Noname manuscript No.
(will be inserted by the editor)

Comments On: Distance Geometry and Data Science

Robert J. Vanderbei

Received: January 31, 2020/ Accepted: someday

This paper provides an excellent introduction to a broad class of modern problems called *distance geometry problems*. The paper begins with a nice and simple definiton of this family of problems and then further motives the study of this class by giving a number of interesting real-world examples where these problems play an important role. These examples range from the analysis of nuclear magnetic resonance (NMR) data to determine the molecular structure of important biochemicals such as proteins and DNA to speech (and writing) recognition and on to the broad class of machine learning problems in which large data sets are mapped into much lower dimensional spaces and then clustered intelligently into meaningful subsets.

After making the importance of distance geometry problems clear, the paper naturally transitions into a discussion of the various algorithms that one can use to solve these problems. The problems are fundamentally nonconvex nonlinear optimization problems and so there are many different approaches one can employ ranging from fast simple heuristics, to linear inner and outer approximations, and on up to interesting semidefinite programming (SDP) approaches.

The semidefinite programming approach often provides the best answer to the problem but, as is well-known, algorithms for solving SDP's don't scale well with problem size and so this leads one to consider heuristic approximations that provide good solutions very quickly. The paper discusses a few different such approximations. My favorite one is the diagonal dominance approximation. This approxiamtion is fairly simple to describe (see the paper for the details) and provides a good approximation.

Robert J. Vanderbei
Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544
Tel.: +609-2585-2345
E-mail: rvdb@princeton.edu

The last section of the paper discusses how the various methods described in the paper perform on clustering sentences in the (available from archive.org) text of H.D. Thoreau's book *On the Duty of Civil Disobedience*.

I have spent decades of my life engaged in research into optimization both from a theoretical perspective and the more applied implementation of efficient algorithms. But, I haven't been much involved in the more modern application of optimization tools to help solve problems in data science and machine learning. However, I do teach an introductory optimization course here at Princeton and in that educational role I have felt that I should introduce the students to some of the more modern applications in the area of data science and machine learning. To that end, I would like to briefly describe the final projects I made for my class over the last few years. The modern tool I introduced in these final projects involved solving support vector machine (SVM) optimization problems to solve a few interesting and practical real-world problems.

For the first SVM project I had each student in the class write out the entire alphabet six times on a piece of paper. I scanned each students alphabets and made a database of about 400 hand-written alphabets for the training set for the SVM. I also had each student write a message that I also scanned. Each student was then tasked with using their support vector machine to "read" the messages written by the other students in the class. The support vector machine did a pretty good job of identifying the handwritten characters. But, it was far from perfect.

The last time I taught the class, I had them do a simpler support vector machine. Rather than trying to classify digitized scans of letters into one of 26 possible letters, I thought it would be easier to do a binary type classification problem. Inspired by that desire and given online access to profile pictures of each student in the class, I thought it would be interesting to see if a support vector machine could easily be trained to do gender recognition. To test this out, I downloaded the profile pictures of the roughly 70 students enrolled in the class. I converted the digital jpeg files to a 226×164 pixel array for each of the red, green and blue channels. Each picture can therefore be vectorized into an array of size $226 \times 164 \times 3 = 111{,}192$. This seemed very much too high dimensional for a simple problem like gender recognition. So, I converted each picture to grayscale by averaging the red, green, and blue frames and then I downsampled the pixels by a factor of 10 in each dimension. The downsampled data consists of arrays of size $23 \times 17 = 391$. Using just 20 randomly selected downsampled images as the training set and then predicting the gender associated with each of the other images, the SVM got the correct answer about 97% of the time. To me this was amazingly surprising. And, it raises an interesting question: might one or some of the dimensional reduction techniques described in Liberti's paper be able to reduce things to a dimension significantly smaller than 391 and still perform as well as the simple binning procedure I used? I hope someone, I myself or Leo Liberti or someone else who reads this paper, will be able to answer this question soon.

I enjoyed reading the paper. It gives a very good explanation of the many ideas and tools and applications of distance geometry. It also provides an

excellent set of references to places in the literature where one can continue learning more about this interesting subject.