EEE549: Statistical ML: From Theory to Practice

Final Project Details

Term Paper Project

- **Final project: 30% of grade**
- Split into two parts:
  - Midterm: Phase I (5% of total points for Final Project)
  - Finals: Phase II (95% of total points for Final Project)

- Phase I: due **October 31, 2023 (11:59pm)**

- **Phase II: due week of Exam – Monday December 4, 2023 11:30 pm**
  - Make sure your team members pull their weight (put their promised effort)
  - Feel free to message me directly if promised goals cannot be met and if team is struggling to work together
  - Do not make travel plans before finishing finals and submitting

Final Term Paper Project

Final Term Paper project will test your understanding of the entire machine learning pipeline

Same project for ALL teams – what differs is what you bring to the project effort

Final Term Paper Project Details

- **Dataset** – **THREE** datasets (all in the classification setting)
  - Categorical/Tabular Dataset
    - Wisconsin Breast Cancer dataset * **AND** UCI Adult dataset
  - Numerical
    - **Fashion MNIST**

- **ML Models** – consider different hypothesis classes to learn a **variety** of models
  - **Existing libraries can be used to learn models – in other words, you don't need to implement GD or backprop or PCA or clustering**
  - **The key effort is in evaluating a range of models within each type, rigorously cross-validate, and make meaningful comparisons**
  - **Exact hypothesis classes on the next slide**

* There is another breast cancer dataset on the UCI repo – make sure you use the right one!

Final Term Paper Project Details

- **ML Models** (**existing libraries can be used to learn models**)
- Six types of models will be learned
- Each of them requires hyperparameter tuning to find best model for that class.
- Model list:
  1. Logistic Regression
  2. SVM (ideally with kernels)
  3. k-nearest neighbors (k-NN)
  4. Clustering
  5. PCA
  6. Neural Networks
     a) Feedforward NNs (for tabular)
     b) CNN (1-layer and 2-layer) for image datasets

Final Term Paper Project Details

For each dataset and model choice:

- **Algorithms** (have to use each of these, as appropriate)
  - Gradient Descent (where applicable)
  - SGD or mini-batch SGD (where applicable)
  - (bonus): Momentum term included in update rule (where applicable)

- **Hyperparameter Tuning**
  - k-fold cross validation
  - General parameters: Learning rate, number of epochs trained, momentum term
  - Additional model dependent parameters also need to be tuned
  - FNN/CNN parameters need to be tuned as well

Immediate To-do List

- **Form a group**
- **Submit a ONE page proposal which includes:**
  - Team members, full details on graduate program etc.
  - List the categorical and numerical datasets to be used
  - General breakdown of tasks across team members including writing and result collation task

- **Proposal due Oct 31, 2023 11:30 pm on Canvas**
- **Get started on the work immediately**

- Final report rubric and expectations – posted by end of October
  - **Key idea is to meticulously design experiments, write results, and draw meaningful conclusions.**

Immediate To-do List

- **Form a group**
- **Submit a ONE page proposal (+references) which includes:**
  1. Team members, full details on graduate program etc.
  2. List the categorical and numerical datasets to be used
  3. General breakdown of tasks across team members including writing and result collation task
  4. **potential references and libraries that will be used (additional page)**
  5. **Font size cannot be smaller than 11 points**
- **One PDF file containing all four details above should be uploaded via Gradescope. Will be setup as group project**
  - Only one member (lead) will do so.
  - Lead has to make sure all members are indicated in the group as otherwise members not indicated/included will get 0 points (since there will be no project assigned to their names).