CSE 512: Open-ended project Due Dec. 7

It's finally time for the project! This open-ended project mainly involves you guys doing something interesting and related to machine learning.

The maximum points is 10 pts: 5 points for fulfilling minimum criteria, 3 points for interesting/surprising qualities, and 2 points for good writing.

1 Choice 1: explore methods

1. Pick one interesting dataset. Must have > 100 samples and > 10 features. Describe why you think it is interesting.

   (Possible keywords: data imbalance, unstructured features, sparse features, repetitive / highly correlated features)

2. Describe a task. It could be obvious (misclassification) or non-obvious (cure the patient). Describe why that task is interesting, and propose 3 ways of measuring a successful task. (e.g. for handwriting disambiguation, I may say "classification rate, soft classification rate, margin-based loss function.") Loss functions may also differ based on methods.

3. Explore *at least 3 methods* for attacking this problem. For each method, show something interesting. Dissect it like an anthropologist, and teach us something new. Adding boosting or bagging counts as an additional method.

   (Possible keywords: ill-conditioned training, slow initial convergence, fast initial convergence, generalization error, model complexity.)

4. **Be a consultant.** Now, you are facing either a president, a doctor, a CEO, or some other high-ranking non-domain expert. Explain why your experiment is interesting to them. Report your results, but also interpret them, to respond to the all-important question "why do I care?"

5. **Be a critic.** Explain 3 ways in which your experiment could be improved. (You can imagine you have access to far greater compute resources and data, but explain what you would need.) Give hypotheses as to what you think would happen if you ran those experiments.

2 Choice 2: explore datasets

1. Pick one interesting method: K-NN, naive-Bayes, logistic regression, SVM, Kernel SVM, linear regression, generalized linear regression, decision tree. Boosting/bagged versions also work.

2. Pick 3 *different* datasets. Each dataset should explore something distinctly different than the other two, e.g. highly unbalanced data, highly sparse features, very numerous features, very rich/bad features, features that should really be processed more, etc.

3. (Alternative to previous) Pick one dataset with > 1000 features. Experiment with the dataset 1) raw, and 2) using 2 different representation schemes. (e.g. Matrix factorization, random data hashing, PCA, ICA)

   - Matrix factorization : feature matrix $X = UV^T$ where U, V are computed via gradient descent, minimizing

     $$\underset{U,V}{\text{minimize}} \quad \frac{1}{2}\|X - UV^T\|_F^2$$

   - Random data hashing: perform some random transform that reduces the dimensionality of your data. For example, if $x \in \mathbb{R}^n$, a linear hashing function performs $f(x) = Ux$ where $U \in \mathbb{R}^{n \times r}$ and $r \ll n$, and U is some random matrix.

4. Run your method on these different datasets, and discuss how the differences affect how well the methods work, what generalization looks like, hints that the method is working well, hints that it's not, etc.

5. **Be a consultant.** Now, you are facing either a president, a doctor, a CEO, or some other high-ranking non-domain expert. Explain why your experiment is interesting to them. Report your results, but also interpret them, to respond to the all-important question "why do I care?"

6. **Be a critic.** Explain 3 ways in which your experiment could be improved. (You can imagine you have access to far greater compute resources and data, but explain what you would need.) Give hypotheses as to what you think would happen if you ran those experiments.

3 Deliverables

- A report (no more than 5 pages, including figures, but not including references) detailing your machine learning adventure.
- Your report must include a short abstract, giving the "elevator pitch" of what's going on.

Interestingness You must devote at least one paragraph to describing why your report is interesting. Potential answers may be:

- This dataset is interesting because the features are very sparse / biased, or labels very unbalanced...
- This task is very difficult because the features are not easy to interpret...
- This solver is usually a terrible idea, but when applied to this specific task, we are able to significantly reduce the computational complexity / significantly improve the performance score...
- ...

We will basically give you full marks if you write something correct and not completely trivial. But do put some thought into this, as when you go on to write conference publications and technical reports, it's probably the only paragraph that most of your readers will actually read.

4 Writing tips

- Do not "dump" things, like giant score tables and code outputs / code samples, in the document. The page limit is there to make you think about how to convey your point concisely and effectively.
- Use the abstract to organize your thoughts. What is the main message you wish to convey? Think of it like: the abstract proposes an idea, the report proves the idea is true, with illustrative examples.
- Start by writing your results section. Give all your figures standalone captions. Write your abstract last.
- Your English does not have to be perfect, but you will be judged based on your writing organization skills and effective communication. I am not looking for grammar and spelling, but I am looking for whether I understand your main idea after just one reading.
- Your results do not have to be perfect / publication worthy, but even the worst results can have very interesting analysis. I am judging your analysis.
- Suggested (not required) sectioning:
  - Abstract (required)
  - Introduction: introduce your main results and a summary of why the problem you chose to present is interesting
  - Methods: What did you decide to do?
  - Results: Plots with captions showing your results
  - Discussion: describe again why what you did is interesting, what questions got answered, and what new questions are raised.

5 Resources

Data

- Kaggle: www.kaggle.com/datasets
- UCI: https://archive.ics.uci.edu/ml/datasets.php
- Stanford SNAP: http://snap.stanford.edu/
- NLP datasets: https://github.com/niderhoff/nlp-datasets

If you find a treasure trove, please also post it on Piazza to share with the class!

Writing

- Writing for scientists and engineers: `https://www.sciencedirect.com/book/9780750646369/writing-for-science-and-engineering`
- Scientists Must Write: A Guide to Better Writing for Scientists, Engineers and Students (In particular, chapter 4, which is available via the book preview)
- (more advanced) `https://ece-kamgarpour-2019.sites.olt.ubc.ca/files/2020/08/ScientificCommunication.pdf`
- (more advanced) `https://www.cs.ubc.ca/~schmidtm/Courses/Notes/writing.pdf`
- (my personal favorite, more advanced) On Writing: A Memoir of the Craft, by Stephen King