DS 4400: Machine Learning and Data Mining I

Fall 2018

Final Project Resources

Public datasets and resources

1. The UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets.html
   Examples:
   - Census Income
   - Spambase
   - Credit
2. Image datasets
   - CIFAR-10: https://www.cs.toronto.edu/~kriz/cifar.html
   - Face recognition: http://www.face-rec.org/databases/
   - German Traffic Sign Dataset: http://benchmark.ini.rub.de/?section=gtsrb&subsection=dataset
3. OpenML: https://www.openml.org/
4. Kaggle competitions: https://www.kaggle.com/
5. Text datasets
   - Enron email data: http://www.cs.cmu.edu/~enron/
6. Healthcare datasets: https://healthcare.ai/broadcast/open-healthcare-datasets/
7. Security datasets:
   - Microsoft malware classification dataset: https://www.kaggle.com/c/malware-classification
   - TREC Public Spam Corpus 2005: https://plg.uwaterloo.ca/~gvcormac/treccorpus/
8. Government datasets: https://catalog.data.gov/dataset
9. Collection of various datasets: https://github.com/awesomedata/awesome-public-datasets

Recommendations

- Select first a problem of interest and express it as an ML task (classification or regression).
- Perform evaluation on at least one large dataset (> 10,000 examples).
- Try different ML algorithms for your problem and compare the results.

Project ideas

1. Design a face recognition system that will be trained with images of multiple subjects. The problem can be modeled as multi-class classification. Evaluate some simple classifier (e.g., logistic regression), as well as more complex ones (e.g., feed-forward neural network and convolutional network).
2. Can you predict who wrote an email using the Enron email dataset? You can try 3 different classification models.
3. Predict sentiment analysis on movie reviews, using a Kaggle dataset: https://www.kaggle.com/c/sentiment-analysis-on-movie-reviews/data
4. Predict if a patient will show up at his appointment using a Kaggle dataset: https://www.kaggle.com/joniarroba/noshowappointments/home
5. Predict earnings and debt of college graduates using the College Scorecard Data: https://collegescorecard.ed.gov/data/
6. Build a classifier for road signs using a subset of images extracted from the German Traffic Sign Dataset.
7. Build a spam email classifier using the TREC Public Spam Corpus dataset.
8. Build a malware family classifier using the Microsoft malware classification dataset.

Project proposal

It should be one page following the template:

- Project Title
- Problem Description
    - What is the machine learning problem you are trying to solve?
    - You can propose your own topic, the ideas above are just examples.
- Dataset
    - Link to data, brief description, number of records, feature dimensionality
- Approach
    - Normalization if any
    - Feature selection if any
    - Machine learning models you will try for your problem
    - Methodology for splitting into training and testing, cross validation
    - Language and packages you plan to use
    - Metrics, how you will evaluate your models
-