Personalized Location Models with Adaptive Mixtures

Moshe Lichman, Dimitrios Kotzias, Padhraic Smyth
Department of Computer Science
University of California, Irvine
{mlichman,dkotzias,smyth}@ics.uci.edu

ABSTRACT

Personalization is increasingly important for a range of applications that rely on location-based modeling. A key aspect in building personalized models is using population-level information to smooth noisy sparse data at the individual level. In this paper we develop a general mixture model framework for learning individual-level location models where the model adaptively combines different types of smoothing information. In a series of experiments with Twitter geolocation data and Gowalla check-in data we demonstrate that the proposed approach can be significantly more accurate than more traditional smoothing and matrix factorization techniques. The improvement in performance over matrix factorization is pronounced and may be explained by the tendency of dimensionality reduction methods to over-smooth and not retain enough detail at the individual level.

CCS Concepts

•**Computing methodologies** → **Mixture models;**

Keywords

Location data, Personalization, Mixture models

1. INTRODUCTION

In recent years we have gained the ability to record human spatio-temporal behavior in increasingly fine-grained detail. At the individual level this type of data holds the promise of *personalization:* delivering information, products and services in a manner that is optimized for each specific individual.

A key challenge in personalization is being able to generalize about an individual's behavior and preferences beyond their historical data. For example, with location data it is likely that individuals will visit both old and new locations. The challenge then becomes how to model their propensity to visit new locations while respecting their tendency to revisit locations from their past. This problem is particularly challenging in the presence of sparse data.

SIGSPATIAL'16, October 31-November 03, 2016, Burlingame, CA, USA

ISBN 978-1-4503-4589-7/16/10…$15.00

DOI: http://dx.doi.org/10.1145/2996913.2996953

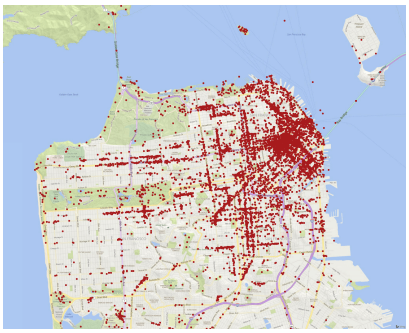

Figure 1: Gowalla check-in data over a 20 month period in San Fransisco, CA.

In this paper we consider the problem of generalizing from sparse location data to learn models that can make accurate future predictions at the individual level. Our goal is to construct for each individual i a model in the form of a multinomial (categorical) distribution $\boldsymbol{\theta}_i$ over M locations, where θ_{ij} is the probability that the location is j given that individual i generated an event, with $\sum_{j=1}^{M} \theta_{ij} = 1$.

Our model uses smoothing to balance historical data from an individual with additional information not in the individual's history. We utilize information from the global population, inter-location proximity, and social relations between individuals, to generate appropriate smoothers, which are then combined with the user's personal history to create a personalized model for location prediction.

The primary contributions of this paper are (1) an adaptive mixture model framework that infers appropriate levels of smoothing, and (2) experimental results demonstrating systematic improvements in prediction accuracy (compared to traditional global methods) on a range of large-scale location data sets.

2. BACKGROUND

Our focus in this paper is on building models from location data in the form of event tuples $\langle i, x, y, \tau \rangle$ where i is the individual who generated the event, (x, y) is the location (e.g. longitude and latitude from GPS) and τ is a timestamp. Figure 1 shows an example of this type of data for San Fransisco. If time is collapsed, this type of data can be represented in the form of a sparse matrix of N individuals by M locations where cell i, j contains a count of the number of times individual i visited location j. From this data we wish to infer predictive distributions $\boldsymbol{\theta}_i$ over the M

locations for each individual.

Matrix factorization and collaborative filtering techniques are widely used in this context, but can oversmooth individual level details and potentially lead to suboptimal predictive performance (e.g., see [1]). One way of incorporating individual-level information is to learn a geographical preference for each individual independently and combine with the smoothed predictions from collaborative filtering or matrix factorization methods. The geographical component lowers the scores of locations that are far away from the individual and increases those that are close by [1, 10, 11]. However, these methods do not provide a direct way to control the degree to which the geographical context component affects the predictions.

An alternative approach is to use the geographical preferences for each individual during the learning process [6, 7, 8, 9]. These methods project the data into a lower dimensional space while taking into account geographical preference. While this improves location prediction over direct matrix factorization, it does not take advantage of other available sources of information such as social ties.

Our contribution in this context is to develop alternatives that are better at retaining the details of each individual's behavior. In particular our model blends an individual's historical data with information about geographic and social similarity using a mixture model approach.

3. MODEL DESCRIPTION

3.1 Adaptive Smoothing with Mixtures

The two key ideas behind our approach are to (1) smooth individual-level information towards population patterns, geographic constraints, and social contexts, and (2) learn to combine these sources in a manner that optimizes predictive performance.

Let $\boldsymbol{\theta}_i = [\theta_{i1}, \ldots, \theta_{iM}], \sum_j \theta_{ij} = 1$, be an unknown probability distribution over M locations for individual i. The maximum likelihood estimate for each probability is $\theta_{ij}^{MLE} = \frac{r_{ij}}{r_i}$, where r_{ij} is the historical count for individual i and for location j and $r_i = \sum_j r_{ij}$. This estimate is problematic for sparse data where r_{ij} values are often zero.

An alternative approach is the Bayesian framework where the $\boldsymbol{\theta}_i$ parameters are treated as random variables with a prior [3]. In particular, given a Dirichlet prior with parameters $\alpha_1, \ldots, \alpha_M$, the mean posterior estimate (MPE) (and also the predictive probability) for the probability of category j given individual i can be written as

$$\hat{\theta}_{ij} = \frac{r_{ij} + \alpha_j}{r_i + \alpha} = \gamma_i \frac{r_{ij}}{r_i} + (1 - \gamma_i)\frac{\alpha_j}{\alpha}$$

with $\alpha = \sum_{j=1}^{M} \alpha_j$ and where $\gamma_i = \frac{r_i}{r_i + \alpha}$. Thus, the posterior mean can be viewed as a convex combination of individual data and a prior, where the mixing coefficient γ_i serves to smooth the individual estimate towards the prior.

With large numbers of individuals, another option for this problem is to use an empirical population estimate instead of a prior, and write the estimate in the form $\boldsymbol{\theta}_i = \pi \boldsymbol{\theta}_i^I + (1 - \pi)\boldsymbol{\theta}^P$ where $\boldsymbol{\theta}_i^I$ is the maximum likelihood (frequency) estimate given individual i's data , $\boldsymbol{\theta}^P$ is the population distribution based on pooling the location counts across all individuals, i.e., $\theta_{ij}^P = \frac{\sum_i r_{ij}}{\sum_i r_i}$, and where $\pi \in [0, 1]$ is a mixing coefficient.

A general weakness of this type of "population smoothing" is that it pulls all estimates in the same direction, i.e., towards the global $\boldsymbol{\theta}^P$ in our example above. An alternative approach is to smooth towards data more relevant to individual i. If we have some knowledge about the similarities of our rows (individuals) and our columns (locations) then two additional smoothing strategies immediately suggest themselves, namely smoothing using data from individuals that are similar to individual i, and smoothing using data from locations (columns) that are similar to the columns that are present in individual i's data. In particular, in this paper we use social network friendships to measure similarities of individuals and spatial distance for similarity of locations.

3.2 Defining the Mixture Components

Given the above we can set up a model (described below) where there are 4 components: one "unsmoothed individual" component, a population component, and two components based on column and row smoothing, I, P, L, S respectively. We represent each individual as a convex combination (finite mixture) of component distributions as $\boldsymbol{\theta}_i = \sum_{k \in \{I,P,L,S\}} \pi_k \, \boldsymbol{\theta}_i^k$, where the π_k are mixture weights[1] with $0 \leq \pi_k \leq 1, \sum_k \pi_k = 1$, and where the components are each a categorical distribution over the M categories (locations), with elements θ_{ij}^k for each component defined as follows:

***I*: Individual Estimate**: $\theta_{ij}^I = \frac{r_{ij}}{r_i}$, where r_{ij} is the count for individual i and location j and $r_i = \sum_{j=1}^{M} r_{ij}$ is the total count for individual i. This is the maximum likelihood (frequency-based) estimate of an individual's distribution.

***P*: Population Smoothing:** $\theta_j^P = \frac{r_j + \alpha'}{r + M\alpha'}$, where $r_j = \sum_{i=1}^{N} r_{ij}$ is the count for location j across all individuals, $r = \sum_{j=1}^{M} r_j$ is the sum of all counts (across all locations and all individuals), and α' is the parameter of a flat symmetric Dirichlet smoothing prior (to allow for non-zero probability mass on locations that have no counts in the historical data). This is a smoothed global estimate of location frequencies across the whole population, i.e., how popular a location is in general. In the results later in the paper we used $\alpha' = \frac{10}{M}$.

***L*: Location (Column) Smoothing:** We use kernel diffusion weight matrices [5] $K^L(j, j')$ and $K^S(i, i')$ for column and row smoothing respectively, where $K^L(j, j') \geq 0$ is a non-negative similarity kernel defined for all pairs of locations $1 \leq j, j' \leq M$, and $K^S(i, i') \geq 0$ is defined similarly for rows, $1 \leq i, i' \leq N$.

We can define $r_{ij}^L = \frac{1}{C_j} \sum_{j', j' \neq j} K^L(j, j') r_{ij'}$ where $C_j = \sum_{j, j' \neq j} K^L(j, j')$ is a normalization constant and the r_{ij}^L values can be viewed as smoothed pseudocounts for a particular location (column) j, obtained from a sum of weighted counts from similar locations (similar columns), with the weights being a function of location-to-location similarity. From these pseudocounts we can estimate a predictive distribution for each individual i and location j as $\theta_{ij}^L = \frac{r_{ij}^L}{\sum_j r_{ij}^L}$.

In the results in this paper, for location (column) smoothing, we use the spatial distance between two locations as an indicator of similarity. The (j, j')th element in $K^L(j, j')$ is

[1]We also experimented with models where weights π_{ik} were learned for each individual, but did not find that they provided any advantage in prediction in our experiments.

proportional to the probability density function of a univariate Gaussian model of the distance (in kilometers) between j and j' (or equal to 0 for $j = j'$) using a fixed bandwidth σ (we used $\sigma = 2km$ in the results below).

***S*: Social (Row) Smoothing:** $r_{ij}^S = \frac{1}{C_i} \sum_{i' , i' \neq j} K^S(i, i') r_{i'j}$, where $C_i = \sum_{i, i' \neq i} K^S(i, i')$ is a normalization constant and the r_{ij}^S values can be viewed as smoothed pseudocounts for a particular individual (row) i, obtained from a sum of weighted counts from similar individuals (rows), with the weights being a function of individual-to-individual similarity. From these pseudocounts we can again estimate a predictive distribution for each individual i and location j as $\theta_{ij}^S = \frac{r_{ij}^S}{\sum_j r_{ij}^S}$. The social (row) smoothing is estimated using the social graph. Each similarity $K^S(i, i')$ is proportional to the cosine similarity between the two binary vectors that represent the individuals' social ties.

3.3 Learning Individual Models

In our learning setup we have three disjoint sets of data: Training data D_{tr} for estimating the components in the mixture model (as described above), validation data D_v for estimating the mixing weights π (described below) and Test data D_{te}.

In the first step of learning the training data D_{tr} is used to compute the four component distributions $\boldsymbol{\theta}_i^I, \boldsymbol{\theta}^P, \boldsymbol{\theta}_i^M, \boldsymbol{\theta}_i^S$ as defined earlier. Three of the four components are individual-specific (the individual I, location M, and social S components) and the other (population P) is shared across all individuals.

In the second step, the validation data D_v is used to estimate the mixture weights $\boldsymbol{\pi}$, given the known components estimated from D_{tr}. We use the Expectation-Maximization (EM) algorithm for this purpose, and a Dirichlet prior with parameters $\beta \cdot m_0$ over the weights, where $m_0 = \frac{1}{4}$ corresponds to a flat uniform prior on the weights and β is a scalar hyperparameter that controls the prior strength. We learn β by using simple gradient ascent steps at the end of each iteration in the EM algorithm. By learning the mixture weights on validation data we are in effect learning the weights that are best in a predictive sense, i.e., combining the different smoothing components in a manner that best generalizes to new data. It is important that the mixture weights are estimated using a data set D_v that is separate from the data D_{tr} used to estimate the components. Otherwise EM will overfit and put all the weight on the individual component $\boldsymbol{\theta}_i^I$ that "memorizes" D_{tr} and will not learn to smooth appropriately.

4. EXPERIMENTS AND RESULTS

We consider different data sets from Twitter and Gowalla in our experiments. The Twitter data set contains tweets from two different regions: Orange County (California), and New York referred to as `TwOC, TwNY`. The tweets were gathered using the Twitter API[2] from May 2015 to February 2016, selecting tweets that have geolocation (GPS) coordinates for each tweet. The Gowalla data set[3]covers San Francisco, and New York, referred as `GoSF, GoNY`. For the experimental results reported later in the paper we retained individuals with information from at least 5 unique days and locations with 3 events or more.

[2]https://dev.twitter.com/streaming/

[3]https://snap.stanford.edu/data/loc-gowalla.html

	Events	Individuals	Locations	Days
`TwOC`	449306	13559	11347	240
`TwNY`	690801	30320	11260	240
`GoSF`	138934	2593	7706	507
`GoNY`	77761	1831	7692	441

Table 1: Number of unique events, individuals, locations and days for the different data sets.

In location-based modeling the GPS coordinates of events are typically categorized into a "vocabulary" of M locations. For the Gowalla dataset the vocabulary is pre-defined by a set of known high volume locations such as businesses and public venues (e.g., [2, 12]). For the Twitter data, we defined our dictionary using reverse geocoding based on publicly-available geoparcels. Geoparcels represents a set of disjoint polygons on a map and represent a specific property or lot (such as a house, a stadium, a store/shop, an airport etc.). The polygons vary in size and shape depending on the function of the property (e.g., typically smaller polygons for houses, much larger polygons for sport stadiums and theme parks). Statistics about the number of users and locations are presented in Table 1.

We compared the performance of our adaptive mixtures with well-known general-purpose baselines as well as with several state-of-the-art recently-published models for location recommendation. We use MPE and MLE variants of the simple multinomial model (as defined in Section 3.1), as well as non-negative matrix factorization (NMF) and hierarchical Bayesian probabilistic matrix factorization [4] (BPF) as our simple baselines. We also compared to well-known methods that incorporate the geographical preferences of an individual in order to increase the level of individual-level personalization: Fused [1], iGSLR [10] and RGFM [6]. We compared all of these methods to our proposed 4-component adaptive mixture model (which we will refer to as **AM4**), with individual, population, location, and social components.

To evaluate the models we use 60% of the data used for training, 20% for validation and the other 20% for test. We then perform evaluation using the test data using Average Percentile Rank (APR) and Log-Probability scores. To compute percentile rank, we sort the locations for each individual using the model's probability for each location. We then find the position (or rank) r for the actual locations that the user visited in the test data. Results below are reported on a scale from 0 to 100 with 0 being best (ranked at the top of the list). The second metric we use is the log-probability (for each individual) for location events in test set for that individual. Both metrics are averaged over all the test location events for an individual and then averaged over all individuals to produce a single score.

We computed the APR test metric using all of the models described above and the average log-probability test metric for models that produce probability distributions as predictions. Tables 2 and 3 shows the results for the different models (rows) for the out-of-sample evaluation. The column shows the results across the 4 different data sets as well as a 5th column that shows the average for each model across the data sets.

	TwOC	TwNY	GoSF	GoNY	Average
MLE	20.08	29.89	33.07	34.89	29.48
MPE	6.40	7.52	15.18	22.33	12.85
AM4	**4.46**	**6.61**	**11.83**	**15.93**	**9.71**
NMF	12.93	13.06	17.91	24.13	17.01
BPF	8.59	9.62	17.01	24.12	14.83
iGSLR	39.50	40.28	38.52	44.53	40.71
RGFM	8.64	8.78	15.85	24.30	14.39
Fused	8.08	9.61	16.96	24.10	14.69

Table 2: **Average Percentile Rank (APR)** on test data across individuals, for each data set and averaged across data sets. Lower score is better.

	TwOC	TwNY	GoSF	GoNY	Average
MPE	-4.50	-5.93	-7.21	-7.43	-6.27
AM4	**-4.45**	**-5.79**	**-6.66**	**-7.01**	**-5.98**
BPF	-6.33	-7.11	-7.53	-8.02	-7.25
Fused	-6.31	-7.10	-7.54	-8.02	-7.24

Table 3: **Average Log-Probability** across individuals, for each data set and averaged across data sets. Lower score is better.

A number of trends emerge from the data in Tables 2 and 3. The simple MPE predictor outperforms all of the baselines (including NMF, BPF, iGSLR, RGFM, and Fused) on all 4 data sets and for both metrics, confirming that including individual-level detail is important for these types of applications. The proposed AM4 method is systematically better than MPE and better than all of the other baselines, across all 4 data sets and under both metrics. This clearly illustrates the superiority of the proposed mixture-based smoothing approach relative to smoothing towards the population (MPE) or using matrix factorization.

In an additional series of experiments we also evaluated all of the models in terms of their predictions only on test data locations that were new (unseen in their historical data), e.g., for point-of-interest recommendation systems. When compared to the other approaches we found that the AM4 method was again systematically the most accurate method overall in terms of APR metric and was among the most accurate in terms of the log-probability metric. The detailed results are omitted because of space constrains.

5. CONCLUSIONS

In this paper we described a general framework for smoothing of individual-level categorical distributions using mixture models and we applied this framework to the problem of learning location models from historical data of individual-location counts. Experimental results on two data sets from Twitter and Gowalla indicate that the proposed mixture approach can lead to significantly improved predictive performance compared to non-mixture approaches. A key aspect of the method is its ability to weight different sources of information and to learn how to generalize beyond observations in the training data. The framework is relatively simple to implement and can be implemented in a computationally efficient manner. Various extensions and generalizations are possible, e.g., conditioning the prediction model on time of day and day of week and extending to continuous-space models such as kernel density representations.

Acknowledgements

The work in this paper was supported in part by the National Science Foundation under award IIS-1320527 and by a Google Faculty Award.

6. REFERENCES

[1] C. Cheng, H. Yang, I. King, and M. R. Lyu. Fused matrix factorization with geographical and social influence in location-based social networks. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[2] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD Conference*, pages 1082–1090. ACM, 2011.

[3] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*, volume 2. Taylor & Francis, 2014.

[4] P. Gopalan, J. M. Hofman, and D. M. Blei. Scalable recommendation with Poisson factorization. In *Proceedings of the Uncertainty in AI Conference*, pages 326–335, 2015.

[5] R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th International Conference on Machine Learning*, pages 315–322, 2002.

[6] X. Li, G. Cong, X.-L. Li, T.-A. N. Pham, and S. Krishnaswamy. Rank-GeoFM: a ranking based geographical factorization method for point of interest recommendation. In *Proceedings of the 38th ACM SIGIR Conference*, pages 433–442. ACM, 2015.

[7] D. Lian, C. Zhao, X. Xie, G. Sun, E. Chen, and Y. Rui. GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation. In *Proceedings of the 20th ACM SIGKDD Conference*, pages 831–840. ACM, 2014.

[8] B. Liu, Y. Fu, Z. Yao, and H. Xiong. Learning geographical preferences for point-of-interest recommendation. In *Proceedings of the 19th ACM SIGKDD Conference*, pages 1043–1051. ACM, 2013.

[9] B. Liu, H. Xiong, S. Papadimitriou, Y. Fu, and Z. Yao. A general geographical probabilistic factor model for point of interest recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 27(5):1167–1179, 2015.

[10] J.-D. Zhang and C.-Y. Chow. iGSLR: personalized geo-social location recommendation: a kernel density estimation approach. In *Proceedings of the 21st ACM SIGSPATIAL Conference*, pages 334–343. ACM, 2013.

[11] J.-D. Zhang and C.-Y. Chow. GeoSoCa: Exploiting geographical, social and categorical correlations for point-of-interest recommendations. In *Proceedings of the 38th ACM SIGIR Conference*, pages 443–452. ACM, 2015.

[12] J.-D. Zhang, C.-Y. Chow, and Y. Li. LORE: exploiting sequential influence for location recommendations. In *Proceedings of the 22nd ACM SIGSPATIAL Conference*, pages 103–112. ACM, 2014.