MathStats: Project Information

Prof. Dr. Tillmann Schwörer

Summer Semester 2020

1 Jokes Analysis

1.1 Setup

- Your team collects a selection of jokes (~ 25), ideally jokes that appeal to different humour preferences.
- Each course participant must rate a mandatory subset of jokes (~15). Additionally each participant rates a subset of the non-mandatory jokes.

	joke1	joke2	...	joke15	bonus1	bonus2	bonus3
participant1	x	x		x	x		
participant2	x	x		x		x	
...							
participant10	x	x		x			x

1.2 Possible analytic qustions

- **Distances**: Measure distances of jokes and/or persons in terms of joke ratings, using alternative distance measures
- **Recommendations**: Make an individualised joke recommendation. Recommend one of the bonus jokes that the closest neighbour liked
- **Clustering**: Use k-means or hierarchical clustering to identify homogenous clusters of jokes or persons
- **Principle component analysis/Singular value decomposition**: Find a lower dimensional representation (latent factors) of persons and/or jokes. Think of, e.g.:

	dark	wordplay	irony
participant1			
participant2			
...			
participant10			

2 Songs, movies, or books analysis

The same setup and the analytic questions as above can be addressed using an alternative domain, e.g.:

- Songs
- Movies
- Books

Note: It won't be feasible to get a full set of ratings for 15 movies or books from each participant. Hence, you will have a *sparse* rating matrix. This makes some of the analytic questions more difficult or unfeasable (notably clustering and PCA/SVD). However, it is possible to measure distances for each pair of movie or book raters, based on the subset of movies/books that both have rated. And thus, to make personalized recommendations.

3 Paper review

Alternatively, you can summarise and present one of the following papers on recommender systems:

- Koren, Bell, Volinsky: "Matrix Factorization Techniques for Recommender Systems" (2009)
- Hu, Koren, Volinsy: "Collaborative Filtering for Implicit Feedback Datasets" (2008)
- Reidy: "An Introduction to Latent Semantic Analysis (2009)"
- Goldberg et al: "Eigentaste: A Constant Time Collaborative Filtering Algorithm (2000)"
- Nathanson, Bitton, Goldberg: Eigentaste 5.0: Constant-time Adaptability in a Recommender System Using Item Clustering (2007)

4 Regression analysis

What drives data scientists' income? Try to answer this question using data from stackoverflows annual developer survey, and linear regression techniques. Pay particular attention to possible determinants of income which we can actually influence. (I can't easily become a women, but maybe I would be willing to learn another programming language in order to earn some more money...). Note that this topic will require a substantial amount of data cleaning (which is a very valuable thing to learn and practice), for instance with respect to outliers.

5 Bring your own topic