Generating Place Recommendations for Travelers

Andy M. Gilbert, Andrew M. Hilger – CS 229 Fall 2017 Final Project

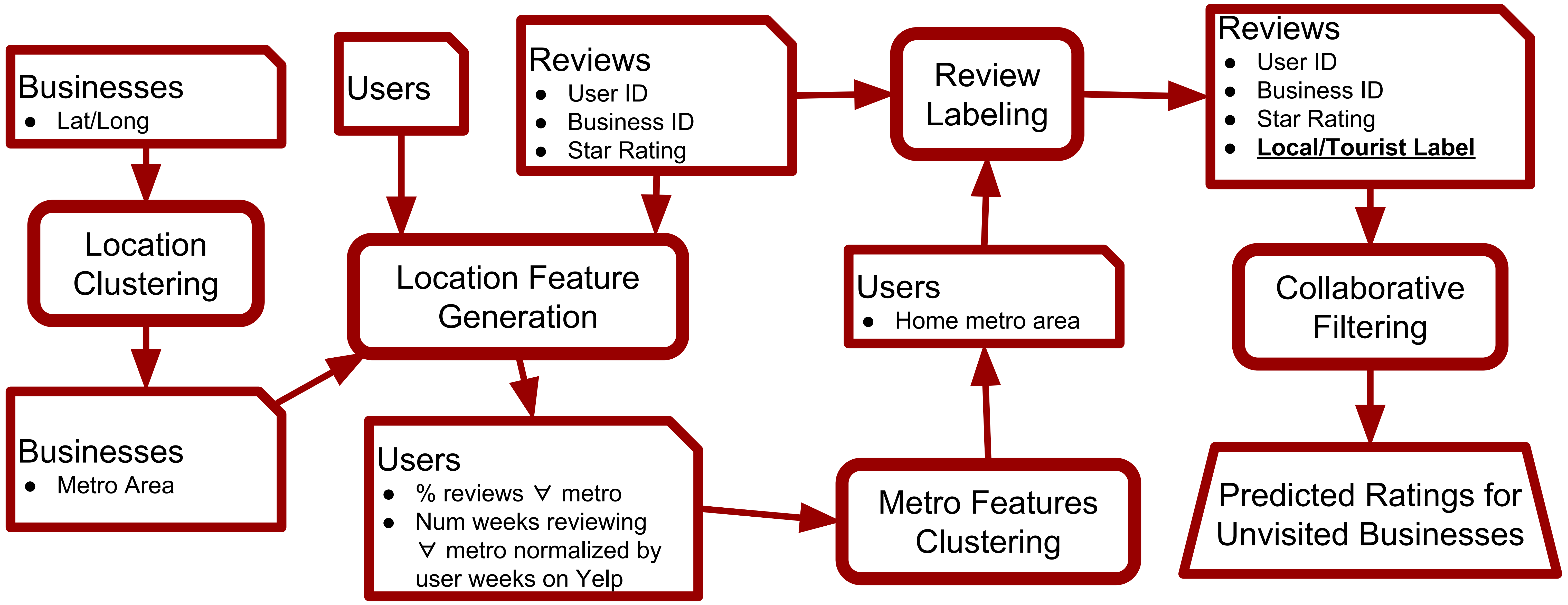
Overview

As vacation time becomes more precious, travelers have less time to plan their vacations. Services such as Yelp as useful for finding good restaurants and other attractions. However, these businesses are primarily reviewed by locals and don't necessarily account for the differing needs and tastes of travelers. Using the Yelp dataset [1], we investigated whether travelers would be better served by a recommendation system that is aware of whether they're traveling.

To make recommendations, we implemented collaborative filtering, with sparse factor analysis. The model predicts business ratings by a user using a set of reviews by that user for other businesses, as well as other user's reviews of those businesses. We trained separate models for traveler reviews and local reviews.

However, the Yelp dataset does not label reviews as being "travelers" or "locals," nor does it even provide any information about the user's home location. Thus, we used unsupervised learning to determine the most likely home metro area of a user, then labeled the user's reviews as "local" or "tourist" reviews depending on whether the business's metro area location matched the user's inferred home metro area.

Methodology



Businesses
- Lat/Long

Location Clustering

Businesses
- Metro Area

Users

Reviews
- User ID
- Business ID
- Star Rating

Location Feature Generation

Users
- % reviews ∀ metro
- Num weeks reviewing ∀ metro normalized by user weeks on Yelp

Review Labeling

Users
- Home metro area

Metro Features Clustering

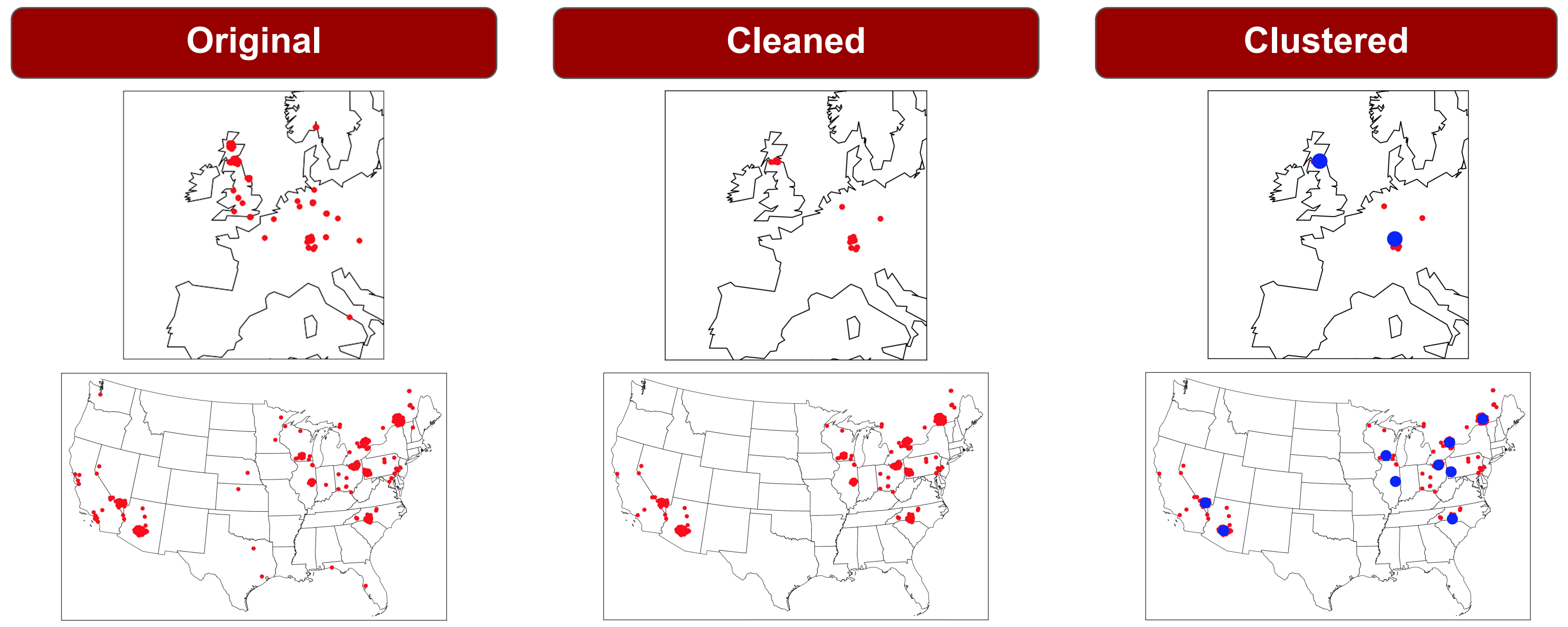
Reviews
- User ID
- Business ID
- Star Rating
- **<u>Local/Tourist Label</u>**

Collaborative Filtering

Predicted Ratings for Unvisited Businesses

The dataset included 1.18M users and 156k businesses. After filtering users with fewer than 20 reviews and businesses with fewer than 30 reviews, 251,675 users and 31,147 businesses remained. The user metro features clustering used 2.02 million reviews corresponding to these users and businesses, using 20 features each corresponding to one of 10 metro areas. For collaborative filtering, we considered 4956 businesses in the Las Vegas area. The reviews for these businesses corresponded to 134k users labeled as locals and 12.2k users labeled as tourists.

Business and User Clustering



Left: Locations of businesses (all points were located in Europe, USA, and Canada). Center: In order to assign businesses to a metro area we first cleaned the data to remove all points that were outside of states with a major metro area represented in the dataset. Right: We then performed k-means clustering using (lat,long) location data. The blue dots correspond to metro areas used to generate user metro features. Users were clustered into metro areas based on percentage of reviews for each user in a given metro as well as number of weeks the user had written reviews in an area normalized by user weeks on Yelp.

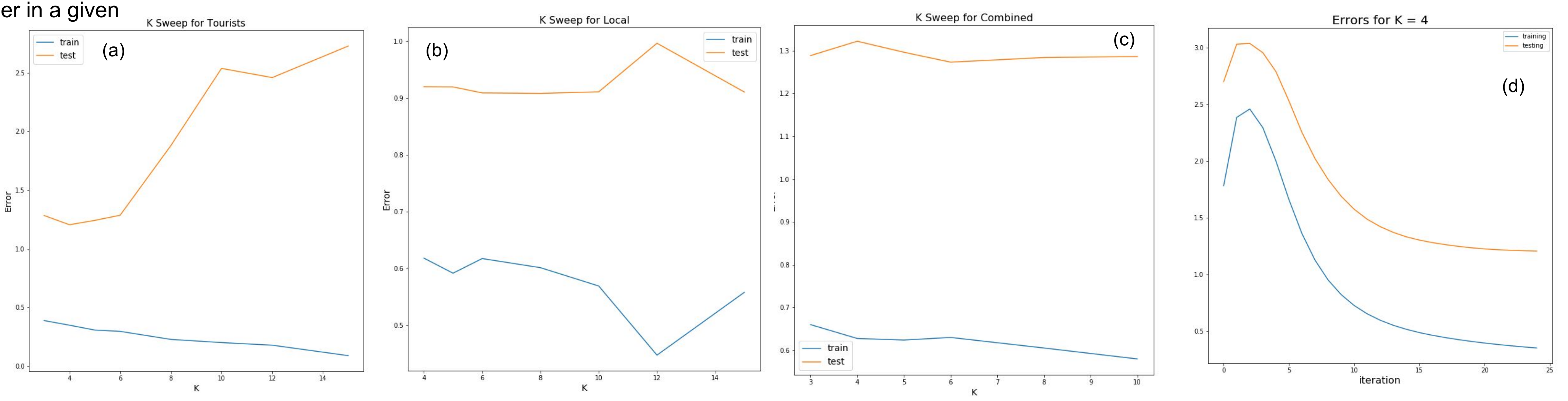
Collaborative Filtering

We implemented collaborative filtering using a sparse factor analysis as described in [2]. Given a set of reviews $Y_{n\times m}$ for m users who have reviewed n businesses, factor analysis identifies k latent factors that explain the observed reviews according to $Y = \Lambda X + N$ where each row of $\Lambda_{n\times k}$ represents the factor loading of a business, and each column $X_{k\times m}$ represents the factor preferences of a user. N represents additive noise.

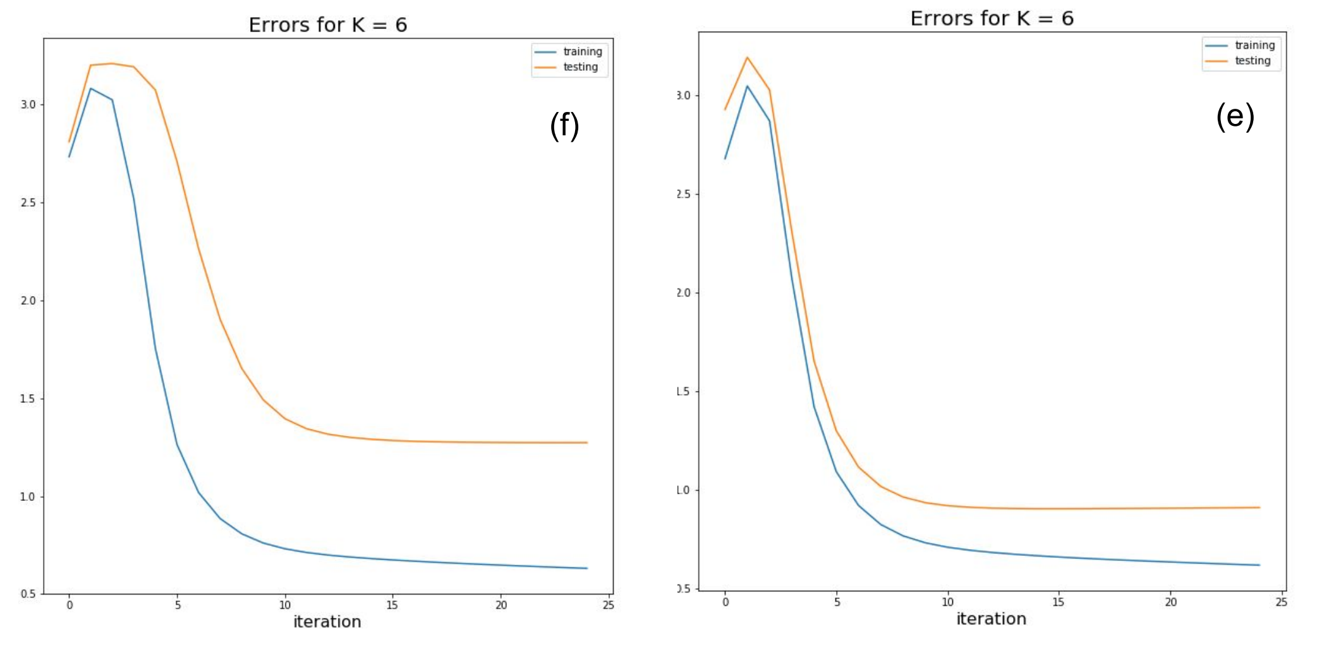
The model is trained using expectation maximization, where the E step determines X for each user, and the M step determines Λ and an estimate of noise variance ψ for all the businesses. This model provides better accuracy than SVD and better scalability than a Pearson-correlation coefficient-based model or a personality-diagnosis model, which accounts for user-specific features [3].

We evaluate the model's performance using mean average error (MAE) between the actual star ratings and the star ratings predicted using the learned Λ and X.

Results



(a)-(c) Training and test MAE as a function of K. (d)-(f): Learning curves for training and test MAE for optimal K. Tourist set is (a)&(d). Local set is (b)&(e). Combined is (c)&(f).



Data Set	Optimal k	Training MAE (80% of n,m)	Test MAE (20% of n, m)
Tourists + Locals (control); M = 146k	6	0.6304	1.273
Tourists Only; M = 134k	4	0.3492	1.205
Locals Only; M = 12k	6	0.6174	0.909

MAE for the Tourists + Locals control was evaluated on reviews from tourists only to ensure comparability with the tourists only set.

Conclusions

- Tourist reviews were best modeled with k = 4 canonical preferences, while local reviews were best modeled with k = 6 canonical preferences. One interpretation is that tourists have less sophisticated preferences than locals. Alternatively, it is more advantageous to use lower k because there are fewer tourist reviews, so a lower k is required to prevent overfitting.
- Training separately on tourists and locals improved predictions for tourists relative to training on both together, but the improvement was not significant (~0.1 improvement in MAE). More tourist reviews for training would likely improve the MAE.
- Future work: Compare performance of tourist/local-aware collaborative filtering across all metro areas to investigate differences in tourists among different cities. Given significantly more data, we would investigate training tourists separately by tourist home metro area.

References

(1) "Yelp Dataset Documentation". Yelp. https://www.yelp.com/dataset/documentation/json

(2) Canny, J. "Collaborative Filtering with Privacy via Factor Analysis." *SIGIR* 2002,August 11-15, 2002, Tampere, Finland. <https://pdfs.semanticscholar.org/9336/1b6326dbd93aa8cc2e54c5260a9216feb039.pdf>.

(3) Su, X. and T. Khoshgoftaar. "A Survey of Collaborative Filtering Techniques" *Advances in Artificial Intelligence*, 2009, 421425 <https://www.hindawi.com/journals/aai/2009/421425/>