MTH/CSE 4224 Intro to Machine Learning: Project 1

Proposal Deadline: April 10
Project Report Deadline: April 21
Presentations: April 22-26

Project Description

Pursue a machine learning project including use of unsupervised learning. You will propose a problem and submit a proposal outlining the goals of your study, intended data sources, and initial plans for cleaning/preprocessing and models you will use.

You may work in teams or individually. Project expectations will be proportional to the number of team members.

Proposal

For each project, you must submit a proposal (max 1 page) including the following:

1. The goals of your study.
2. Your intended data source(s) and initial plans for cleaning and/or preprocessing the data.
3. Initial plans for methods you will use. (You should test multiple methods.)
4. Your team: If you want to work in a team, list who will be on the team, where each person plans to focus in the project.

Note: Methods we have not covered in class are permitted, but you should use some methods we used for benchmarking at least.

Note: If you propose to form a team, you may not break up after proposal approval.

Proposal Grading

The proposal is worth 15 points. The grading criteria are:

- 5 points: Goals of the work are described in detail.
- 5 points: Appropriate data has been located for use or you have a plan to gather the data.
- 5 points: There is a reasonable plan for cleaning/preprocessing the data and methods to use.

The purpose of the proposal is to design a feasible study at an appropriate level. I will evaluate them and give feedback to ensure the level of difficulty is appropriate given the team size.

Project

Each project is different and different projects will require more work on some parts than others, but all should have four main parts **(must include some unsupervised learning)**:

1. **Data preparation:** Create or locate the dataset and preprocess the data.
2. **Benchmarking:** Initial results from basic, un-tuned machine learning methods.
3. **Training and tuning:** Train models and tune the hyperparameters to solve your problem as well as possible.
4. **Communicating the work:** Write a report thoroughly describing your work including diagrams, decisions you made along the way, progress from the benchmarks, and your general findings, and do a presentation of your work with the professor.

Project Grading

The project is worth 150 points. (Students within teams will receive the same grade.)

- 15 points: Proposal
- 15 points: Problem description in report (What are you trying to do? What is your data?)
- 20-40 points: Data preparation
- 50-70 points: Training and tuning your models
- 15 points: Conclusions, descriptive diagrams, etc.
- 15 points: Present the work

The amount of points dedicated to some categories depends on the nature of the project.

Data

These are just some options for data sources:

- Google Dataset Search
- UCI Machine Learning Repository
- Kaggle
- Spotify API
- `https://www.interviewqs.com/blog/free_online_data_sets`

You can turn most data into vectors, so you can study audio, images, readings from sensors, large databases, text, video, etc.

You need to keep the dimension of the data under control for computational constraints, but a few thousand is fine. (You could degrade or crop high res images, take small snippets of audio, etc.)

Guidelines: Data with at least 1000 data points with at least 20 dimensions tend to work best.