Carnegie Mellon University
Department of Electrical and Computer Engineering

18-752 Estimation, Detection and Learning Spring 2023

Project ideas

1 Data collection

Raw data can be obtained from various agencies, as you can find with a simple search. Do not use datasets that have already been prepared for Machine learning, such as those in the UCI repository or Kaggle datasets that already have features extracted. Some examples of where you can find datasets are:

1. Kaggle `https://www.kaggle.com/`, but only raw datasets. i.e., those that have not been pre-processed and/or have had features extracted. Several of the projects listed below have relevant raw datasets available in Kaggle.
2. Western Pennsylvania Data portal WPRDC
3. Open Data New York
4. Social feed such as Twitter, etc.
5. National Oceanic and Atmospheric Administration (NOAA)

Data can also be obtained by doing your own experiments.

1. Microphone to record voice, music
2. Camera to record images or video
3. Smartphones have lots of sensors such as magnetometer, accelerometer, gyroscope, light sensor, GPS, etc. Apps are typically available that can record these sensor values over extended periods of time. You can obtain the data from the phone and process it offline in a computer.
4. Downloading online books or newspaper articles for text-based inference project .
5. Using any laboratory equipment you may have access to, if that is allowed for class work.

2 Projects

The proposed project must be reasonably interesting as an application, and also, reasonably complex. Some examples are shown below. You may choose one of these, or you can propose your own project. In the latter case, the instructor may ask you to modify your project if it seems unreasonably simple (or too complex).

Weather prediction

You can build your own weather forecast engine. Use multi-channel data (that have at least 5 attributes such as temperature, etc.) obtained from NOAA. You can set up the problem as time series prediction of the future. Alternatively, you can set up the problem as predicting the weather in a city using information in other cities.

Movement inference

Use accelerometer and/or gyroscope sensor data from your smartphone to recognize different types of movements or gestures.

Localization using non-GPS sensors

Use your smartphone's accelerometer and/or compass sensor data to track your location when you move. You can use the GPS periodically to correct for errors that may accumulate.

Background cancellation

Record a video with a webcam that is at least 30 seconds long, cancel the true background and replace it with an artificial background (such as a beach). In addition to the Normalized MSE, the background-substituted video should look authentic.

Target tracking

Train the computer to track the football in 10 sample videos of NFL gameplays. You are allowed to manually mark the location of the football in the first video frame.

Speech activity recognition

Record a dialog between you and your partner (such as a theatrical play) and train the computer to identify the time intervals where each of you is talking, and also intervals where neither or both of your are talking. The entire interval where one person (or both or none) is talking must be correctly identified, even if it based on feature vectors corresponding to 25 millisecond audio segments.

Handwritten character recognition

Train the computer to recognize your handwritten letter characters. This project will require reading literature on feature extraction that is typically used for handwritten character recognition.

For this project, you *must* use a camera to collect samples of your handwriting. If you use additional handwritten datasets that others have collected (such as from Kaggle), the samples must be images only, and further, the Testing data to measure performance must only be samples of *your* handwritten characters.

Optical Character Recognition (OCR) of scanned pages

Scan pages of a book and design software that can fix scanning error (such as rotated pages) and then recognize the characters in the scan. This project will require reading literature on feature extraction that is typically used for optical character recognition.

If you use additional OCR datasets that others have collected (such as from Kaggle), the samples must be images only, and further, the Testing data to measure performance must only be samples from *your* scanned data.

Face detection in images

Train the computer to locate faces in a Test set of at least 20 images. You will need a substantial number of Training images to design the inference scheme. Also, recollect that any project using images must use at least one Manual feature extraction method. You can try the Manual method discussed in class, or read face detection/recognition publications to get ideas.

Newspaper classification

Obtain newspaper *editorials* from four newspapers and train the computer to recognize which newspaper published that editorial, without relying on the names of the editors.

Tweet classification

Obtain tweets using the Twitter API from several political commentators. Then, train the computer to predict who posted the Test data tweets.

There are python packages available to use the Twitter API. But be very careful to adhere to the various rate limits imposed by Twitter to avoid getting banned by the service.

Music genre recognition

Obtain several recordings of different music pieces and train the computer to classify the genre of a piece (e.g., disco, pop, rock).

Traffic/trajectory prediction

Use GPS trajectory of users/vehicles to predict the future actions/movements, or the future traffic loads in a specific area. You may use the following dataset:

- User trajectory data:
  `https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-u`
- Vehicle trajectory data: `https://dl.acm.org/doi/10.1145/3277868.3277870`

Network attack identification

Use sensor/network measurements to identify if a network attack is happening or not. You may use the following datasets:

- iTrust: `https://itrust.sutd.edu.sg/itrust-labs_datasets/`

Network traffic prediction

Use network measurements (e.g., bandwidth, packet loss rate) to forecast future network conditions, bandwidth, etc. You may use the following datasets:

- FCC MBA: `https://www.fcc.gov/general/measuring-broadband-america`
- CAIDA: `https://www.caida.org/data/`

Cluster workload/failure prediction

Use cluster measurements (e.g., CPU/memory measurements) to predict future resource usage, cluster workload, job failure, etc. You may use the following datasets:

- Google cluster dataset: `https://github.com/google/cluster-data`

- Alibaba cluster dataset: `https://github.com/alibaba/clusterdata`
- Microsoft cluster dataset: `https://github.com/Azure/AzurePublicDataset`