit a 1 page description of your intended project. The description should include the following:

1. Title
2. Problem Description and Motivation
3. The data that you will be using and how are you going to generate or obtain them
4. The code that you will need to implement or download
5. What will you need to read and understand in order to execute the plan of your project

It's only 1 page, but you need to think, gather information, and plan carefully.

Final Report

The final report must follow the template used for the NIPS conference.
Instructions for Latex (*recommended*) or Word templates can be found at
`http://leon.bottou.org/nips/` or `http://nips.cc/PaperInformation/StyleFiles`.
Check out the publications of NIPS to get an idea.

The report must be organized as follows:

- Introduction – Motivation
- Problem Definition (including data description)
- Approach Followed/Methods Used.
    a. Methods and algorithms used, data representation, explanation and discussion of the choices made
    b. Description of the implementations and software used.
    c. Relation to the state-of-the-art in the area
- Experiments
    a. Description of the questions to address / what are your claims
    b. Detailed description of the experiments and results.
    c. Results of the experiments and conclusions.
- Conclusions –Discussion.

With regards to the reports, try to be short and concise but always complete. You should preferably (but optionally) use the English language which is the most common publication language even in Greek conferences.

Your report should convince that you addressed an interesting, important problem, explain what the problem is on an intuitive but also mathematical level, have at least some innovative aspect, convince that the choice of your methods are reasonable, and that your experiments are methodologically correct and support your claims or answer your research questions.

Your report should contain all necessary details for someone with background in machine learning but not knowledgeable in the specific task, domain, or even methods used to comprehend what you did. You should have enough detail and be precise so that an expert **could repeat your experiments** if desired. The report should be interesting to someone: it is not meant to be a proof that you have worked a lot (this only interests me I guess and no one else). It is meant to contribute something to the area. Your readers in general do not care if you worked a lot, but whether you have something new to teach them, something of interest to tell them. Have you found something interesting about the final model, have you understand something new about the problem domain, have you learn something useful to other researchers by performing the experiments?

Presentation

Congratulations! Your paper has been accepted to a conference in the field and now you will travel to attend the conference and present it (conferences in nice exotic places are preferable of course). You have strictly 15' for the presentation and 5' to answer questions from the audience. The purpose (and possibly structure) of the presentation is the same as the one for the paper. After 20' the Session Chair of the conference will stop you and get you off the podium and have the next speaker take the floor.

It is **strongly suggested** you rehearse your presentation multiple times so you are sure you can stay in time and achieve your presentation goals. The language is preferably English. You will not have time to discuss all the details of your work so you need to choose the level of abstraction, what is important and what to leave out. Look at your audience when you speak. Speak loud and clear, do not rush, leave a couple of seconds the slide on before changing to the next one. Your slides are supposed to guide you and the audience, they are not supposed to have all the text you would like to say. You say the text and the slides just guide the conversation and illustrate your basic points and figures.

Plagiarism

Throughout the project and the reports you will be using algorithms, techniques or maybe even text written by others. In science, we always stand on "shoulders of the giants" to progress. Plagiarism refers to "use or close imitation of the language and thoughts of another author and the representation of them as one's own original work" [1]. You must therefore "Render unto Caesar the things which be Caesar's" and when using someone else's ideas or text refer to the source (unless it is obvious). Other people's quotes must be in "" or denoted otherwise. For example, the famous quotes used here are within "" and the references that are not obvious are cited.

When you cite, **use the work that introduce or invented the idea or algorithm you cite**. Not just any other paper that employs this algorithm!

[1] qtd. in Stepchyshyn, Vera; Robert S. Nelson (2007). Library plagiarism policies. Assoc of College & Resrch Libraries. p. 65. ISBN 0838984169.

Project Topic Selection

Select a dataset, task, problem of your choice. I encourage you to select a classification task (or a regression task), like the ones studied mostly during the course. But you could also use a clustering task, or study a novel algorithm or something different after discussing your idea with me. Many problems may not strike you immediately as classification tasks, but could be cast as such with the appropriate representation. Any problem that is suitable for solving with machine learning techniques is permissible. I present some general categories below but you are welcome to surprise me with something novel.

- Algorithm implementation. You will have to select and implement some interesting

machine learning algorithm and of course, empirically study its performance and compare it against other related algorithms in the field.

- Data Analysis. You can find or generate data for a prediction or diagnosis task. The analysis should have a specific focus and task. For example, it could be to construct a prediction model or identify the important variable sub-set. I strongly suggest to analyze new datasets that appear in one of the webpages of scientific data-analysis competitions. Such competitions also come with prizes! You could also analyze data from an old competition and compare with the winner. You can find such information in the following websites and not only:
  - `http://www.kaggle.com/`
  - `http://www.innocentive.com/`
  - `http://www.chalearn.org/`
  - The annual KDD conference competitions, but also from other conferences such as NIPS.
  - `http://wiki.c2b2.columbia.edu/dream/index.php/Challenges`
- System Implementation. You will construct a system that solves a specific task and improves its performance based on its experience and historical data.

For classification and regression tasks you can find many real data publicly on the web at many sites. For example, at the UCI Machine Learning Repository, the Gene Expression Machine Learning Repository, machine learning and data mining competitions in several conferences. In addition, you can create datasets on your own. Have your friends grade their favorite movies and try to learn a model that predicts whether they will like a given movie. Take their pictures and build a system that learns from their pictures to classify men and women. Measure the load on your computer and try to predict it based on some features that you record. Be inventive.

You are also encouraged to consider a subject that is related to your Ph.D. or M.Sc. thesis. If you would like, I also have several real datasets that you could analyze. At this point, you may have an interesting idea, but not know enough about how to solve it.

Don't be discouraged. I can guide you as to what methods to read, learn and use.