CMP712 Machine Learning
Project Topics

- In your project, you should explore a machine learning problem using a real-world data set which you should obtain before your project proposal. In a typical project, you should select a real-world data set and apply appropriate machine learning techniques for a task using selected data set. If you can find new approaches for your task in order to improve the performance, it will be a good project.

- You may choose your project proposal from the list given below or you may suggest any other project in Machine Learning field. In addition to following project list, you can find a lot of different machine learning project ideas on the web. In your project, you should apply machine learning methods to a problem domain and your project should include a computational work.

- A single student or two students together can select a project topic. If two students work on the same project together, they should clearly indicate their contributions in a separate page attached to their final project report. Projects that are done by a group should be more substantial than projects done individually.

- You should pick your project topic as soon as possible. You should write one page document for your project proposal, and submit ***your project proposal*** (a pdf file). ***You should obtain or create your data set for your project before your project proposal.*** Your project proposal should include the followings:
    - Project title and the student(s) in the project team.
    - The data set
    - Project idea
    - The computational work that will be done.
    - The relevant papers with your project proposal. You should read at least one of them before your project proposal.

- You should find at least two-three relevant major papers in your project topic and read them.

- At the middle of the semester, you will submit your ***midway project report***. This means that you should finish some of your project work before the midway point. The midway project report should be in the format of the ***final project report*** and its length should be 3-5 pages. Your midway project report should include the related work section and describe the machine learning techniques that you tried upto the midway point and their results. The midway project report is basically a short version of your final project report.

- At the end of the semester, you will submit your ***final project***. You should submit the pdf file of your final report
    - Prepare your final project report in the format of a conference article using IEEE double column format (6-8 pages).
    - In your final project report, you should use your own words. *Do not cut and paste from the papers that you read.*
    - Your final project report should contain at least the following sections in addition to a *title* and an *abstract*:

      Introduction – Describing the problem that you are attacking

Related Work – Describe the related works here together with their relations with your work.
Sections describing your computational work in detail – Describe the details of your computational work in these sections. If your computational work needs evaluation, do not forget to include evaluation sections.
Conclusion – Give your concluding remarks and possible future works in this section.
References – Give the references that are cited in your paper.

- With your final project report, you should send an electronic copy of each of the ***major papers that you read in your survey***.

- You should send your ***source files of your project*** at the end of the semester.

- You should make a ***demo of your project*** to me at the end of semester.

Machine Learning Data Sets:

- UC Irvine has a repository where you can find a data set for your project. You can find a data set for your project from the following locations or other machine learning sites.
  - http://www.ics.uci.edu/~mlearn/MLRepository.html
  - http://www.cs.toronto.edu/~roweis/data.html
  - http://www.cs.pitt.edu/mpqa/
  - https://pub.towardsai.net/best-datasets-for-machine-learning-data-science-computer-vision-nlp-ai-c9541058cf4f
  - https://towardsdatascience.com/top-sources-for-machine-learning-datasets-bb6d0dc3378b
  - https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research

Some Possible Project Topics:

- You may select a project from the following list or you can suggest a project related with Machine Learning.

Text Categorization

Each written document can be categorized according to its content. For example, the category of a newspaper article can be econmy, sport, etc. A text categorization system determines the category of a given document.

Author Identification

Determining the author of a given text.

Opinion Mining

Deciding the polarity (positive or negative) of views in customer review texts or texts in social media environments. This can be done for Turkish or English texts

Extraction of protein names in biomedical texts

Extraction of protein names in biomedical texts.

Extraction of protein interactions in biomedical texts

Extraction of relations between protein names in biomedical texts.

Weakly-Supervised Learning

Weakly-Supervised Learning for Relation Extraction

Optical Character Recogntion

Recommendation System

Object Recognition in Images

Speech Recognition

Weather Prediction

Some Possible Project Topics from CMU Machine Learning Course:

Using Vector Space Models for Natural Language Processing

Using NP-context co-occurrence matrix, finding synonyms, finding members of categories (i.e., "is this noun phrase an athlete?"), or clustering noun phrases to automatically induce categories.

- http://www.cs.cmu.edu/~tom/10709_fall09/RTWdata.html
- http://qwone.com/~jason/20Newsgroups/

Peter D. Turney and Patrick Pantel (2010). From Frequency to Meaning: Vector Space Models of Semantics. Journal of Artificial Intelligence Research 37, pp. 141-188.

Image Segmentation Dataset

The goal is to segment images in a meaningful way.

- https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/

Character recognition (digits) data

The goal is optical character recognition and the simpler digit recognition.

- http://ai.stanford.edu/~btaskar/ocr/

Precipitation data

- This dataset includes 45 years of daily precipitation data from US:
  - http://research.jisao.washington.edu/data_sets/widmann/

Weather prediction: Learn a probabilistic model to predict rain levels.

WebKB

This data set contains webpages from 4 universities, labeled with whether they are professor, student, project, or other pages.
Project idea: Learning classifiers to predict the type of webpage from the text.

- http://www.cs.cmu.edu/~webkb/

Email Annotation

The goal is to identify which parts of the email refer to a person name. This task is an example of the general problem area of Information Extraction. Model the task as a Sequential Labeling problem, where each email is a sequence of tokens, and each token can have either a label of "person-name" or "not-a-person-name".

- http://www.cs.cmu.edu/~einat/datasets.html
- Paper: http://www.cs.cmu.edu/~einat/email.pdf

Netflix Prize Dataset

The Netflix Prize data set gives 100 million records of the form "user X rated movie Y a 4.0 on 2/12/05".

- https://www.netflixprize.com/

Project idea:

- Can you predict the rating a user will give on a movie from the movies that user has rated in the past, as well as the ratings similar users have given similar movies?
- Can you discover clusters of similar movies or users?

Object Recognition

The Caltech 256 da taset contains images of 256 object categories taken at varying orientations, varying lighting conditions, and with different backgrounds.

- http://www.vision.caltech.edu/Image_Datasets/Caltech256/

Project idea:

- You can try to create an object recognition system which can identify which object category is the best match for a given test image.
- Apply clustering to learn object categories without sup ervision.

Enron E-mail Dataset

The Enron E-mail data set contains about 500,000 e-mails from about 150 users.

- http://www.cs.cmu.edu/~enron/

Project idea: Can you classify the text of an e-mail message to decide who sent it?