Latent Rating Regression Model I

According to Hongning Wang (2010), Latent Rating Regression Model is to derive criterion ratings based on texts. Specifically, the overall rating is not directly determined by word frequency features, but instead, determined by different criterion ratings. Those criterion ratings are more directly derived by word frequency features.

For each aspect k, a feature matrix $\mathrm{F_k}$ is generated based on corresponding reviews. $\mathrm{F_k}$ is a $n \times \mathrm{d}$ matrix for n hotels and d words.

i—the ith hotel, i=1,2,3…n
j—the jth feature/word, j=1,2,3…d
k—the kth aspect, k=1,2,3,4,5

A criterion rating S_k is derived from the feature matrix $\mathrm{F_k}$ and the word sentiment polarities β_k on criterion k by the equation:

$$\mathrm{S}_k \sim \sum_{j=1}^{d} \beta_{kj} \mathrm{F}_{\,kj}$$

A criterion weight w_k, is known that $\sum \mathrm{w}_k = 1$. Besides, the multivariate Gaussian distribution is employed to capture the unpredictability and dependencies of w_k.

$$\mathrm{w} \sim Normal(\mu, \Sigma_1)$$

The overall rating of the dataset r is related to criterion weight w and criterion ratings S. Moreover, each r is sample from a Gaussian distribution:

$$r \sim Normal(\sum_{k=1}^{k} \mathrm{w}_k \mathrm{S}_k\,, \Sigma_2)$$

Latent Rating Regression Model II

A. Model Development

In this report, slightly differently from the model above, we focus on how to derive S_k. The estimated value S_k could be obtained by Maximum Likehood estimator, using VI algorithm.

In this report, we focus on how to derive S_k. Hence, we employ from previous model:

$$\mathrm{S}_k \sim Normal(\sum_{j=1}^{d} \beta_{kj} \mathrm{F}_{kj}, \Sigma_3)$$

We now consider five criteria, including Price, Room, Location, Cleanliness and Service. It could be presented digitally:

$$\mathrm{S} = \{\mathrm{S}_1, \mathrm{S}_2, \mathrm{S}_3, \mathrm{S}_4, \mathrm{S}_5\}$$

To simplify the problem, we use equal weights in this project, which may also make sense for the average hotel rating. Every commenter may have different criterion weights. And for each hotel, we just simplify it to be equal weighted. It satisfies the condition of dependencies between weights. We have:

$$\mathrm{w} = 0.2$$

B. Project Model Detail

Specifically, we employ a model for VI algorithm based on knowledge of mean distribution and variance distribution.

$$\mathrm{S}_k \sim Normal(\,\mathrm{F_k} \beta_k^T, \lambda^{-1})$$
$$\beta_k \sim Normal(\mathbf{0}, diag(\,\alpha_1,\, \alpha_2, \dots,\, \alpha_d)^{-1})$$
$$\alpha_j \sim \mathrm{Gamma}(\mathrm{a_0}, \mathrm{b_0})$$
$$\lambda \sim \mathrm{Gamma}(\mathrm{e_0}, \mathrm{f_0})$$

Where
F_k is a $n \times d$ matrix for n hotels and d words, showing the word feature of the whole hotel reviews;
β_k is a $1 \times d$ vector for d words, representing word sentiment polarities; α_j is parameter of β_k distribution;
λ is a $n \times n$ variance matrix;

With the model, we further derived:
$P(\beta_k, \lambda, \alpha_1, \alpha_2, \ldots, \alpha_d | F_k) \propto P(S_k, \beta_k, \lambda, \alpha_1, \alpha_2, \ldots, \alpha_d | F_k)$
$P(S_k, \beta_k, \lambda, \alpha_1, \alpha_2, \ldots, \alpha_d | F_k) = P(S_k | \beta_k, F_k, \lambda) \cdot P(\beta_k | \alpha_1, \alpha_2, \ldots, \alpha_d) \cdot P(\lambda) \cdot P(\alpha_1) \cdot P(\alpha_2) \cdot \ldots \cdot P(\alpha_d)$

Derive the optimal form of each q distribution and then obtain relative updating equations:

For q distribution of β_k:

$$q(\beta_k) = \mathrm{Normal}(\beta_k | \mu', \Sigma')$$
$$\Sigma' = \left(diag\left(\frac{a'}{b'_1}, \frac{a'}{b'_2}, \ldots, \frac{a'}{b'_d}\right) + \frac{e'}{f'} \sum_1^n F_{ki}^T F_{ki}\right)^{-1}$$
$$\mu' = \Sigma' \cdot \frac{e'}{f'} \sum_1^n r_i F_{ki}^T \quad (\text{assuming } S_{ki} = r_i \text{ for any aspect k})$$

For q distribution of λ:

$$q(\lambda) = \mathrm{Gamma}(\lambda | e', f')$$
$$e' = \frac{N}{2} + e_0$$
$$f' = 0.5 \sum_1^n ((r_i - F_{ki}^T \mu')^2 + F_{ki} \Sigma' F_{ki}^T) + f_0$$

For q distribution of α_j, j=1,2,3…d:

$$q(\alpha_j) = \mathrm{Gamma}(\alpha_j | a', b')$$
$$a' = \frac{1}{2} + a_0$$
$$b' = \frac{1}{2}(\mu'[j])^2 + \Sigma'(j, j) + b_0$$

Finally, we get the objective function

$$L_t = \frac{1}{2} \ln(|\Sigma'_t|) - e' \ln f'_t - a' \sum_{j=1}^{d} \ln(b'_{j,t}) + \mathrm{Const.}$$

In the end, when L_t converge, we get the corresponding β_k : β_k is equal to the mean in the q distribution of β_k; the steady value of μ' is the β_k we want; $\beta_k = \mu'_t$ in final iteration t , a $1 \times d$ vector. The criterion rating $S_k = F_k \beta_k^T$ and S_k is a $1 \times n$ vector.

Latent Rating Regression Model : Coding

1. Data processing (Python)

Input: millions of hotel reviews

Output : F_1, F_2, F_3, F_4, F_5 (F_k is a $n \times d$ matrix for n hotels and d words, each row sums to 1) and r (r is a $1 \times n$ vector, each $r_i = rate_i - avg\_rate$)

2. VI Algorithm (MatLab)

2.1 Initialize parameters $\Sigma', \mu', a_0, b_0, e_0$ and set

$$e' = \frac{N}{2} + e_0$$

$$a' = \frac{1}{2} + a_0$$

2.2 For iteration t=1, ……, T

-Update q(λ)

$$f'_t = \frac{1}{2}\sum_{1}^{N}[(R_i - F_i^T \mu'_{t-1})^2 + F_i^T \Sigma'_{t-1} F_i] + f_0$$

$$E[\lambda]_t = \frac{e'}{f'_t}$$

-Update $q(\alpha_k)$, for iteration k=1, …, d

$$b'^k_t = \frac{1}{2}(\mu'_{t-1}[k])^2 + \Sigma'_{t-1}(k,k) + b_0$$

$$E[\alpha_k]_t = \frac{a'}{b'^k_t}$$

-Update q(ω)

$$\Sigma'_t = \left(diag(E[\alpha_1]_t, E[\alpha_2]_t, \dots, E[\alpha_d]_t) + E[\lambda]_t \sum_{1}^{N} F_i F_i^T \right)^{-1}$$

$$\mu'_t = \Sigma'_t \cdot E[\lambda]_t \sum_{1}^{N} R_i F_i$$

-Evaluate L

$$L_t = \frac{1}{2}\ln(|\Sigma'_t|) - e' \ln f'_t - a' \sum_{k=1}^{d} \ln(b'_{k,t}) + \text{Const.}$$