CIS 419/519 Introduction to Machine Learning
Course Project Guidelines

1 Project Overview

One the main goals of this course is to prepare you to apply machine learning algorithms to real-world problems. The final course project will provide you the opportunity explore such an application of machine learning to a problem of your own choice.

Projects must be completed in **teams of two to three students**. All teams (regardless of size) are expected to produce a project of equivalent scope, so you should only work in a team of two people if you are willing to take on substantial additional work; I would strongly recommend teams of three. If you have a particularly ambitious project idea that cannot be completed by a team of three people, you may propose a team of four students, but you must have a strong justification for such a larger team. You may not complete the project solo.

Milestones and Deadlines

- **Project Proposal**: due Monday, Oct. 6, 2014 11:59pm
- **Project Status Report**: due Monday, Nov. 10, 2014 11:59pm
- **Final Report & Summary Slides**: due Monday, Dec. 8, 2014 11:59pm (no late days!)

Grading Breakdown

- Project proposal: 10%
- Project status report: 10%
- Final summary slides: 10%
- Final report: 70%

Evaluation Criteria

- Technical quality (i.e., Does the technical material make sense? Are the things tried reasonable? Are the proposed algorithms or applications clever and interesting? Do the authors convey novel insight about the problem and/or algorithms?)
- Significance (Did the authors choose an interesting or a "real" problem to work on, or only a small "toy" problem? Is this work likely to be useful and/or have impact?)

- Novelty of the work (Is the proposed application and approach novel or especially innovative?)
- Clarity of presentation (Is the presentation clear? Could we reconstruct the method entirely from the report?)

Students enrolled in the graduate version of the course (CIS 519) will be expected to complete a project of significantly higher scope, quality, and polish than students in CIS 419. Specifically, CIS 519 projects are expected to be of sufficient quality for a machine learning workshop publication. Teams may include students from both CIS 419 and CIS 519, but projects from combined undergraduate/graduate teams will be graded under the CIS 519 criteria.

Although I encourage you to implement your project in python using scikit_learn, you may use other software or programming languages if you have a particularly compelling reason.

2 Choosing a Topic

Your first task is to identify a topic for your project. One of the best ways to identify a topic is to choose an application domain that interests you and identify problems in that domain. Then, explore how to apply learning algorithms to best solve it. Let the problem drive your choice of technique, rather than the other way around. Most projects will be based on particular applications.

Alternatively, you can also choose a problem or set of problems and then develop a new learning algorithm (or novel variant of an existing learning algorithm) to solve it. Although CIS 520 is intended more to prepare you to develop novel learning methods than CIS 419/519, you may choose to develop a novel learning method (or novel variant) if you want a challenge.

Regardless, most projects will combine aspects of both applications and algorithms. Your project **must** include an evaluation on real-world data (i.e., not a “toy” domain or synthetic data).

2.1 Ideas

Many fantastic course projects will come from students choosing either an application that they’re interested in, or picking some sub-field of machine learning that they want to explore more, and working on that topic. If you’ve been thinking about starting a research project, this project may also provide you an opportunity to do so.

Alternatively, if youre already working on a research project that machine learning might be applicable to, then working out how to apply learning to it will often make a very good project topic. Similarly, if you currently work in industry and have an application on which machine learning might help, that could also make a great project.

Here are a few other sources of project ideas:

Course projects/suggestions from similar courses at other universities

- Stanford, 2013: `http://cs229.stanford.edu/projects2013.html`
- Stanford, 2012: `http://cs229.stanford.edu/projects2012.html`
- C. Guestrin, CMU: `http://www.cs.cmu.edu/~guestrin/Class/10701/projects.html#datasets`
- Ray Mooney, UT: `http://www.cs.utexas.edu/~mooney/cs391L/project-topics.html`
- Amy McGovern, OU: `http://www.cs.ou.edu/~amy/courses/cs5033_fall2014/index.html`

Eric's list of project suggestions

- Extend an active learning technique (which queries the user for labels) to use other sources of feedback that are richer than binary labels, such as equivalence sets, distribution examples, measures of ”typicality” of the instance, or some other idea of your own.
- There are multiple ways to combine kernels together to create new kernels (addition, multiplication, etc.). Develop an SVM-based learning algorithm that tries a number of kernels and their combinations in a principled manner to find the optimal separator for a data set.
- Multi-view learning is typically applied to supervised or semi-supervised classification scenarios. Instead, apply it to unsupervised clustering or constrained clustering.
- Write a reinforcement learning agent to play Mario or Tetris using the RL-Glue framework. The framework is available at `http://glue.rl-community.org/wiki/Main_Page`, and you might be interested in the steps described in `http://www.eecs.wsu.edu/~taylorm/2010_cs414/Project1.pdf`.
- Design an algorithm for transfer learning that improves image classification in some categories of the Caltech 256 data set based on transfer from other categories, or object recognition in the MIT objects and scenes data set, or indoor scene recognition. Transfer could also be used to improve image segmentation in the Berkeley image segmentation data set.
- Often times, users have an idea of the classifier they are looking for, even if the data does not directly support it. Design an interactive method for building a model in collaboration with a user. For example, perhaps the user knows that particular attributes should be in the first few splits of the decision tree, even if there isn't enough data to support it, so the tree could be interactively built in collaboration with the user. Or, perhaps the user knows that particular factors are especially important.

Look through papers from recent machine learning conferences

- Int. Conf. on Machine Learning 2014: `http://jmlr.org/proceedings/papers/v32/`
- Int. Conf. on Machine Learning 2013: `http://jmlr.org/proceedings/papers/v28/`
- Neural Information Processing Systems: `http://papers.nips.cc/`

Final Advice Pick a topic that you can get excited and passionate about! Be brave and feel free to propose ambitious things that you're excited about. Finally, if you are not certain what would make a good project, we encourage you to e-mail us or come to instructor/TA office hours to talk about project ideas.

3 Project Proposal

Your first deliverable is a one-page project proposal that includes the following information: project title, names of all teammates, and a description of what you plan to do. Your proposal must be one page in length, single-spaced with 12 point font.

You should write a compelling proposal that describes your project in detail and demonstrates that you have the understanding and ability to complete it. Your proposal should also discuss sources of real-world data for your chosen application or how you plan to obtain real-world data. Since you may wish to use machine learning methods that we have not yet covered, you may need to read ahead. Do not worry if there are particular aspects of the project that you can't answer currently (such as which ML method is best); this is a proposal for future work, after all. However, your proposal should demonstrate that you've started to think through the various issues involved with your project and present a compelling argument in support of it.

If you are not certain exactly what the proposal should include, I recommend that you consult Heilmeier's Catechism[1], excluding the cost and time estimate). Imagine that you are bidding for funding, so your proposal should be a compelling argument that convinces me your project is a good idea, important, and that you have the capability to complete it successfully. And, you must do all of that in only one page.

Save your proposal as a PDF file named `proposal-Lastname1Lastname2Lastname3.pdf` (where the filename lists the lastnames of all team members in alphabetical order; for example, "proposal-JonesSmithXu.pdf"). Make certain that your filename follows this naming convention.

You should also submit a plain text file named `README` that has exactly three lines of text: the first line should be a comma-separated list of the names of all teammates, the second line should be a comma-separated list of PennKeys (in the same order), and the third line should be the title of your project. For example,

```
Jane Doe, John Smith
jdoe, jsmith
Title of My Legendary Machine Learning Project
```

Submit your project proposal by running the following command on eniac:

```
turnin -c cis519 -p projectproposal README proposal-Lastname1Lastname2Lastname3.pdf
```

Only ONE person from each team should submit the proposal and README. Note that even students in CIS 419 should submit to the cis519 course, as shown in the command above (we have only one course account, cis519, for both sections of the course). You can check that your submission was received by running:

```
turnin -c cis519 -p projectproposal -v
```

to list the files submitted. You should see one listing for each file you submitted. Subsequent submissions will overwrite earlier submissions; if you re-submit files after the due date, your submission will be counted as late regardless of whether or not the file contents changed.

4 Project Status Report

The project status report is due approximately one month before the final submission, as is intended to make certain that your project is on-track. It should describe what you've accomplished so far and very briefly state what you have left to do.

You should write your status report as if it is an early draft of your final project report. Specifically, you can write it as if you're writing the first few pages of the project report, so that you can re-use most of the text in your final report. Your status report should be at most 2 pages long.

[1] `http://en.wikipedia.org/wiki/George_H._Heilmeier#Heilmeier.27s_Catechism`

Please write the status report (and final report) keeping in mind that the intended audience is Prof. Eaton and the TAs. (Thus, for example, you should not spend two pages explaining logistic regression.)

Your status report should be in the same LaTeX template as your final report (see the next section for details) and should be saved as a PDF file named `status-Lastname1Lastname2Lastname3.pdf` (make certain to follow the naming convention exactly).

You should also submit a plain text file named `README` that has exactly three lines of text: the first line should be a comma-separated list of the names of all teammates, the second line should be a comma-separated list of PennKeys (in the same order), and the third line should be the title of your project.

Submit your status report and README by running the following command on eniac:

```
turnin -c cis519 -p projectstatus README status-Lastname1Lastname2Lastname3.pdf
```

Only ONE person from each team should submit the status report and README.

5 Final Submission

Your final submission will consist of two deliverables: (1) a final report, and (2) a set of summary slides. Remember that late days cannot be used for the final project submission.

5.1 Final Report

Your final project report can be at most 5 pages long (include all text, appendices, figures, references, and anything else), and must be written in the provided LaTeX template. If you did this work in collaboration with someone else, or if someone else (such as another professor) had advised you on this work, your report must fully acknowledge their contributions.

At a minimum your final report must describe the problem/application and motivation, survey related work, discuss your approach, and describe your results/conclusions/impact of your project. It should include enough detail such that someone else can reproduce your approach and results. For inspiration on what should be included, see the project reports available on the links provided in Section 2.1. You will likely end up with a better report if you start by writing a 6-7 page report and then edit it down to 5 pages of well-written and concise prose.

In addition, your report must also include a figure that graphically depicts a major component of your project (e.g., your approach and how it relates to the application, etc.). Such a summary figure makes your paper much more accessible by providing a visual counterpart to the text. Developing such a concise and clear figure can actually be quite time-consuming; I often go through around ten versions before I end up with a good final version.

We know that most students work very hard on the final projects, and so we are careful to give each report sufficient attention. We (specifically, Prof. Eaton) will personally read every word of every report. After the class, we are also considering posting the final reports online so that you can read about each others work. If are okay with having your final report posted online, be sure to give us explicit permission to post it in the README file, as described below.

5.2 Summary Slides

In addition to the final report, you are also required to prepare a two-slide overview of your project. Think of these slides as a concise presentation of your project, highlighting the problem you worked on, your approach, and your results / contributions.

You may use any format you wish for the slides, but you are limited to only two slides. The goal is not to cram as much as possible into two slides, but to provide a clear and concise presentation of the main points of your project. You should avoid any font smaller than 14 pt, and most of your text should be around 18pt or larger. The best slides will use lots of graphics along with some text. You are welcome to re-use these graphics in your project report, and you may reuse the summary figure from your report in your slides.

Although this is only two slides, you should be aware that it is actually quite difficult to present an entire project in such a concise manner while still being clear. Do not leave these slides to the last minute; you will likely need to make several versions of these slides until you narrow them down to the essentials, and so they might actually take a while.

5.3 Submission Instructions

Save your report as a PDF file of 5 pages or less named `report-Lastname1Lastname2Lastname3.pdf`. Save your summary slides as a PDF of 2 pages named `summary-Lastname1Lastname2Lastname3.pdf`. (As before, lastnames should be in alphabetical order.)

You should also submit a plain text file named `README` that has exactly FOUR lines of text: the first line should be a comma-separated list of the names of all teammates, the second line should be a comma-separated list of PennKeys (in the same order), and the third line should be the title of your project. The fourth line of text should be either "We grant permission to post our CIS 419/519 final project report and summary slides on the publicly accessible CIS 419/519 course website" if you are willing to have your final project report posted publicly, or "Do not post our final project report" if you are unwilling to have your report posted. (Note that this choice will not affect your grade, and projects will only be posted after final grades are submitted.) For example,

```
Jane Doe, John Smith
jdoe, jsmith
Title of My Legendary Machine Learning Project
We grant permission to post our CIS 419/519 final project report and summary slides
on the publicly accessible CIS 419/519 course website
```

Submit your project by running the following command on eniac:

```
turnin -c cis519 -p projectfinal README report-Name1Name2Name3.pdf summary-Name1Name2Name3.pdf
```

Only ONE person from each team should submit the final project submission. You can check that your submission was received by running:

```
turnin -c cis519 -p projectfinal -v
```

to list the files submitted.