CS 475: Final Projects

Prof. Mark Dredze

Version 1.0

Important Dates:

Friday November 4: Project proposals due
Monday November 28: Progress reports due
Monday December 12: Final project writeups due

All submission should be by email to cs475@cs.jhu.edu.

The class project is worth 25% of your final grade. You should plan on spending an appropriate amount of time on the project (30 hours per person, 60 hours per two person project.) Projects will be done in groups of 2. Both people on a project will receive the same grade, so you are responsible for choosing a responsible project partner.

Project Description

Implement and apply a machine learning algorithm to solve a problem. In doing so, you must demonstrate knowledge of something we've learned in class, or a general topic in machine learning. A simple black box application of machine learning is insufficient. You may also implement a machine learning algorithm based on a research paper of interest. You should explain how you will implement the algorithm and what new experiments and settings you will explore.

This is a good opportunity to work on a research project, especially if you are a PhD student. My general advice to PhD students is to work on something that can become a paper, although you probably won't get that far in the class project. Projects on learning theory are also acceptable, as well as any machine learning topic even if we didn't cover it in class.

Your proposal should identify which machine learning concepts and ideas you'll be testing, the data you will need for this project, and an expected outline for your final writeup. Data collection cannot be a significant part of the project, so be clear as to where the data will come from and what state it is in. Additionally, please include:

- Resources you will use (libraries, data, etc.)
    - State what libraries you will use. If you will implement everything yourself, say so. You are welcome to use any available software library, and write your project in any programming languages you' like.
    - Say what data you will use. In almost all cases, you CANNOT collect your own data. This is too time consuming. Select projects for which data is available or easy to obtain.

    - The most common reason I find for not approving a project is a bad data plan. When you hand in the proposal, you should already have the data you plan to use in place. You should not rely on collecting it yourself, creating it, etc.
- Methods: a description of the machine learning method you will use
    - Be specific. Do not write "we will try various methods." Say what you are going to use.
- Milestones: the deliverables for the project. This should include must achieve, expected to achieve, and would like to achieve
    - A milestone CANNOT be "obtain X% accuracy." A milestone needs to be something you know you can achieve. A better milestone would be "implement X algorithm" or "develop features for Y dataset."

- An outline of the final writeup. This is a description of what I can expect you to deliver in your final report.
- Bibliography- a list of relevant citations

Project Proposals

Your project proposal should be at most two pages. This isn't a hard limit, but if you are writing more than two pages you are likely writing too much. A single proposal is needed for each team and will be submitted by email. Your proposal should include your name, email, JHED and the details listed above. **Please complete the provided project proposal template.**

Progress Reports

The progress reports should be short. 1 paragraph or less is sufficient in most cases. Acceptable progress reports are: "We've done nothing yet" or "Everything is going according to plan." One report is needed per group. Submit by email. Your report should be written in the text of the email. Only include a PDF is necessary. Your progress report should:

- briefly summarize what you've done so far
- compare to your proposal: are you on track, a bit off track, major changes needed
- describe any changes you anticipate in your final report from your proposal

Final Project Reports

The final report should be a SIX page PDF (not including references). This is not a requirement but a strong suggestion. Slightly longer or shorter writeups are acceptable, and you may include as many appendices as needed. The format of your report should use the NAACL 2010 format, available in both Latex and Microsoft Word: http://naaclhlt2010.isi.edu/authors.html

Please be sure to include your names in the report.

Your report should follow your proposed outline. In addition, it should include the following:

- An introduction and background section that explains relevant material and situates your work in the context of existing literature. Do not assume the reader knows the details of your field. You must include sufficient references.
- An explanation of all machine learning techniques used as well as an explanation of why they were used. This should reflect the knowledge you've learned by taking this course. Don't just say "I used an SVM" but explain what an SVM is in detail. Remember, this report is for a machine learning class, and that should be the focus of the writeup.
- A detailed description of the work done for the project. Note that in a conference or journal publication, this section is focused on the science and not the details of the actual work. For these writeups the focus is shifted. I want to know what you did, i.e., how you prepared data, the libraries you used, code you wrote, problems you had, etc. This is where you explain exactly how you spent your time on the project. This should convey exactly how much you did.
- A description of the results you obtained. Examples of actual output (pictures, sentences, etc.) would be great if appropriate. Please include details for any tests you ran, even if they weren't "interesting" or didn't work. Again, this should highlight the work you did. You should include an analysis of the results using knowledge of machine learning (I think this worked or didn't work because this algorithm assumes...)
- Comparison to proposal. This should include a detailed list (bulleted lists are appropriate) of your progress measured against your proposal. This part is similar to your progress report in that it includes a list of what you were able to accomplish, changes to your original plans, etc.

I found the following writeup advice from Tommi Jaakkola for his 2004 machine learning class. I think it's very relevant so I'll just quote it:
"We expect that the ``size'' of your project should be equal to about the amount of work required for ...(Mark: 25% of your grade). The project, however, should be in some sense ``complete''. By this we mean that you should not ignore relevant machine learning issues. In the final report you shouldn't just say what you did but also why it was a reasonable thing to do given the course material... You shouldn't worry about getting ``great'' results. The idea and your understanding of the machine learning issues involved are much more important than getting ``great'' results."

Grading
The project is worth 25% of your grade You are expected to spend an equivalent of hours per person on the project (about ~2.5 homework assignments). I will be evaluating your project based on the writeup using the following guidelines.

Did the project constitute a sufficient amount of work for a final report?
Is the writeup clear and understandable?
Does the writeup demonstrate an understanding of machine learning and are relevant machine learning concepts and algorithms explained?
Does the writeup include sufficient references and descriptions of related work?

Was progress made along the proposed goals and if not, is there sufficient explanation given?
Does the writeup clearly present the performed work?
Are the results analyzed and understood using machine learning concepts?

Code
Please submit your code along with the writeup. You do not need to include directions for running the code nor do I expect it to work if it needs external libraries, data, etc.

Frequently Asked Questions
I want to work on a project that I am already doing for another course or independent study. Is that allowed?
All work done for the project cannot be double counted for another class or credit. However, you can do synergistic work. For example, suppose you are working on some fancy new model for another course's project. For this project, you can propose to develop new training methods for this fancy model. The model itself doesn't count as your project, but the training methods do. You can work on them both and, provided each is sufficient work for a project, that's fine. However, I will insist that you have a plan such that if one fails, you won't fail the other project. In this case, I'll ask that you work on training for a simpler model which you know works. If you end up working with the advanced model that's great, but not required.

Does the project have to be novel?
No. You can do anything you want, even if its been done before. This isn't research; it's a class project. A perfectly acceptable project is to implement a paper you like.

What programming language should I use? Can I use existing libraries?
You can program in anything you want. You can use any library you want. Please indicate these decisions and dependencies in your proposals and write-ups. The total proposed work must be reasonable. For example, if you propose to implement an SVM, but are using an external SVM library, that's not enough work. However, you can propose an application of an SVM and then use an existing library.

Does the project need to use topics covered in class?
No. Anything in machine learning is fair, even if we don't cover it in class.

Can I do a project that has been done already? For example, I see an old Kaggle competition that has published systems already. Can I implement one of those systems?
Yes, you need not work on anything novel. You can copy a previous system or previous paper that has already been published. If you do, I encourage you to do something new in the work, which could be new features, a new evaluation, etc. Your work doesn't have to be novel.

How much time can I spend on data preparation?

This should be a minor part of the project. I prefer if you have the data on hand before the project starts.

Once I get started, are small changes to the project ok? When do I have to tell you when I make a change?
Changes are fine and expected. You only need to inform me if it is a major change to your plan, like you need to come up with a new project, or the very first step totally failed and you need to come up with a new plan. This is rare.

Can I build an app (mobile, facebook, etc) as part of the project?
No. Building these tools require a lot of work that has nothing to do with machine learning. The goal of the project should be to focus on machine learning methods. You are welcome to build these methods with some other goal in mind which you will implement after the project is completed.