HMC CS 158
Final Project

1 Project Overview

One of CS 158's goals is to prepare you to apply state-of-the-art machine learning algorithms to an application. The class's final project will offer you an opportunity to do exactly this.

Projects can be done in pairs or solo. (I strongly encourage working with a partner.)

2 Project Topics

Your first task is to pick a project topic. If you are looking for project ideas, please come to office hours, and I would be happy to brainstorm and suggest some project ideas. I encourage you to be creative, and feel free to ask me if your idea is appropriate. In the meantime, here are some suggestions.

Most students do one of the following kinds of projects:

1. **Application project**. Pick an application that interests you, and explore how best to apply learning algorithms to solve it. If you choose this option, your project *must* include a solid experimental examination. (This is by far the most common, and I highly recommend you choose it for this course.)
2. **Algorithmic project**. Pick a problem or family of problems, and develop a new learning algorithm, or a novel variant of an existing algorithm, to solve it.
3. **Theoretical project**. Prove some interesting/non-trivial properties of a new or an existing learning algorithm. (This is often quite difficult, and so very few, if any, projects will try to do this. I highly recommend against this option for this course.)
4. **Teaching project**. Pick a machine learning algorithm that interests you but that will not be covered in class. Motivate the algorithm and teach the fundamentals to your classmates. If you choose this option, you *must* also apply the algorithm to at least one problem domain. (The difference between the application and teaching project is that the former tends to compare different algorithms while the latter focuses on a specific algorithm.)

Some projects will also combine elements of applications and algorithms and theory.

The only caveat regarding your project topic is that **your project must be specific to this course**. That is, you are not allowed to use your clinic project or a project from another course. This being said, if you have an existing project (past or current) that would benefit from ML *and* you can identify a concrete component developed specifically for this course, talk with me to make sure there is no "double-dipping".

Many fantastic class projects come from students picking either an application that they are interested in, or picking some sub-field of machine learning that they want to explore more, and working on that as their project.

For inspiration, you might look at some machine learning data sets and competitions:

This document is based on the final project handout from Andrew Ng's CS 229 course at Stanford University and David Kauchak's CS 451 course at Middlebury College.

- Kaggle
- UC Irvine Machine Learning Repository
- Yahoo Webscope

Please organize groups under Canvas Project Groups. If multiple groups are interested in the same data set or competition, we might consider allowing larger teams.

Projects will be evaluated based on:[1]

- **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 the project scope and difficulty "worthy" of a course project?)
- **Novelty**. (How creative is the project/experiment?)
- **Clarity** of project deliverables. (Do authors provide sufficient detail about their project? Is the deliverable organized and well-prepared?)

Regarding any technical work, you may code in whatever language you would like and may (and are encouraged to) use any external resources you would like, including both code and data.

Lastly, a few words of advice: Many of the best class projects come from students working on topics that they are excited about. Remember, this is your main assignment for the final weeks of the course, and I expect you to put in regular time on the project. (**Do not procrastinate!**) So, **pick something that you can get excited and passionate about!** Be brave rather than timid, and do feel free to propose ambitious things that you are excited about. Finally, if you are not sure what would or would not make a good project, please also feel strongly encouraged to either email staff or come to office hours to talk about project ideas.

3 Project Logistics[2]

Refer to the course calendar for due dates. **Submit all deliverables in Canvas in PDF or PowerPoint format** (or the Mac-equivalent). (In particular, MS-Word, OpenOffice, PostScript, or any other document format will not be accepted.)

3.1 Project Proposal Presentation [10 points]

Submission instructions: If you use a slide deck, submit online as `LastNames-ProjectTitle-ProposalSlides.pdf` (or `.ppt`). For example, if the project partners are John Doe and Jane Smith, and your project title is "Learning to Recognize People," then name your PDF file `DoeSmith-LearningToRecognizePeople-ProposalSlides.pdf`.

Each team will have 5 minutes in-class to present the following information (so make sure you have thought about each of these things):

[1] Don not overthink these criteria, nor worry too much if you are not sure that you can do well on all of them. Just think of this as an "ideal" that you should aspire to. For example, I expect more than just "I ran this classifier on this data set and got X accuracy".

[2] While I include a rough point breakdown, these are meant only as reference. I will evaluate the entire project based on the above evaluation metrics when determining a grade.

- team members
- project description
  - What is your data set?
    - Where did it come from?
    - What are the features, and what is being predicted?
    - What is the dimensionality of your data set? How many examples, and how many features?
  - What initial questions do you have about the data set?
  - What is the desired input / output to your predictive model?
- project plan
  - Will you conduct a literature review?
  - If you are not doing a theoretical project, provide an outline of your experimental evaluation.

Your goal here is to convince the staff that you have taken the time to understand the data set, where it came from, and potential issues involved. We will then have a few minutes of class discussion to help you finalize any ideas.

3.2 Project Proposal Write-up [10 points]

Submissions: Submit online as `LastNames-ProjectTitle-Proposal.pdf`.

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

- **Header**: Provide the title of the project and the full names of all your team members.
- **Summary**: A one paragraph (300-500 word) 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 cannot 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 are having trouble finding appropriate data.

3.3 Status Update [10 points]

Submissions: Submit online as `LastNames-ProjectTitle-Update.pdf`.

Your status report should describe what you have accomplished so far and what else you plan to do. You should use this report to help you make sure you are on-track and to provide staff with an update on your current progress.

For example, if you are doing an application project, then by this point, you should have tried some initial visualizations, defined the features / subsets of your data, tried out at least one machine learning algorithm, and visualized and summarized the results.

An example report is provided at the end of this document. This is not meant to take you a long time, but please do spend a little bit of effort putting this together. Please keep in mind that the intended audience is the course staff.

3.4 Project Presentation [30 points]

Submissions: If you use a slide deck, submit online as `LastNames-ProjectTitle-Slides.pdf` (or `.ppt`).

Each team will present their work to the class. Provide some motivation for your project, discuss your approach and results, and give insight into anything interesting that you find.

The time limit for this presentation will be set after I see how many projects we end up with. I would estimate 20-30 minutes per presentation for now. While I recommend using a slide deck, you are welcome to use whatever presentation style you think is most suitable for your project.

3.5 Project Write-up [40 points]

Submissions: Submit online as `LastNames-ProjectTitle-Writeup.pdf`.

This is a professional document. Consider it as a report that might be passed off to another team. Someone unfamiliar with your project should be able to understand your problem and achievements and continue or extend your work.

Project writeups can be at most **5 pages** long (including appendices, figures, references, and everything else you choose to submit). Include a document header with pertinent project information, and structure the document for readability. (**Quality matters!**) If you include any diagrams, make sure they are captioned and properly referenced in the main text.

If you did this work in collaboration with someone else, or if someone else (such as another professor) advised you on this work, your writeup must fully acknowledge their contributions.

As an example of information that you should consider including, if you are doing an application project, you might have the following sections:

- **Introduction**: Briefly the technique/problem/application that you investigated.
- **Methods**: 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?
- **Results**: Concisely state your findings, including any supporting tables, graphs, and figures.
- **Conclusions**: Summarize your findings. What did you “learn” about the dataset?

4 Miscellaneous

We may make final presentations or write-ups available online so that you can read about each others' work (and so future classes can read about your projects). If you prefer to opt out, then please contact the instructor.

The Writing Center

The Writing Center provides a welcoming space for writers to get feedback on their composition projects, whether written, spoken or visual pieces. Writing Center Consultants are prepared to assist students in any discipline with any stage of the writing process, from developing an idea to polishing a final draft. Even the most accomplished writers benefit from seeking feedback at the writing center. The center is open Sunday through Thursday evenings from 7-11 and Saturday and Sunday afternoons from 3-5. It is located in Shanahan 1470, just up the walkway from the cafe. You may schedule an appointment through their website, or you may simply drop in during normal hours. If you'd like an appointment outside of normal hours, you may email with your request.

YOU WILL LIKELY FIND YOUR WRITING CENTER VISIT MORE VALUABLE IF YOU GO EARLIER THAN THE NIGHT BEFORE SUBMISSION.

Example Status Report

<table>
<tr><td><b>Members</b></td><td>Dave Deriso, Bryan McCann, Vincent van Gogh</td></tr>
<tr><td><b>Title</b></td><td>Classification of Van Gogh's Irises</td></tr>
<tr><td><b>Hours</b></td><td>Dave (6 hrs), Bryan (7 hrs), Vincent (6.5 hrs).</td></tr>
<tr><td><b>Predicting</b></td><td>We aim to classify the kind of irises in Van Gogh's garden</td></tr>
<tr><td><b>Data</b></td><td>Data are publicly available here (1 .csv file). There are (150) rows, each representing a unique example flower. There are (5) columns: Sepal Length (continuous), Sepal Width (continuous), Petal Length (continuous), Petal Width (continuous), and Species (class). Species is labeled and represent our ground-truth data with which we will train our model.</td></tr>
<tr><td><b>Features</b></td><td>We have a 10 dimensional feature space consisting of: [the first (4) columns of the raw input data, 3 features based on PCA, 3 features based on gaussian kernels]. Our literature review has indicated that a lower dimensional representation of these data will improve the model more than the raw inputs alone (see report). An L1 shrinkage parameter is employed in the regression to discern whether or not a feature is informative.</td></tr>
<tr><td><b>Models</b></td><td>We have continuous inputs and have a categorical output which go well with SVM, LDA, multinomial logistic regression, and naive bayes classifiers. We have implemented our models in Python.</td></tr>
<tr><td><b>Results</b></td><td><i>Training set = 75 random samples, Test Set = remaining 75 samples</i>
<table>
<tr><th>Model</th><th>Training MSE</th><th>Test MSE</th><th>N Iterations until Convergence</th></tr>
<tr><td>SVM</td><td>0.9432</td><td>0.7563</td><td>53</td></tr>
<tr><td>LR</td><td>...</td><td>...</td><td>...</td></tr>
</table></td></tr>
<tr><td><b>Problems</b></td><td>The baseline algorithm achieves test performance of 0.80, and we are having issues brainstorming how to improve our test performance.</td></tr>
<tr><td><b>Future</b></td><td>We are planning to continue adding new features, and we may even try some more advanced algorithms. We'd like to tap into some of the deep learning/neural network approaches the professor mentioned, but we're not sure that we have enough data. If this doesn't go well, we are going to focus on improving existing models by training hyperparameters (such as the shrinkage term).</td></tr>
<tr><td><b>Specific Questions</b></td><td>We have no specific questions, but will come to office hours for more help. If you must give us feedback, we'd like to know what other methods we should look at, but at this time there are no burning questions.</td></tr>
</table>

Details About Each Section

Hours	The number of hours each person put into the project since the last checkpoint.
Predicting	What are you building? Specifically, what are the inputs and outputs. Don't explain the motivation for your topic. **(1-2 sentences max)**
Data	Exactly where did your data come from and what does your contain? (i.e. What are in the rows and columns? Are examples labeled with ground truth? If you have images, are they color, normalized, etc?) **(2-3 sentences max**)
Features	How many features do you have and which features are the raw input data (ex. color, weight, location, etc) vs. features you have derived (ex. PCA, Gaussian Kernel)? Why they are appropriate for this task? **(3-4 sentences max)**
Models	Exactly which model(s) are you using? Why they are appropriate for this task? Why have you chosen to compare/contrast each model? **(3-4 sentences max)**
Results	Make a compact table of results. Each row should be a different model. The columns should be the training error and the test error. List how many samples are in each of the training and testing data sets. Obviously, these sets should be different. **(1-2 sentences max + 1 table max)**
Problems	Any problems/issues that have arisen that might keep you from finishing your project. **(3-4 sentences max)**
Future	List plans for the future. **(2-3 sentences max)**
Specific Questions	In brief, if we could provide specific feedback about some aspect of your project, what could it be? Please make this succinct. **(1-2 sentences max)**