Skip to content

CSE 446 Machine Learning

Emily Fox

University of Washington

MWF 9:30-10:20, THO 101

Home | Lectures | Homework | Exams | Project | People | Discussion Boards New Old

Important Dates

- Mon., Feb 6 at 9:30am: Project Proposals
- Fri., Feb 24 at 9:30am: Project Milestone
- Fri., Mar 10, time/location TBD: Poster Session
- Mon., Mar 13 at 9:30am: Project Report

Your Course Project

Your class project is an opportunity for you to explore an interesting machine learning problem in the context of a real-world data set. We are providing some seed project ideas below. You can pick one of these ideas, and explore the data and algorithms within and beyond what we suggest. You can also use your own data/ideas, but, in this case, you have to make sure you have the data available now and a nice roadmap, since a quarter is too short to explore a brand new concept.

Projects can be done by you as an individual, or in teams of two students. You can discuss your ideas and approach with the instructors, but of course the final responsibility to define and execute an interesting piece of work is yours.

The final project is worth 20% of your grade, which will be split amongst three deliverables:

- A project milestone (20% of the grade), due on February 24th at 9:30am.
- A project poster presentation (20% of the grade), on March 10th, time and place TBD.
- A final report (60% of the gradue), due on March 13th at 9:30am.

Your project will be evaluated by three criteria:

- Technical Depth: How technically challenging was what you did?
- Scope: How broad was your project? How many aspects, angles, variations did you explore?
- Presentation: How well did you explain what you did, your results, and interpret the outcomes? Did you use the good graphs and visualizations? How clear was the writing?

The Technical Depth and Scope are complementary criteria, e.g., if you develop a single elaborate algorithm or model on a small dataset, you may score high on depth but low on scope, while if you try many very simple methods on different datasets, your scope would be higher but the depth lower.

Project Proposal

You must turn in a project proposal on **Monday February 6th at 9:30am through Catalyst**.

Read the list of available data sets and potential project ideas below. If you prefer to use a different data set, we will consider your proposal, but you must have access to this data already, and present a clear proposal for what you would do with it.

Project proposal format: Proposals should be one page maximum. Include the following information:

- Project title
- Data set
- Project idea. This should be approximately two paragraphs.
- Software you will need to write.
- Papers to read. Include 1-3 relevant papers. If you are doing something different then one of the suggested projects, you will probably want to read at least one of them before submitting your proposal.
- Teammate: will you have a teammate? If so, whom? Maximum team size is two students. One proposal per team.
- Milestone: What will you complete by the milestone? Experimental results of some kind are expected here.

Project Milestone

A project milestone should be submitted on **February 24th at 9:30am via Catalyst**. Your write up should be 3 pages maximum in NIPS format, not including references (the templates are for LaTex, if you want to use other editors/options please try to get close to the same format). You should describe the results of your first experiments here. Note that, as with any conference, the page limits are strict! Papers over the limit will not be considered.

Poster Session

We will hold a poster session on **March 10th** from 2:00-4:00pm in the Atrium of the Paul Allen Center. Each team will be given a stand to present a poster summarizing the project motivation, methodology, and results. The poster session will give you a chance to show off the hard work you put into your project, and to learn about the projects of your peers.

Here are some details on the poster format:

- We will provide poster boards that are 32x40.
- Suggested ways to make your poster:
  - Create a bunch of presentation slides (using powerpoint, beamer, etc), and print out each side on a piece of letter-sized paper. Then, put them all together on the provided poster board.
  - If you have access to a poster printer, you are welcome to create a single huge slide and print it out on a poster printer. However, you are not expected to have access to a poster printer, and you will not receive extra "presentation points" if you use this method.

Project Report

Your final submission will be a project report on **March 13th at 9:30am via Catalyst**. Your write up should be 8 pages maximum in NIPS format, not including references (the templates are for LaTex, if you want to use other editors/options please try to get close to the same format). You should describe the task you solved, your approach, the algorithms, the results, and the conclusions of your analysis. Note that, as with any conference, the page limits are strict! Papers over the limit will not be considered.

Project Ideas

The course staff has outlined several potential project ideas below. This should give you a sense of the datasets available and an appropriate scope for your project. You can either pick one of these or come up with something of your own to work on.

Netflix Challenge
From 2006-2009, Netflix sponsored a competition to improve its movie recommendation system. Their system is based off of predicting what rating a user will give to a particular movie (using a 1-5 star system). In effect, what we have is a matrix where each row represents a user and each column represents a movie. Some elements are filled with past ratings, but most of them are unknown. Students can use matrix factorization or clustering methods to predict the missing values in this matrix.

- **Data:** Ratings (Lines are of the form "userid movieid rating")

- **Task:** Predict users' ratings of movies they haven't seen.
- **Background:** netflixprize.com Wikipedia article
- **Methods:** matrix factorization/SVD, collaborative filtering, clustering, side-information

fMRI Brain Imaging
Brain scans were taken of a subject in the process of a word reading task. We want to be able to predict what word the participant is reading based off of the activation patterns in their brain. To do this, we have 218 semantic features for each word in our dictionary (where each feature is a rating from 1-5 answering a question such as "Is it an animal?"). Thus, we can use the fMRI image to predict the semantic features of the word, and then use our dictionary to find our best guess as to which word it is. In this way, we can predict words without ever having seen them in our training set.

- **Data:** fmri.zip (See 3.3.3 of this homework from last quarter's Big Data class for a description of the dataset)
- **Task:** Given an image and two candidate words, predict which of those words was being read by the subject.
- **Background:** CMU Background
- **Methods:** LASSO, dimensionality reduction, stochastic coordinate descent

Document Clustering
Documents taken from sources like Wikipedia or the online discussion board Usenet often have word choice that reflects the topic being dicussed -- for example, the word "clustering" is much more likely to show up in a document on Computer Science than one about motorcycles. Given the collection of words in a document, it should be possible to predict what subject it is contained in. Alternatively, we might want to try to find documents similar to the one in question: if someone is reading a Wikipedia article on classification, perhaps they would also like to know that there is an article on regression.

- **Data:** Usenet: Original Useful Subsets , Wikipedia: Wikipedia dataset
- **Task:** Search for articles similar to an example, or divide the dataset into clusters.
- **Background:** KM Chapter 25
- **Methods:** unsupervised learning, K-Means, LDA, spectral clustering

Digit Recognition
Implement handwriting recognition by classifying pictures (stored as pixel data) as the appropriate digit. This project is based off a tutorial ML competition hosted on kaggle.com

- **Data:** Kaggle Dataset
- **Task:** Classify an image of a digit
- **Background:** Kaggle Description MNIST database
- **Methods:** multinomial logistic regression, k-nearest neighbor, svm, PCA, cross-validation.

Federalist Papers
This task involves the famous disputed federalist papers. Some of the papers we know who wrote them and others we do not. The task involves converting the text into a "bag of words" (see attached for more info) feature vector and then attempting to classify the remaining essays as Hamilton or Madison. Students can use cross validation to test predictive power on known essays. This project involves mining text, which may be a challenging component, but it shows a very cool connection between machine learning and history.

- **Data:** Project Gutenberg
- **Task:** Classify disputed papers as Hamilton or Madison
- **Background:** Background
- **Methods:** logistic regression, decision trees, svm, PCA, cross-validation

Job Salary Prediction
This is another task taken from a Kaggle competition. Given an advertisement for a job opening, the goal is to predict the starting salary for the job being posted. Much of the data about the ads is unstructured text (like the ad content itself), but some structured data is given as well. A tree of the geographic relationships between the job locations is also provided. This task is similar to the running example in lecture of predicting starting salary, and has real-world usefulness to the company that posted the problem.

- **Data:** Kaggle Dataset
- **Task:** Predict a salary from a job posting
- **Background:** Kaggle Description
- **Methods:** regression, bag-of-words, dimensionality reduction, neural networks

Eigenfaces
The goal of this task is to learn how to recognize faces. We have a set of pictures of 20 people in various directions and expressions, some of which have sunglasses. One major problem with image data is that our input features are individual pixels, which are high-dimensional but not terribly meaningful in isolation. Using PCA, we can decompose our images into eigenvectors, which are linear combinations of pixels (nicknamed "eigenfaces"). Students can explore different classification tasks, from determining the presence of sunglasses to identifying individuals.

- **Data:** Faces Directory
- **Task:** Classify images of faces
- **Background:** .PGM format specification
- **Methods:** dimensionality reduction, PCA, SVM, neural networks