Approaches to Clustering Under Stability with an Expert Oracle
Project Proposal for 15-300, Fall 2016

Apoorva Bhagwat

1 Logistics

Title. Clustering Under Stability with an Expert Oracle

Website. I will use `abhagwat.github.io` to organize material related to my project.

2 Description

What? Clustering is an unsupervised machine learning problem. Given a set of points in some form (for instance, as vectors in a Euclidean space, or as a graph structure), we want to group these points into clusters so that 'similar' points are in the same cluster. Here, the notion of similarity depends on the domain of our problem. Two common formulations of this problem are k-means clustering and k-medians clustering. These are examples of *center based clustering*, because they involve finding centers for each cluster by minimizing some distance-based objective function within each cluster.

Unfortunately, both k-means and k-medians clustering are NP-hard, so we have to turn towards approximation algorithms, and investigate reasonable properties of data sets that can let us approximate clustering efficiently. A lot of research has already been done about criteria that yield efficient algorithms.

Another way to make clustering computationally feasible is to introduce an 'expert' oracle (making the learning model semi-supervised) that answers queries of the form 'do these two points belong to the same cluster?'. It turns out that introducing this oracle makes gives us an arbitrarily good polynomial time approximation scheme for clustering under a certain stability condition known as γ-margin separability, using a small number of oracle queries.

I will be exploring the relationship between query complexity and computational complexity in the presence of an oracle, and how we can solve clustering efficiently using a small number of queries. It is also interesting to see how expert oracles affect the complexity of related problems such as correlation clustering and ranking. Even though I have written about multiple things I discussed with Avrim, I plan to start by focusing primarily on correlation clustering.

One aspect of expert oracles that I haven't seen being explored yet is allowing the oracle to be noisy (perhaps the oracle makes mistakes with a certain probability, or responds 'don't know' for a certain fraction of inputs). This may be a more accessible and interesting research direction.

Who? I will be working with Prof Avrim Blum. One of his PhD students, Nika Haghtalab, is currently away, but I plan to meet her when she returns to CMU, and I hope to get her guidance as well.

So what? Clustering with expert oracles can be used to model a number of real world scenarios, including :

1. Learning algorithms that can use humans as assistants (facial clustering algorithms, etc.)
2. Learning algorithms in computational biology that use lab experiments as 'expert advice'. The experiments can be very expensive to carry out, so we need to design low query-complexity algorithms for this domain.
3. Learning algorithms that use services like Amazon Mechanical Turk to gather expert advice.

3 Goals and Milestones

75% Goal. At the very least, I would like to come up with an algorithm for correlation clustering using an expert oracle with a low query complexity. Such an algorithm was recently published (by Ashtiani et al) for k-means clustering, so hopefully I will be able to generalize the result to correlation clustering.

100% Goal. My ideal goal would be to come up with a low-query complexity algorithm for correlation clustering, along with a hardness result that lower bounds the number of queries to efficiently solve correlation clustering.

125% Goal. If I manage to achieve my 100% goal, I will try to improve upon the bounds given by Ashtiani et al. These are upper and lower bounds on the number of oracle queries necessary to approximate k-means clustering in polynomial time. It would probably make sense to try to improve bounds in the correlation clustering case instead of completely switching domains.

End-of-semester milestone. By the end of this semester, I hope to become familiar with existing work in the area of clustering with an expert oracle, get up to speed on existing work on correlation clustering.

January 30th Investigate which stability conditions make correlation clustering computationally feasible, and become familiar with algorithms that exploit these conditions. Understand the related hardness results, and figure out if these stability conditions can be weakened when we have access to an oracle.

February 13th Start conjecturing algorithms for simple cases of correlation clustering. Experiment on a computer to decide if these simple algorithms are actually correct/efficient.

February 27th Iterate on the simpler algorithm(s) and generalize them to harder instances. Continue with experimentation, and try to prove the algorithm correct, along with its approximation guarantees.

March 20th Do a careful analysis of the running time of the algorithm and come up with a clean formulation of the algorithm that makes it easy to prove that it is efficient.

April 3rd Conjecture a lower bound on the number of queries required for efficient clustering. Make sure that this bound actually holds on small examples / by experimenting in a program.

April 17th Try to come up with a proof of this lower bound. This will most likely involve showing NP-hardness of the problem when the number of queries allowed is low.

May 1st Refine the work done during the semester and write it up for presentation.

4 Literature

I've already read these two foundational papers about clustering under stability in detail (as a part of assignment 2 for 15-300) :

- Awasthi, Pranjal, Avrim Blum, and Or Sheffet. "Stability yields a PTAS for k-median and k-means clustering." Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on. IEEE, 2010.
- Balcan, Maria-Florina, Avrim Blum, and Anupam Gupta. "Clustering under approximation stability." Journal of the ACM (JACM) 60.2 (2013): 8.

I am currently reading the following paper and other material it refers to :

- Ashtiani, Hassan, Shrinu Kushagra, and Shai Ben-David. "Clustering with Same-Cluster Queries." arXiv preprint arXiv:1606.02404 (2016).

I also plan to read the following papers to get more background on stability conditions, and hardness of clustering. These are more general in scope than the previous papers :

- Ben-David, Shai. "Computational feasibility of clustering under clusterability assumptions." arXiv preprint arXiv:1501.00437 (2015).
- Balcan, Maria-Florina, Nika Haghtalab, and Colin White. "k-center Clustering under Perturbation Resilience." arXiv preprint arXiv:1505.03924 (2015).
- Dasgupta, Sanjoy. "The hardness of k-means clustering." Department of Computer Science and Engineering, University of California, San Diego, 2008.