Statistics meets Data Science

The June meeting of the Victorian branch was – at least for this reporter – unusually exciting. First, we need to say that SSA-Vic has recently embraced Meetup, and so intending attendees were invited to announce this fact in advance on the web, and supply a photo and some insightful comments. Over 50 did so. This had the effect of generating an unusual level of pre-meeting excitement. Secondly, the meeting was held jointly with Data Science Melbourne. This conjunction was the result of members of the branch council reaching out to the local data science community. It worked. Lots of data scientists, business analysts and others rolled up, people who would be most unlikely to attend a regular branch meeting. And thirdly, the meeting itself was not held in a campus lecture room, but on the 23rd floor of a city building, at the location of a hip software company known as ThoughtWorks. People started interacting over wine, cheese and other snacks about 45 minutes before the meeting proper commenced, and by that time about a hundred people were eating, drinking and all talking at once. If the Meetup site generated a buzz, the pre-meeting interaction generated a clamour. Final attendance came in at 111 people, and though people weren't quizzed about whether they were statisticians, data scientists or neither, it was pretty clear that branch members were in the minority. But having said that, the branch turnout was very impressive: many young and old members were there.

It was with great anticipation that the meeting began around 6:15pm. The main speaker, Alec Stephenson was preceded by a short message from Maia Sauren, a hacker for humanity representing the hosts Thoughtworks. Currently a Research Scientist in CSIRO, Alec is a card-carrying statistician with a strong record of success in Kaggle data mining competitions. His title was "A Statistical Viewpoint on Data Science, Data Mining and Big Data." The notice of the talk promised that "In this non-technical talk he will examine the relationship between statistics and data science. He will illustrate the skills needed to be a successful data miner for both consulting and for predictive competitions. And he will give his two cents on the big data hype." I'm happy to report that he did all that, and more.

For Alec, as for Drew Conway, data science is the intersection of "Math and Statistics Knowledge, Substantive Expertise, and Hacking Skills". More fully, in Alec's view, essential skills for Data Science are "Statistical Modelling: e.g. R, Matlab, Python" and "Data munging: e.g. Perl, Python, Ruby", while additional skills are "Fast computation: C, C++, Java; Data Storage: SQL, noSQL; and Big Data: MapReduce, Mahout, Hive, Pig." Moving on to the skills needed to win data mining competitions, Alec said there are only three: "Data Munging, Statistical Modelling, and Ensembling." I'm sure more than a few people thought as I did: if only it was that simple to win a competition! This list of three skills was supplemented by several pieces of advice, one of which was "If something takes more than one minute to run, do you really need to run it?" Alec illustrated "ensembling" on Fisher's iris data by creating a prediction problem with it, and then averaging the results of random forest, glmnet and gradient boosting predictors. It helped. Later he mentioned the idea of stacking

predictors. Following this demonstration of the power of averaging, Alec went on to contrast the important issues when dealing with clients with those needed to do well in competitions. His point: clients are typically concerned with more than winning.

The final part of Alec's talk concerned Big Data. His statistical viewpoint on Big Data is that there is only a very limited role for statisticians in that domain. Why? For him, Big Data begins at about .5TB, which he suggested is well beyond the experience of most statisticians. Further, he told us, if you need to "touch all the data (0.5TB+)" you are probably "restricted to linear (or logarithmic) algorithms" (sums, averages, graph search or sorting). In other words, you couldn't do anything very statistical with such a large amount of data anyway. He did concede that statisticians may have a role in deciding what data is relevant to a given question, to subsetting or sampling the Big Data and modeling those subsets, but that was about all. He ended his talk by pointing out that some statisticians are paying only lip service to the notion of Big Data, and remarking that he personally thought it was over-hyped. With that, he wound up to applause, and we turned to the discussion.

I know I wasn't the only one in the room that viewed Alec's very limited role for statisticians in Big Data with some dismay, if not skepticism. It seemed to be rooted in a rather narrow view of statisticians (which of course might be wholly justified), an assumption that a technology gap will remain fixed (0.5TB as too big for statisticians), and his personal reluctance to run programs on computers for more than a minute. It is worth pointing out that Alec did not give a single example of what was for him a Big Data problem. In my line of work, a single 11-day run of a HiSeq 2500 DNA sequencing machine in high output mode can produce 600GB of data. In such cases, we do need to "touch all the data", but by running algorithms (which are not linear or logarithmic) for hours or days on a cluster, the DNA reads can be mapped to a genome, and the original data set reduced dramatically in size, ready for more careful statistical analysis. This doesn't seem to be such an uncommon scenario: a lot of data and a lot of computing leads to summarized data, ready for statisticians to go to work on: not only DNA, think images, audio or text.

Some of these issues were touched upon in the discussion after the talk, and we all left wiser than we arrived, greatly heartened by the upsurge of interest in – if not *Statistics*, then statistics by another name.

Terry Speed, 5 July 2014