Fundamentals of Machine Learning - Group Project
CSE 326/426
Prof. Sihong Xie
Due Date: 11:59pm, Sep 24, 2018

1 Final Project Requirements

This is a group project for you to implement, evaluate and document a published machine learning model/algorithm. The goal is to reproduce the algorithmic performance claimed by the paper. Eech team will prepare a one-page proposal, a one-page milestone report, an 8-page final report along with the softwares and data. All reports are in PDF format prepared using Word/Page/Latex. There is no presentation required.

The project has several due dates: please check the schedule on the course website for the most update-to-date information. Roughly, the proposal is due one month after the class starts (find your teammate(s) before that), the milestone report is due one and a half month later, and the final report and other deliverables are due the week before the final week. All due dates are 11:55pm ET and the submissions go to Coursesite.

2 Choosing a project

The philosophy is to stretch you a bit beyond the fundamentals. All projects should be based on a paper published on a top AI/ML conference/journals (JMLR, JAIR, NIPS, ICML, UAI, ICLR, AAAI, IJCAI, and AIStat). You can reproduce the paper's results, or apply the models/algorithms to an interesting problem you have in hand (let's say you're working on autodriving/robotics/healthcare/financial/NLP projects, you can apply machine learning to them). The paper should be pertaining to and deeper than what we will cover in the course. A good selection can be a mixture of theoretical and empirical treatments. Purely theoretical/empirical papers are also welcome but they have to be deep and/or comprehensive. I listed relevant papers on the course webpage (with brief explanations) as good examples. You can also find papers on the above listed conferences.

Make sure that you can understand the scope of the paper: you should discuss your selection with the instructor about the rationale and your background, and obtain permission from the instructor, to make sure that you don't pick something that is beyond your capability and our computing resources (Sandbox has reasonably powerful GPUs, and we will NOT pay for cloud computing or any other third-party resources beyond Sandbox). For examples of topics,

- Neural networks: you can pick a deep learning paper on image or natural language processing. You can use PyTorch/Tensorflow here.
- Graphical model: you can implement a Gibbs sampler, a variational inference or a message passing algorithm. You can't use machine learning package here.
- PCA: you can explore other dimension reduction techniques, such as autoencoder, LDA, ICA or their Bayesian treatments (we will teach Bayesian methods when covering graphical models). You can use numerical computing package here.
- Gradient descent: you can explore coordinate descent, variance reducation techniques, parallel/distributed/asynchorous optimization on large datasets, or on deep networks.

- Boosting: you can pick a variant of other boosting algorithms or other ensemble methods (e.g., XGBoost).
- EM algorithms: this is a large family of algorithms and many paper applies this idea to applications, such as HMM.
- Reinforcement learning: you can choose to implement a policy gradient, Q-learning, or their marriage (actor-critic). You are encouraged to use environments defined in OpenAI Gym rather than implement your own. Bandit algorithms for recommendation system are also good examples.

3 Writing the proposal

The purpose of the one-page proposal is for your team to reach consensus on a paper and to measure its feasibility. The proposal will consist of the following 5 components.

- Goal: What general machine learning problem the paper tried to address? What's particular issues?
- Gap: Why have the issues not been addressed by previous work at the time the paper was published? What's the technical challenges?
- Resolution: What methodology do the papers propose to address the gap? Do their theory and/or experiments work? Why they work?
- Outcomes: How much the proposed method will work better than the previous methods?
- Resources: can you access relevant datasets (or a close approximation) used by the paper to conduct your experiments. Do you need a multi-GPU workstation, or just a laptop?

For a made-up example, let's say you pick a paper on kernel SVM that extends large margin linear classfiers. The goal is to improve classification accuracy, the gap is that linear models can't handle non-linearity, and the resolution is to formulate a kernel SVM that introduces non-linearity through kernel functions. You expect that, if the data are linearly non-separable, a kernel SVM shall work better than a linear model. You can find an abundance of datasets that are linearly non-separable.

At this stage, we don't expect you to have a very deep understanding of the paper, but to expose you to another level of machine learning and to give you a goal, that you should be able to attain as we move towards the end of the course.

A small amout of the project grade will be assigned to the proposal, based on the completeness of the proposal.

4 Milestone Reports

You should submit a one-page report including the following:

- What progress have your team made so far? It should be on your understanding of the theory/model/algorithm with some exploratory experiments on a dataset.
- What are the challenges your team are facing now? Any plan to address that.

This is a great opportunity for your team to communicate with the instructor/TA to discuss any difficulty you have come across and for you to move forward. Make sure that your team will have a clear picture about the next steps after the milestone. If you're making very good progress and would like to earn some extra credit, you may propose further steps, such as comparisons with more baselines.

For example, your team might have derived and implemented a linear SVM in the primal form. At the same time you've collected a couple of small datasets. However, you don't quite understand the math behind the dual form and how to implement that with a kernel function. You ask the instructor and figure out the math.

A small amout of the project grade will be assigned to the milestonr report, based on the progress you have made and how you communicate with the teaching staffs.

5 Final Submission

The final submission is a zip package, including your final software implementation, the datasets to test your codes, a final report (8 pages, single column) describing the algorithmic details (how you preprocess the data, how you design and implement the algorithms, any parameter tuning, advanced data structure sworth mentioning, etc.), experimental results/analysis (do your results support or invalidate the claim of the paper, and why?), any non-trivial improvements you have tried and you feel worth mentioning. Figures and tables are highly encouraged to demonstrate your ideas and results. We also ask you to indicate the contribution of each team member at the end of the report.

5.1 Grading metrics

A large portion of the grade of the project will go here, using the following metrics:

- Correctness: we want your codes to implement the ideas of the proposed paper rather than something else (don't submit a simple logistic regression on text data, while you have proposed to implement a multi-layer perceptron on images).
- Depth of your understanding: do you show insight into the algorithm/model? Does your team isolate the key factors, such as layer of deep network or number of iterations in Gibbs sampling, that lead to performance differences? We can tell from the analysis of your experimental results/analysis, and the development of the algorithm in your codes.
- Difficulty and complexity of the project: depth of the math and effective number of lines of codes contributed by the team.
- Individual contributions: based on your report, we will decide on how much each team member shall earn. This is not a zero-sum game but we try to discourage free-riders. If the workload are roughly balanced among team members, there should not be a problem.