

Final Project Guidelines

Nathan Jacobs, UKY, Machine Learning F2012

October 9, 2012

1 Overview

During this course you will conduct an open-ended research project of your choosing. This is an opportunity for you to dig deeper into an interesting machine learning problem in the context of a real-world data set. Below, you will find some project ideas, but the best idea would be to combine machine learning with problems in your own interest/research area.

1. You will submit a proposal, a mid-term report and a final report. In addition, you will present a brief summary of your proposal, an informal status update and final presentation in class.
2. Your class project must be based on the topics you are learning in this course.
3. You may work individually or with a partner. If you choose to work with a partner, you must submit documentation with each report that describes the contributions of each team member.
4. No latework will be allowed for any aspect of the final project.

2 Grading

This project accounts for a significant portion of your final grade¹. Evaluation of your final project will be based on the following break down:

	Possible Points	Due Date
Project Proposal	30	Oct 17 23:59:59 (csportal)
Paper Summary	50	Oct 19 23:59:59 (csportal)
Midterm Progress Report	40	Nov 8 23:59:59 (csportal)
Midterm Progress Update	10	Nov 8 (in class, 1-4 minutes, no slides)
Final Presentation	30	Dec 4 or 6 (in class, 10-15 minutes, slides required)
Final Report	140	Dec 13 23:59:59 (csportal)
Total	300	

3 Detailed Requirements

This section provides guidelines for what you need to do for each element of the final project.

Project Proposal

You must submit a brief project proposal (single column, 11pt font, probably less than 1 page). A key aspect of the proposal is identifying a set of data you will use for your project. You are encouraged to use an existing dataset that is linked to below, because we know that they have been successfully used for machine learning in the past. If you prefer to use a different data set, we will consider your proposal, but you must have access to this data already, and present a clear proposal for what you would do with it.

Project proposal must include the following information:

- Your name and **student id** on top of the page
- Project title
- Data set description
- Project idea (approximately two paragraphs)
- Software you will need to write and libraries you will use.
- Related work: You must include the bibliographic reference for at least one paper related to your project (you will should read it before submitting your proposal). I find that using [Google](#) or [Google Scholar](#) is the easiest way to find papers. I encourage you to pick paper(s) from recent top machine learning, or related, conferences (ICML, NIPS, ICCV, CVPR). If applicable, this can be the same paper you use for the Paper Summary.
- Teammate: will you have a teammate? If so, who? Maximum team size is two students.
- What will you complete by the midterm project milestone?

¹Each homework is worth 100 points and this project is worth 300 points.

Submit the document (pdf,word) via CS Portal. I will provide feedback, possibly with suggestions for how to adjust your project to the proper scope.

Paper Summary

- Choose a full paper from ICML (<http://icml.cc/2012/papers/>, <http://icml.cc/2012/schedule/>). These are eight pages long, make sure you do not select a workshop paper.
- Spend exactly three uninterrupted hours (no email, no facebook, no TV) trying to understand the paper. You probably want to do a quick search to see if there is a presentation or video already online for this paper. You may also want to track down some of the referenced documents or consult other resources.
- Spend exactly one hour summarizing your understanding of the document in the following format (all in your own words, no copy/paste):
 - Header information: add your name and include the bibliographic reference (Author(s), Title, Publication Venue, Year).
 - Answer the following questions (attempt to write this without looking back at the paper): What problem is being addressed? What is the technical approach? What is new in this work (you may want to compare to previous work)? Why is that interesting or useful? How do they show that it works?
 - For each major section of the paper, and there should be about 6, answer the following questions: What is the goal of this section? What percentage of this section did you understand²? Where did you get confused, for example what equations did you not understand? What could you do to get past this block (two common options: reading a previous paper or consulting a textbook/wikipedia on a basic topic)?
 - (optional) How would you improve this? How was the writing? What did you think about the figures or experiments?

Submit the document (pdf,word) via CS Portal.

Midterm Report and Presentation

In order to increase the quality of the final projects we have a milestone to encourage early progress. You are expected to have initial results by this point, which means working code and visualizations. You will write a short paper and give a brief in-class presentation. The paper will likely serve as the starting point for your final report, so I encourage you to take it seriously.

Your midterm progress report will be a write up in the format of an academic publication³. If you use a single column format with reasonable margins I anticipate the report being between 3 and 5 pages (it will likely be shorter in a two-column format). You must include the following items in your report (with rough page length guidelines):

- Introduction (1 page)
 - Why is your problem interesting/important?
 - Briefly describe the primary contributions of your work
 - Describe some related work: Three papers (probably recent papers from ICML or NIPS) that you read and are related to your project. You must include a bibliographic reference and a short description of how it works and how it relates to your project.
- Problem statement (1/2 page)
 - A description of the problem you are addressing
 - Applied and/or theoretical background for understanding the problem
- Description of progress (2 pages)
 - Features, Methods, Algorithms, Visualizations and Evaluation
- Plans (1 page)
 - Challenges and obstacles
 - Plans to address them

See the Final Report section below for a more detailed description of some of these sections.

For your midterm progress presentation you will give a brief overview of the material you present in your midterm report. Your presentation should be about 4 minutes and will not include visual aids (although you may draw on the board if it helps). Your goal should be to describe your project, progress, challenges and plans so the class can give useful feedback and suggestions.

Submit your report (pdf,word) and the interesting aspects of your source code using CS Portal.

²Where understand means, I could probably figure out how to write the code to implement this given a few days of focused effort?

³See the [NIPS paper style guide](#)

Final Report and Presentation

Your final report will be a write up in the format of an academic paper (see [the NIPS style guide](#) for an example (8 pages maximum, 6 pages minimum, including references). The final report should include an expanded description of your methods and results as well as a conclusion that summarizes the importance of your work (it may be appropriate to speculate about future work). Your report should not include the plans/challenges section, but a description of the challenges may be necessary to sufficiently describe your contributions (this is ok, but should be written appropriately to support this). If you are lucky, the introduction and problem statement may not need to change much from the mid-term report, however, you should feel free to change these as you see fit.

Suggested Format for Final Presentation ⁴You will prepare a 10-15 minute (more if you are a group) presentation, probably using Powerpoint or Keynote, that summarizes your project. You should address each of the topics you will include in your final report, but you can omit some of the boring details. I encourage you to attend the CS Keeping Current Seminar on Wednesdays at 4pm in the Marksbury building if you want to see some example talks for guidance.

Suggested Format for Final Report Below are guidelines on how to write up your report for the final project. **These are only guidelines**; you will need to adjust it to the problem you are investigating, but try to structure your report along these suggestions and make it look like an article. Suggested⁵ sections and layout for your final report:

(1) Introduction

Motivate and abstractly describe the problem you are addressing and how you are addressing it. What is the problem? Why is it important? What is your approach? What is the goal of your paper?

Provide a short discussion of how it fits into related work in the area.

Summarize the basic results, conclusions and contributions that you will present. All these apply equally to experimental papers, survey papers or theoretical papers.

(2) Related Work

This part need not be exhaustive, but you need to know about some of the related work. Discuss the problem and method in the related work. How is your problem and method different? Why is your problem and method better? You **must** include a reference to at least three alternative sources (preferably academic publications that you have read, this may not include the course textbook or a web page). For each reference you **must** provide a short summary (several sentences should suffice).

(3) Problem Definition and Algorithms

Task Definition Introduce the model and/or problem you are studying and define the notation you are going to use. Precisely define the problem you are addressing (e.g., formally specify the inputs and outputs). Elaborate on why this is an interesting and important problem.

Algorithm Definition If you study learning algorithm(s) experimentally this is the place to present it. Describe in reasonable details the algorithm(s) you are using. A pseudo-code description of the algorithm you are using is frequently useful. Depending on the context, it may be useful to trace through a concrete example, showing how your algorithm processes this example.

Expectations In case of an experimental study, discuss what you hope to achieve. How do you expect each algorithm to behave and why. Try to justify your hypothesis as rigorously as possible. Discuss how your expectations drive your experimental design.

(4) Experimental Evaluation (if relevant)

Methodology What are the criteria you are using to evaluate your method? Describe the experimental methodology that you used. What is the training/test data that was used, and why is it realistic or interesting? What performance data did you collect and how are you presenting and analyzing it?

Results Present the quantitative results of your experiments. Graphical data presentation such as graphs and histograms are frequently better than tables. What are the basic differences revealed in the data. Are they statistically significant?

Discussion Is your hypothesis supported? What conclusions do the results support about the strengths and weaknesses of your method compared to other methods? How can the results be explained in terms of the underlying properties of the algorithm and/or the data.

⁴By this point you should have all of your source code written.

⁵Originally by Dan Roth.

(5) Theoretical Evaluation (if relevant)

If you are writing a theoretical paper and/or a survey, this is the place for your analysis and contribution. Try to make it clear what parts of the work are presentation of known work, what is given a new look by your presentation and what is novel in your view of the problem.

(6) Conclusion

Briefly summarize the important results and conclusions presented in the paper. What are the most important points illustrated by your work? What are the major shortcomings of your current method?

For each shortcoming, propose additions or enhancements that would help overcome it.

As usual, submit your slides (ppt) and report (pdf,word) using CS Portal prior to the dates listed above.

4 Project Guidance

There are many possible final projects for this course and you are free to choose your topic, as long as it represents sufficient effort and is related to machine learning. The key is to identify the problem you will solve, the data you will use, the methods you will use to evaluate your work and a reasonable plan.

Some ideas to get you started:

- Implement a learning system for a specific application. The key here is defining a interesting problem of suitable difficulty:
 - choose data and a problem from your research area, or
 - choose data and a problem that interests you:
 - * classifying handwritten digits or faces ([Matlab format data files](#))
 - * [various computer vision datasets](#)
 - * [UCI Machine Learning Repository](#)
 - * [Public Machine Learning Data](#) from [Datamob](#)
 - * [Machine Learning Datasets](#) from [KDNuggets](#)
 - * [UK Governmental Data](#)
 - * [US Governmental Data](#)
 - * Predict future [crimes](#)?
- Extend some recent work on machine learning: Find a recent paper at [ICML](#) or [NIPS](#), and do it better.
- You should spend some time thinking about it first, but if you don't know what to do, feel free to ask me for advice (be sure to tell me what interests you, a particular kind of data? a particular method?). I will prepare a couple of reasonable pre-defined projects for such cases.

Some examples of final results from similar projects:

- [Cornell 6780: Fall 2010 projects](#)
- [Cornell 6780: Fall 2009 projects](#)
- [CMU 10-701: Fall 2007 projects](#)
- [CMU 10-701: Fall 2009 projects](#)

Programming Languages and Helpful Libraries

You are free to use any programming language you want for your final project. We recommend using a higher level language such as Matlab or Python (as opposed to C or C++). You may find that you want an even higher-level API that is strictly focused on machine learning (several examples are listed on the [Piazza class resources page](#)).