Ideas for research projects

For each idea, we (i) summarize the idea, (ii) discuss what we have already done, (iii) summarize what remains to be done, (iv) discuss venues and prospect of success.

Streaming PCA

(i) *Idea.* The task of streaming feature selection is as follows: Suppose you have a stream of features and you want to select an optimal subset for some task. When you first see each feature, you are able to either accept or reject it. After you've rejected it you cannot accept it in the future (though you might be able to drop a feature you've already accepted).

There already exist some solutions to this problem: α-investing (Zhou, Foster, Stine, Ungar, "Streamwise feature selection", JMLR 2006), grafting (Perkins, Theiler, "Online feature selection using grafting", ICML 2003), and OSFS (Wu, Yu, Wang, Ding, "Online streaming feature selection", ICML 2010). However, I think all these approaches only select individual features, like forward selection does in classical feature selection.

We propose Streaming PCA. In contrast to the above approaches Streaming PCA also considers affine transformations of subsets of variables. This has the same potential benefits that standard PCA has over forward selection: it is able to find subsets of variables that are informative in some way jointly but are not informative separately.

(ii) *Done.* An algorithm has been designed. There is an analysis of the algorithm in terms of computational complexity and also number of variables. Some experiments have been run.

(iii) *Todo.* The algorithm could possible be improved to allow the user to control the number of variables selected. The analysis from the two pdfs above needs to be combined and could be improved. Need to compare to the three algorithms mentioned above; can copy the setup (datasets, criteria) of Wu et al. for this.

(iv) *Venues.* ICML. I think the prospect of success is high since the paper is close to done already and papers on similar topics and of similar ambition have already been published in ICML.

Temporal SVM

(i) *Idea.* We all know the idea.

(ii) *Done.* We have an algorithm, and we have an implementation (using CVX). As I recall, we tested it on toy datasets and it works.

(iii) *Todo.* We need several things to make this into a good paper.

(a) We need to find a domain and dataset where there there is actually a suitable change over time. Ideally we could find a medical or biological dataset for this. I found yesterday a biological dataset that looked potentially good, but I couldn't really understand the task. And now I've lost it. I think I found it in a paper about dynamic Bayes nets.

(b) We need to justify our approach. Two objections: (1) Why not just use SVM and use the timestamp as another feature. (One answer: You could indeed do this by designing an appropriate kernel function but it would be tricky and hacky, whereas our approach is straightforward.) (2) Linear is too restrictive. (One answer: It's just a first step, and it is still useful even as-is.)

(c) We need to find a few competitors and design an experiment where we will do well by comparison. Of course this task relates to (a) in that our choice of dataset needs to match our experimental design, and it relates to (b) in that when we find competitors this will suggest new objections to our approach.

(d) It would be good to include a few theorems about guaranteed generalization, if possible. The book I have out from the library by Scholkopf and Smola will be helpful for this.

(iv) *Venues.* Seems like a perfect ICML paper if we can do the above items, so if we can pull them off I think the chances of success are high. However we will have to work hard to do the above. I worry that none of us knows classical time series analysis - we need to remedy this ASAP. But I think it is feasible to do it.

Strong consistency of Ulrike's clustering framework

(i) *Idea.* In "Nearest Neighbor Clustering: A Baseline Method for Consistent Clustering with Arbitrary Objective Functions" (Bubeck and von Luxburg, JMLR 2009), the authors present a framework for clustering which is like empirical risk minimization. They present an algorithm called nearest neighbor clustering (NNC), and they show that NNC is weakly consistent under certain conditions.

Recall that a rule is said to be weakly consistent if the empirical risk converges to the true risk in probability as the sample size tends to infinity, while a rule is said to be strongly consistent if the convergence happens almost surely. Our idea is to try to upgrade NNC to strong consistency (possibly under more restrictive conditions than Bubeck and von Luxburg stipulate). Bubeck and von Luxburg point out a sticking point to strong consistency, but they do not seem to have spent a lot of effort overcoming it.

The nice thing about this idea is that (a) it is well defined and of some importance, but seems not to have been well-explored, (b) I spent a lot of time last semester learning how to prove convergence almost surely, so I feel that it is within our reach, and (c) it could lead to a connection with Ulrike.

(ii) *Done.* Nothing, but I have read the paper fairly closely and have some ideas about how to proceed.

(iii) *Todo.* See above.

(iv) *Venues.* Seems like a solid COLT paper.

Correcting OCR

(i) *Idea.* The idea is to make an undirected graphical model, or other model, to correct OCR in 18th century documents. As training/testing data, we have a moderately sized collection of documents each in two versions, where one version has been manually cleaned up and one is not.

(ii) *Done.* We discussed last fall a model, and we have the data (including cleaned up versions).

(iii) *Todo.* Need to implement the model. There is software by Hal Duame that may make this easy. Need to match cleaned up and raw versions of each document. I have software which we can modify to mostly automate this. Need to come up with a baseline or competitors and an experimental methodology. Need to justify even attacking this problem in 2011.

(iv) *Venues.* We could try to submit this to AAAI - they apparently have a few NLP papers each year, and since it uses an undirected graphical model it seems at least a little AIish. Otherwise it would make a lot of sense - perhaps more sense - as a ACL short paper (deadline Feb. 25), and we can do the whole thing after AAAI/ICML/COLT. I think the prospects at AAAI are moderate but the chances at ACL are fairly high.

Deadlines and other ideas for projects

jan 19	ijcai abstracts
jan 24	ijcai papers
feb 1	icml
feb 3	aaai abstracts
feb 8	aaai papers
feb 11	colt
feb 25	acl short papers

COLT

One idea. What is the unique (if any) clustering algorithm that gives you a "best" clustering in terms of Meila's VI measure? For this look at "A nonparametric information theoretic clustering algorithm" in icml 2010.

compare approaches to multiview clustering

stability of hierarchical clustering

persistent k-means

show *strong* consistency of the thing in bubeck09a, by finding/estimating the constant in lemma 11

one of the open problems in that website - see first paper in colt10 proceedings for link

ICML

New feature selection approach based on homology. Like my paper for bioinformatics.

make new approach to multiview clustering

thorough empirical comparison of approaches to clustering stability esp normalization

temporal erm

AAAI

cs760 paper

tensor treelets

ACL

correcting OCR