SM4 Data Mining and Machine Learning - Week 5 Practical

Dino Sejdinovic

February 16, 2017

1 Collaborative Filtering on UK Parliament Data (unassessed)

We will analyse data collected by Public Whip on the votes each Member of Parliament (MP) has cast during the 2010-2015 parliament in the House of Commons and try to infer latent factor representations for MPs in order to visualise them in 2 dimensions. Each vote corresponds to a particular piece of legislature and is called a *division*. Download these data here and to load use:

```
In [ ]: library(readr)
        df <- read_csv("~/data/publicwhip/parliament2010.dat")
        votes <- df[,2:1227]
```

There are 664 MPs and 1226 divisions. In the data matrix `votes`, $+1$ indicates an `aye` on the division, -1 a `no`. 0 indicates that the MP did not vote in that particular division. In reality, it is a bit more complicated than this as there are also *tellers* as well as the MPs who voted both ways (!), but I cleaned up the original format so that we can treat it as a simple binary dataset. Now let us consider the number of votes each MP has cast. It turns out that there is a huge range in these: there is an MP who voted only once(!), whereas another one voted in 1179 out of 1226 divisions. Consider removing the MPs who voted 5 times or less from the dataset (can you explain why they voted only a few times?)

To start with, let us ignore the fact that votes which were not cast are essentially missing data (due to MPs not attending sessions, for example) and simply compute the principal components projections on our data matrix. Note that $p > n$ here and do not forget to center your data matrix.

Consider the projections to the top two principal components – they should already indicate a good split between the major parties. What can you say about MPs in each of these parties who are located towards the middle of the plot? Is it an indication of outlier MPs - some form of rebelliousness? Or just an indication that those MPs did not vote much? Can you recognise the names of any of those MPs? Investigate the correlation between the number of votes cast and the distance of the projected points from the mean of the principal components projections – which is $(0, 0)$.

To deal with this issue treat the problem as collaborative filtering instead and make two enhancements to the latent factor learning: first, treat the 0s as missing data; second, make use of the fact that the observations are binary (`aye`, `no`) and use a generalised linear model with logistic links. Formulate the problem either as alternating logistic regressions or write down the overall objective function and implement stochastic gradient descent to learn both the latent factors of MPs and of divisions. Investigate the correlation between the number of votes cast and the distance of the projected points from their mean in the new latent space.