SDS 385 Exercises 9: Matrix factorization

Xinying Hao

Department of Marketing, The University of Texas at Austin

Nov.11, 2016

1 Application to marketing

1.1 Latent Dirichlet Model

LDA is to model documents as arising from multiple topics, where a topic is defined to be a distribution over a fixed vocabulary of terms. The basic idea is that we can only observe the count of categories for each document (customer) and we want to know the hidden variables represent the latent topic structure, i.e. the topics themselves and how each document (customer) exhibits them (Blei & Lafferty, 2003). Now we assume that there are d=1…D documents (customers), n=1..N, words per document, and k= 1…K topics (segments).

1. ***Each document(customer) is a random mixture of topics(segments).*** As in Figure 1,$\theta_d \sim dir(\alpha)$, which confirms to a Dirichlet distribution with parameter α,θ_d is a vector of topic proportions. In other word, $\theta_d$indicate the probability of document d belonging to each topic.

2. ***Each category is drawn from one of topics.***$Z_{d,n} \sim multi(\theta_d)$, $Z_{d,n} \in \{1, ..., K\}$is the topic assignment for each category in each document. The estimate of $Z_{d,n}$ is a list, where each element of the list, is an integer vector indicating the topic assignment for each category.

3. ***We only observe the categories of each document (customer),***$W_{d,n} \sim multi(\beta_{z_{d,n}})$, which is the observed data.

4. ***So our goal is to infer the underlying topic structure,***$\beta_k \sim dir(\eta)$,the distribution of

Figure 1: Graph Representation of LDA

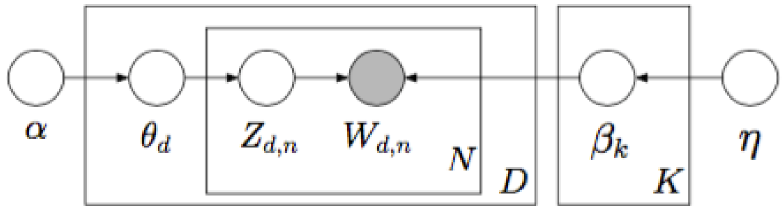
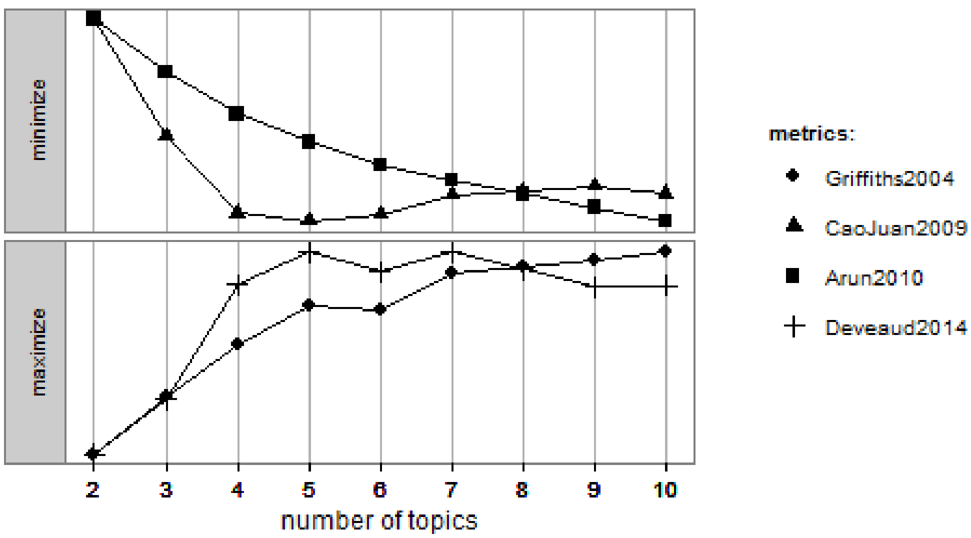


Figure 2: Find the Optimal Number of Topics



each topic over categories. The estimates of β_k are shown in Figure 2. Note that we have the estimates of the number of times a categories was assigned to a topic.

1.2 Sparse Matrix Factorization

- Transform the data using square root, i.e. $x_{ij}^{Transform} = \sqrt{x_{ij}}$

- $\lambda_u = 1.8$, $\lambda_v = 1.8$ (larger λ may result in large set of non-zero categories)

Figure 3: Market Segments Distribution over Categories

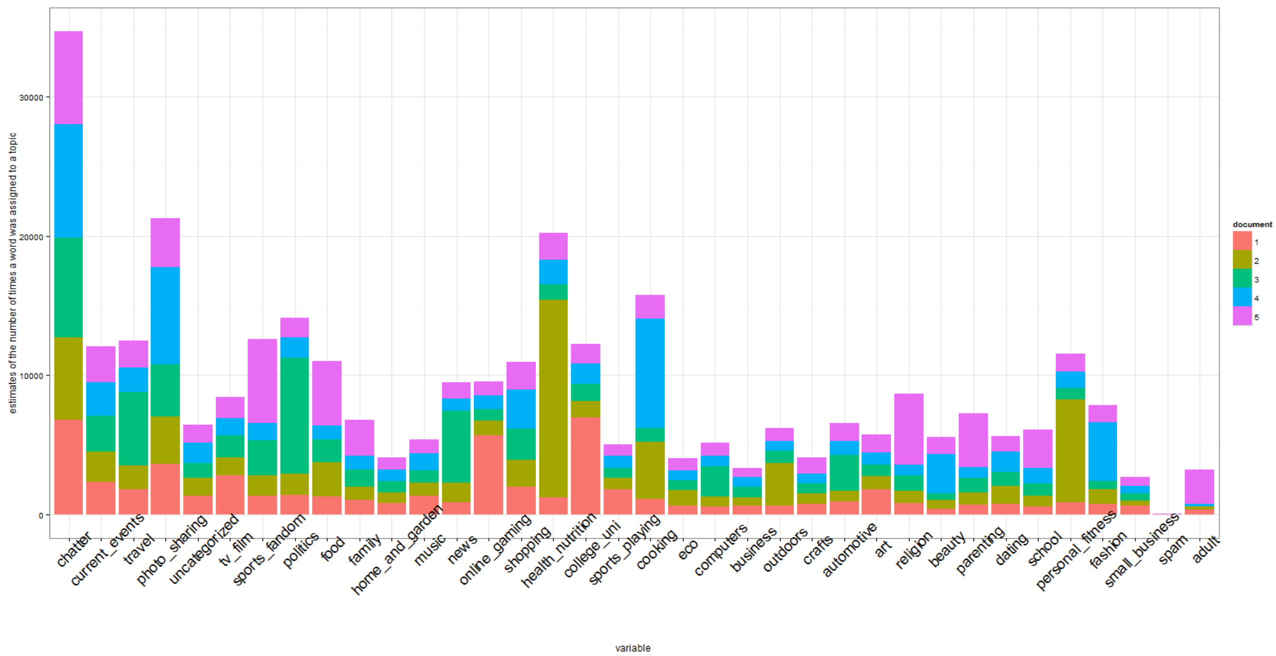


Table 1: Top Categories_LDA

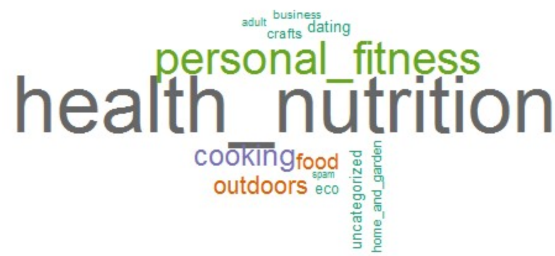
Segment1	**Segment2**	**Segment3**	**Segment4**	**Segment5**
religion	**health_nutrition**	politics	college_uni	cooking
adult	**personal_fitness**	news	online_gaming	fashion
sports_fandom	outdoors	travel	tv_film	photo_sharing
parenting	cooking	automotive	sports_playing	beauty
food	food	computers	art	chatter
school	adult	chatter	adult	shopping
family	eco	sports_fandom	music	dating
crafts	dating	current_events	chatter	uncategorized
beauty	spam	business	uncategorized	music
home_and_garden	uncategorized	tv_film	current_events	school

Table 2: Top Categories_Sparse Matrix Factorization

Factor1		**Factor2**		**Factor3**		**Factor4**		**Factor5**	
health_nutrition	0.87	chatter	0.68	politics	0.76	cooking	0.82	cooking	0.76
personal_fitness	0.34	health_nutrition	0.65	travel	0.52	photo_sharing	0.41	photo_sharing	0.44
chatter	0.28	politics	0.28	computers	0.34	beauty	0.31	fashion	0.37
outdoors	0.20	travel	0.19	news	0.18	fashion	0.26	politics	0.27
tv_film	0.04							travel	0.11
cooking	0.03							beauty	0.05
food	0.02								
photo_sharing	0.01								

Figure 4: Word Clouds. How to target the right audience with the right messages?

Segment1 & Segment2



Segment3 & Segment4

Segment5