EECS 349 Machine Learning: Group Final Project

A Group Assignment

For this assignment, you may form groups of 3 (preferred) or 2 (possible in some cases). No solo projects are allowed. Groups are to complete the assignment as a unit.

Project

Your project is an opportunity for you to explore an interesting machine learning problem of your choice in the context of a real-world data set. This project can be whatever you want, as long as it is focused on getting a machine learning system to learn. Example projects might be building a decision-tree-based spam filter, making a neural-net based face recognition system, or building a tic-tac-toe player that learns from experience with reinforcement learning.

There are no restrictions on programming language or machine learning method.

Use of existing datasets and software packages is fine.

Handing in Deliverables

- Due dates and point values for project deliverables are on the course calendar.
- All deliverables are due by the start of class on the specified time on the specified date.
- The final paper (extended abstract) must be made available on your project website.
- The initial project proposal, the revised project proposal, the status report, and the extended abstract must all be in the form of an ACM conference paper.
- Templates for ACM conference paper format can be found here: http://www.acm.org/sigs/pubs/proceed/template.html
- Pay attention to the page limits for each written document. I will stop reading your work when I reach the page limit.
- The only acceptable file formats for written documents is Adobe Acrobat. Documents in other formats will not be accepted.

Project Proposal

You must turn in a brief project proposal (2 page maximum, 1 page minimum in the ACM conference paper format). The initial submission will be critiqued. You will revise the project proposal in response to this critique and resubmit it.

Include the following information: *(italics indicate hints and examples as to how to answer these questions)*

- Project title & group membership
- What learning problem do you want to solve? *We would like to make a weather prediction system that predicts future weather given recent observations...*
- What is the intellectual interest and practical utility of doing this project? *This is an interesting problem because….This is a useful problem to solve because….This is a challenging problem because….*
- Describe the dataset you will use to test and train your system.
  - Where will you get it? Be specific
  - Do you have to label it yourself or is it already labeled?
  - How big is it? How many examples?
  - How is it encoded?

EECS 349 Machine Learning: Group Final Project

*We plan on collecting <temperature, barometric pressure, humidity> readings taken every hour in Chicago for the last year. This is available on the weather.com website. Unfortunately, it is not presented as a spreadsheet of comma separated values. Instead, the user must type in an hour and a day to see those values displayed as XML for a website. We will screen-scrape this data (see a following section on how) and collect the data for every hour in the last year. This will give us 365*24 =8760 hours of data and about 8750 sequences of 10 hours that we can train and test the system on. This data will be stored in a csv file as follows*

Month	*Day*	*Hour*	*Temperature*	*Barametric*	*Humidity*
3	*9*	*10*	*45*	*30*	*.75*
3	*9*	*11*	*48*	*30*	*.70*

- What is the EXACT TASK that the system will learn to do? (What is the exact input, what is the exact output?). This should be more specific than "We will predict the weather". It should be more like:

  We will predict the temperature, humidity and barometric pressure, given past weather readings. Given the statistics for the past 10 hours, the system will output a prediction for the temperature, barometric pressure and humidity on the next hour. The system will take 10 triplets, <temperature, barometric pressure, humidity>. Each triplet will represent the readings for those three values at hour 1 to hour 10. Our system will be a regression system that will predict the value at hour 11.

- What aspect of the task is your system going to optimize/improve/learn?

  This is a prediction task on a time series. We are going to use a Markov model and the thing it will learn is to output a most-likely 11th hour's set of values, given a sequence of 10 previous readings. This will be learned using the forward-backward approach. Once the model is learned, we'll feed it a sequence of 10 readings and have it predict the most likely 11th.

- What measure will you will use to evaluate performance of the system? Be specific. Don't say just "We'll measure performance."

  Our ground-truth is the actual readings (barometric pressure, temperature, humidity) at hour 11. We'll treat (barometric pressure, temperature, humidity) as 3 coordinates in a Euclidean space. We can then measure the Euclidean distance from our actual readings to the predicted values. That will be the error measure. We will average this over a large number of trials to get a mean error.

- What is the baseline approach you will compare to? Random performance? An existing system? Human performance? Be specific.

  The baseline approach we will compare to is to just assume that the weather readings at hour 11 are exactly the same as at hour 10. This is reasonable, since, lacking any other information, this is what someone would normally do. If our system can do better than this, we'll feel confident it learned something

- What software will you need to write?

  Who will write it?

  How long will that take?

  We will need to write a screen-scraper to pull the weather data down from the site weather.com. Beatrice did something similar for another class, therefore, she will be in charge of this portion of the project. The expected completion date

to collect 1 year of data (24 readings per day) for our target city (Chicago) is shown in the project milestones section.

- Describe any existing software packages you will use. Answer the following questions.

  Where will you get them?

  Have you already tried them out?

  We will use the HTK Markov model toolkit (http://htk.eng.cam.ac.uk/docs/history.shtml) . Larry downloaded it and tried training a Markov model on one of their toy problems provided with the kit. We're confident this will work.

- Related papers to read.

  Include a minimum of 4 relevant research papers.

  Give bibliographical info.

  READ a couple of them before submitting the proposal.

  Try scholar.google.com as a starting point.

  Go back and look at the ACM template. This shows you how to format your citations. I want full bibliographic information.....AND hyperlinks to papers. A good place to find papers is scholar.google.com

- What is the job of each team member?
- Milestones

  What will you complete by the date of the status report?

  What will you complete by the date of the final?

  Who is responsible for each milestone?

  The table below outlines our milestones and expected completion dates.

Milestone	*Group Member(s)*	*Date*
Data collected from web	*Beatrice*	*Nov 16*
Testing software written	*Lourdes*	*Nov 20*
Baseline weather predictor done	*Larry*	*Nov 20*
Initial sanity tests of baseline & experimental setup	*Lourdes, Beatrice*	*Nov 21*
Initial Markov Model done	*Larry*	*Nov 28*
Experiments complete	*All 3*	*Dec 2*
Poster Printed	*Larry*	*Dec 7*

Meeting With Professor

You will meet **as a group** with the professor on several dates. See the course calendar for dates and times. You will give a short verbal report on the progress your group is making. Your grade will depend on your ability to clearly and briefly describe the high-level approach (your task, the approach, how you are measuring success, etc), as well as your ability to answer concrete questions on your progress. This is also a chance to get feedback, so use it as such.

EECS 349 Machine Learning: Group Final Project

Written Status Report

You will hand in a **one-page** report describing how you did or did not meet the milestones you put in your project proposal. You will describe issues you see that may cause problems in completing the project as planned. You will also lay out a set of clear, concrete milestones to reach by the final project due date.

Project Website

You are to build a website that gives an overview of your research project.

You are free to format the website as you see fit.

You are free to host the website wherever you like.

That said, there will be certain required elements that we will look for. The website **must** contain the following:

1. The project title (prominently displayed)
2. The name of each project member (prominently displayed)
3. At least one contact email address (prominently displayed)
4. The name of the course, university and professor
5. A 3-4 paragraph synopsis of what this work is "about"
    - motivate the problem: (what is the thing you're trying to solve and why anyone should care
    - describe your solution in very high level terms (what kind of learner did you use, how did you apply it)
    - describe how you tested and trained it (what your dataset was, how you measured success)
    - describe some results (how well it did in no more than a paragraph)
    - points them to the extended abstract for more detail
6. A minimum of one pretty picture or graph that illustrates your work…with a caption and to explain what the viewer is looking at.
7. A link to a copy of your Extended Abstract final report (see below). This is the kind of thing that you would submit to a conference where your work is accepted for poster presentation.

Note, these are minimal requirements. Having all these features guarantees an OK grade, not an A. Doing a good job with these things guarantees a good grade.

Note also, the plan is to make a unified final project website that contains links to all class projects. This website will be reachable from the class website and I will email the URL to the class mailing list so that you can all see each other's projects.

Poster/Demo

The final exam for the class will be a poster/demo session. Each group will bring a poster outlining their project. Those groups for whom a demo might be appropriate are welcome to provide a demonstration. All members of the group must attend the poster session. Grading will be based on the work, the effectiveness of the presentation, the poster and (where applicable) the demonstration.

Items 1 through 6 from the "Project Website" requirements are ALSO required for the poster….although item 5 does not necessarily have to be in paragraph form. As with the website, these are minimal requirements. They do not guarantee an A. Doing a good job in the work, the poster and the presentation are the way to get an A.

EECS 349 Machine Learning: Group Final Project

Extended Abstract

The final extended abstract must be **no less than two full pages and no more than three full pages**. It should be structured like a research paper in a computer science conference. The same points discussed in item 5 of your website requirements must be hit in this abstract.

Note: You CAN and SHOULD reuse text AND images between the paper abstract, the poster and the website.