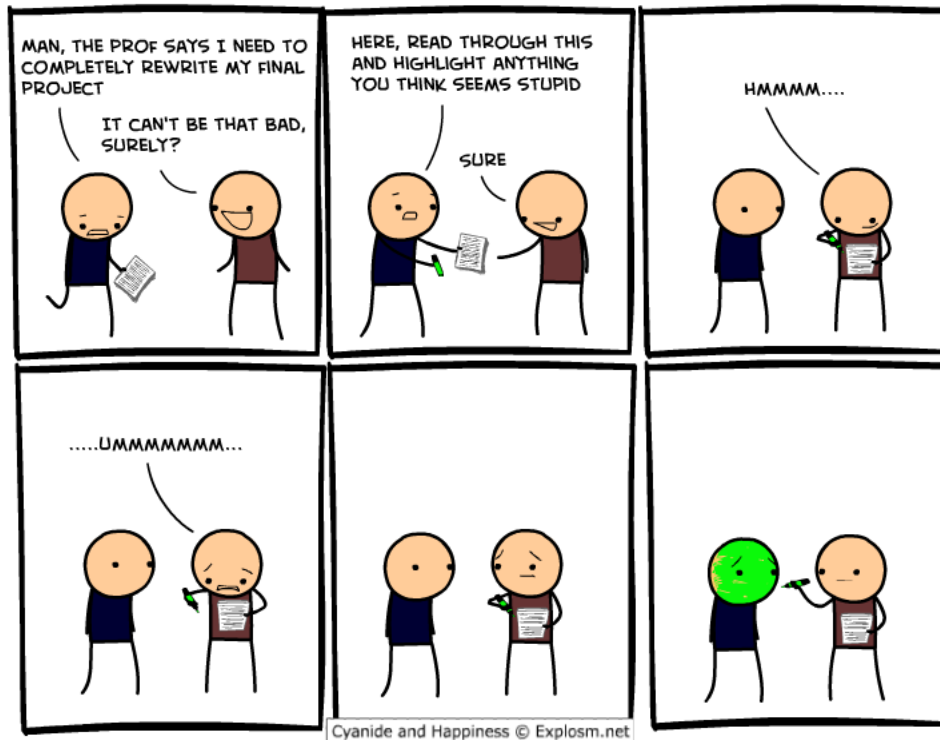# CS451 - Final Project
Fall 2013

Overview

In this class we have looked at a number of machine learning techniques, and have examined a few in-depth in assignments. The purpose of the final project is to explore a topic we have examined (or not examined - but related to machine learning) that's interesting to you in more depth as an experimental project.

The project should meet the following guidelines:

- Your project should relate to something related to machine learning. I give a few ideas below, but I encourage you to be creative. Feel free to ask me if your idea is appropriate. *Find*

something that you're excited about and interested in since you'll be working on this for the rest of the semester!

- Your project *must* include a solid experimental examination. For example, if you implement a classification algorithm, I would expect to see optimization of hyperparameters (on the development set), an evaluation on at least one data set **and** a brief analysis of the model learned (e.g. looking at the weights of a linear model).

- Your project should be in a pair or group of three. If you'd like to do it solo or in a group of four, please come talk to me.

You may code in whatever language you would like and may (and are encourage to) use any external resources you would like including both code and data.

Schedule

date	description
11/20	In-class project proposal
11/22	Project proposal write-up
11/27	Status report
12/6	In-class presentations
12/8	Writeup and code

Project proposal presentation [5 points]

Your first task will be to come up with your project group and figure out what you'd like to work on.

In class on Wednesday you will have 3 minutes to present the following information (*make sure you've thought about each of these things*):

- Who is on your project.

- What do you plan to do for your project:

  – Brief overview of what you plan to investigate
  – What experiments you plan on running
  – The data you will use

- Any other relevant information

We will then have a few minutes of discussion from the class to help you finalize any ideas.

Project proposal write-up [10 points]

Submit online by 6 pm as "proposal"

Your project proposal should be a 1-2 page write-up with clear section headings containing the following information:

- **Team:** Members of the team. I'm *strongly* encouraging groups of 2 or 3. If you want to work solo, please come talk to me.

- **Summary:** A one paragraph description of your project including:

    - What you plan do to for the project. Be as specific as possible!
    - What experiments you will run and what metrics you will use for evaluation.

- **Resources:** What resources you will use/need including code, data, etc. You may use any resources you can find, including code you have written for this class or other classes, code provided with the book, data you find on the web, etc. If you would like a resource and can't find it, ask and I might be able to help you. However, you must have found *ALL* resources by the time you submit your proposal. Come talk to me (early) if you're having trouble finding appropriate data.

Status report [10 points]

Submit online by 6pm on due date as "status"

Your project proposal *must* include the following (make explicit headings):

- **Members:** Names of team members

- **Summary:** A one paragraph summary of what was accomplished since last checkpoint.

- **Results:** One or more numerical results. This could be some analysis of a data set, a preliminary result from your system, etc.

- **Problems:** Any problems/issues that have arisen that might keep you from finishing your project.

- **Hours:** The number of hours each person put into the project since the last checkpoint

- **Code:** A snapshot of your current code-base. You may submit this as a link to an online repository (e.g. GitHub) or just a directory of code.

This is not meant to take you a long time, but please do spend a little bit of effort putting this together.

Paper and code [80 points]

Submit online by 11:59pm on due date as "final"

Code: You may submit this as a link to an online repository (e.g. GitHub) or just a directory of code.

Paper: You should submit a short (2 pages max) paper that includes:

- Introduction. *Briefly* the technique/problem/application that you investigated

- Your experimental setup. What data did you use? How did you setup the experiment (10-fold cross-validation, etc.)? What did you use for evaluation? How did you decide if results were significant? We have talked a lot about proper experimentation in this class, so an important component of this project will be that you have setup a proper experiment (see the notes for a refresher on this).

- Your results. These should be stated concisely and should include supporting tables, graphs and figures.

- Conclusions. Summarize your findings.

Presentation [10 points]

Each group will give a short (3 minute) presentation of their work during the last day of class. Your presentation must include 1-2 powerpoint slides which *must be e-mailed to me no later than 9am on the last day of class*.

Your presentation must include the following information:

- What you investigated

- Your experimental results

Grading

- Project proposal - Meets specifications above.

- Status reports - Meets specifications above. How much work was accomplished during the time period? This is your work for the next 2.5 weeks and I expect you to be putting in regular time on the project. **Don't procrastinate!**.

- Project and paper

    - The scope/difficulty of your project.
    - How creative is your project/experiment?
    - How complete is your project? Did you accomplish what you set out to do?

- Paper meets specifications above
  - The quality of your write-up

- Presentation

  - Covered content
  - Organized and well-prepared
  - Presentation style

Project Ideas

- Implement one of the approaches/techniques we mentioned in class, but didn't do as an assignment

  - decision trees with other pruning/stopping criteria
  - feature selection
  - detecting outliers
  - dealing with imbalanced data
  - modifying classifiers to support weighted examples
  - ranking
  - linear regression
  - logistic regression (classifier)
  - ensemble approaches (boosting, bagging, ...)
  - random forests

- Dig in more into Hadoop or the MapReduce framework.

  - Look at other machine learning techniques that could be implemented (e.g. ensemble approaches or multi-class classification)
  - Do an in depth timing experiment on one or more techniques and optimize our existing algorithms.
  - Data analysis

- Do a more in-depth experiment of something we've covered already

  - Look at the performance of our current classifiers on a broad range of data sets
  - Look at the impact of feature selection over our classifiers

- Explore some problem/issue/technique that we haven't dealt with in class (e.g what to do with data with missing feature values, collaborative filtering).

- Try and tackle a problem data set and see how well you do. For example `www.kaggle.com` has some interesting data sets. If you go this route I expect more than just "I ran this classifier on this data set and got X accuracy."