Latent Rating Regression Derivation

Yufeng Ma

April 5, 2016

1 Latent Rating Regression Model

As r_d and α_d both satisfy the Gaussian distribution, we can have the following formula:

$$r_d \sim N(\sum_{i=1}^{k} \alpha_{di} \sum_{j=1}^{n} \beta_{ij} W_{dij}, \delta^2)$$

$$\alpha_d \sim N(\mu, \Sigma)$$

Combining them using the Bayes and partition rules, we can get

$$\begin{aligned} P(r|d) &= P(r_d|\mu, \Sigma, \delta^2, \beta, W_d) \\ &= \int p(\alpha_d|\mu, \Sigma) p(r_d| \sum_{i=1}^{k} \alpha_{di} \sum_{j=1}^{n} \beta_{dij} W_{dij}, \delta^2) d\alpha_d \end{aligned}$$

Here $\sum_{i=1}^{k} \alpha_{di} = 1$ is required for each document d. So in order to get rid of this constraint, auxiliary variables $\{\hat{\alpha}_{d1}, \hat{\alpha}_{d2}, \ldots, \hat{\alpha}_{dk}\}$ are introduced here, where

$$\alpha_{di} = \frac{e^{\hat{\alpha}_{di}}}{\sum_{j=1}^{k} e^{\hat{\alpha}_{dj}}}$$

Also to avoid negative aspect rating S_{di}, we can impose similar trick as follows:

$$S_{di} = e^{W_{di} \beta_i^T}$$

While for doing inference, we would prefer the one that maximizes a posteriori.

2 EM Updating

Now let's get the complete-data log-likelihood function for the newly derived problem.

$$\begin{aligned}
\mathcal{L}(D) &= \sum_{d \in D} \log p(r_d | \mu, \Sigma, \delta^2, \beta, W_d) \\
&= \sum_{d \in D} \left[\log p(\hat{\alpha}_d | \mu, \Sigma) + \log p(r_d | \sum_{i=1}^{k} \alpha_{di} \sum_{j=1}^{n} \beta_{dij} W_{dij}, \delta^2) \right] \\
&= \sum_{d \in D} \left[\log \frac{1}{\sqrt{(2\pi)^k |\Sigma|}} exp(-\frac{1}{2}(\hat{\alpha}_d - \mu)^T \Sigma^{-1} (\hat{\alpha}_d - \mu)) \right. \\
&\quad \left. + \log \frac{1}{\sqrt{2\pi\delta^2}} exp(-\frac{(r_d - \alpha_d^T S_d)^2}{2\delta^2}) \right]
\end{aligned}$$

If we skip the constant and multiply it by 2, then we can get

$$\mathcal{L}(D) = \sum_{d \in D} \left[-\log |\Sigma| - (\hat{\alpha}_d - \mu)^T \Sigma^{-1} (\hat{\alpha}_d - \mu) - \log \delta^2 - \frac{(r_d - \alpha_d^T S_d)^2}{\delta^2} \right]$$

In order to deal with the overfitting problem, we should also impose regularization on the global parameter β.

$$\mathcal{L}(D) = \sum_{d \in D} \left[-\log |\Sigma| - (\hat{\alpha}_d - \mu)^T \Sigma^{-1} (\hat{\alpha}_d - \mu) - \log \delta^2 - \frac{(r_d - \alpha_d^T S_d)^2}{\delta^2} \right] - \lambda \beta^T \beta$$

Also to ensure numerical stability and good initialization of random variables, an additional term is added into the log-likelihood function per each document d,

$$\begin{aligned}
\mathcal{L}(D) = \sum_{d \in D} \left[-\log |\Sigma| - (\hat{\alpha}_d - \mu)^T \Sigma^{-1} (\hat{\alpha}_d - \mu) - \log \delta^2 - \frac{(r_d - \alpha_d^T S_d)^2}{\delta^2} \right. \\
\left. -\gamma \sum_{i=1}^{k} \alpha_{di} (S_{di} - r_d)^2 \right] - \lambda \beta^T \beta
\end{aligned}$$

Therefore our goal becomes minimizing the following function and get the corresponding parameters:

$$\begin{aligned}
\hat{\Theta} = \arg\min_{\Theta} \sum_{d \in D} \left[\log |\Sigma| + (\hat{\alpha}_d - \mu)^T \Sigma^{-1} (\hat{\alpha}_d - \mu) \right. \\
\left. + \log \delta^2 + \frac{(r_d - \alpha_d^T S_d)^2}{\delta^2} + \gamma \sum_{i=1}^{k} \alpha_{di} (S_{di} - r_d)^2 \right] + \lambda \beta^T \beta
\end{aligned}$$

2.1 E-step Updating

2.1.1 Updating $\hat{\alpha}$

In the E-step, we will have to infer $\hat{\alpha}_d$ for every document d by minimizing the following objective function. As we are only interested in $\hat{\alpha}_d$ here, so unrelated parameters like Σ and δ can be ignored here.

$$L(d) = (\hat{\alpha}_d - \mu)^\intercal \Sigma^{-1} (\hat{\alpha}_d - \mu) + \frac{(r_d - \alpha_d^\intercal S_d)^2}{\delta^2} + \gamma \sum_{i=1}^{k} \alpha_{di} (S_{di} - r_d)^2$$

To apply the conjugate-gradient-interior-point method on the above function, we need the derivatives with repect to $\hat{\alpha}_d$

$$\frac{\partial L(d)}{\partial \alpha_{di}} = \frac{2(\alpha_d^\intercal S_d - r_d)}{\delta^2} \frac{\partial \alpha_d^\intercal S_d}{\partial \hat{\alpha}_{di}} + \gamma \frac{\partial \sum_{j=1}^{k} \alpha_{dj} (S_{dj} - r_d)^2}{\partial \hat{\alpha}_{di}} + \frac{\partial (\hat{\alpha}_d - \mu)^\intercal \Sigma^{-1} (\hat{\alpha}_d - \mu)}{\partial \hat{\alpha}_{di}}$$

Doing some derivation with the exponetial function, we can get

$$\frac{\partial \alpha_d^T S_d}{\partial \hat{\alpha}_{di}} = \alpha_{di} \sum_{j=1}^{k} [\tau(j = i) S_{dj} (1 - \alpha_{di}) - \tau(j \neq i) S_{dj} \alpha_{dj}]$$

$$\frac{\partial \sum_{j=1}^{k} \alpha_{dj} (S_{dj} - r_d)^2}{\partial \hat{\alpha}_{di}} = \alpha_{di} \sum_{j=1}^{k} \left[\tau(j = i)(S_{dj} - r_d)^2 (1 - \alpha_{di}) - \tau(j \neq i)(S_{dj} - r_d)^2 \alpha_{dj}\right]$$

$$\frac{\partial (\hat{\alpha}_d - \mu)^\intercal \Sigma^{-1} (\hat{\alpha}_d - \mu)}{\partial \hat{\partial}_{di}} = 2(\hat{\alpha}_d - \mu)\Sigma^{-1} \cdot I \cdot \frac{\partial \hat{\alpha}_d}{\partial \hat{\alpha}_{di}} = 2 \sum_{j=1}^{k} \Sigma_{ji}^{-1} (\hat{\alpha}_{dj} - \mu_j)$$

As we can get S_d pretty straightword by $\beta^\intercal W_d$, we are done with the E-step.

2.2 M-step Updating

Now while in the M-step, we need to get the paramenters which maximize the probability of observing all the $\hat{\alpha}_d$ and r_d.

2.2.1 Updating μ and Σ

First for maximizing the probability of observing all the $\hat{\alpha}_d$, we should minimize the following objective function,

$$\arg\min_{\mu, \Sigma} \sum_{d \in D} \left[\log |\Sigma| + (\hat{\alpha}_d - \mu)\Sigma^{-1} (\hat{\alpha}_d - \mu)\right]$$

which leads us to

$$\mu_{(t+1)} = \arg\min_{\mu} \sum_{d\in D} (\hat{\alpha}_d - \mu)\Sigma^{-1}(\hat{\alpha}_d - \mu) = \frac{1}{|D|}\sum_{d\in D}\hat{\alpha}_d$$

$$\begin{aligned}\Sigma_{(t+1)} =& \arg\min_{\Sigma} \sum_{d\in D} \left[(\hat{\alpha}_d - \mu)\Sigma^{-1}(\hat{\alpha}_d - \mu) + \log|\Sigma|\right] \\ =& \frac{1}{|D|}\sum_{d\in D}(\hat{\alpha}_d - \mu_{(t+1)})^\intercal(\hat{\alpha}_d - \mu_{(t+1)})\end{aligned}$$

2.2.2 Updating β and δ

Then we want to maximize the probability of observing all the r_d. With skipping the unrelated variables in the function, we have to minimize the following objective function

$$L(D) = \sum_{d\in D}\left[\log\delta^2 + \frac{(r_d - \alpha_d^T S_d)^2}{\delta^2} + \gamma\sum_{i=1}^{k}\alpha_{di}(S_{di} - r_d)^2\right] + \lambda\beta^\intercal\beta$$

So in terms of the parameter β, our objective function will be like

$$L(\beta) = \sum_{d\in D}\left[\frac{(r_d - \alpha_d^T S_d)^2}{\delta^2} + \gamma\sum_{j=1}^{k}\alpha_{dj}(S_{dj} - r_d)^2\right] + \lambda\beta^\intercal\beta$$

where the derivative with respect to β_i will be

$$\begin{aligned}\frac{\partial L(\beta)}{\partial\beta_i} =& \sum_{d\in D}\left[\frac{2(\alpha_d^\intercal S_d - r_d)}{\delta^2}\frac{\partial\alpha_d^\intercal S_d}{\partial\beta_i} + 2\gamma\alpha_{di}(S_{di} - r_d)\frac{\partial S_{di}}{\partial\beta_i}\right] + 2\lambda\beta_i \\ =& 2\sum_{d\in D}\alpha_{di}\left[\frac{(\alpha_d^\intercal S_d - r_d)}{\delta^2} + \gamma(S_{di} - r_d)\right]\frac{\partial S_{di}}{\partial\beta_i} + 2\lambda\beta_i\end{aligned}$$

Here $\frac{\partial S_{di}}{\partial\beta_i}$ can be get from $S_{di} = exp(W_{di}\beta_i^T)$, which leads us to

$$\frac{\partial S_{di}}{\partial\beta_i} = S_{di}W_{di}$$

While for getting the optimal parameter δ, we need to minimize the following objective function,

$$\delta^2_{(t+1)} = \arg\min_{\delta}\sum_{d\in D}\left[\log\delta^2 + \frac{(r_d - \alpha_d^\intercal S_d)^2}{\delta^2}\right] = \frac{1}{|D|}\sum_{d\in D}(r_d - \alpha_d^\intercal S_d)^2$$

3 Multivariate Gaussian Maximum Likelihood Estimator

Here we applied several times of the ML estimator of Gaussian Distribution, now let's have a detailed look at how this result can be derived.

$$L(D) = \sum_{d \in D} \left[\log |\Sigma| + (\hat{\alpha}_d - \mu)^\intercal \Sigma^{-1} (\hat{\alpha}_d - \mu) \right]$$

First for estimating the mean vector μ, we would take the partial derivative with respect to μ, and set the derivative to zero.

$$\begin{aligned} \frac{\partial L(D)}{\partial \mu} = \sum_{d \in D} \frac{\partial (\hat{\alpha}_d - \mu)^\intercal \Sigma^{-1} (\hat{\alpha}_d - \mu)}{\partial \mu} &= \sum_{d \in D} 2(\hat{\alpha}_d - \mu)^\intercal \Sigma^{-1} \frac{\partial (\hat{\alpha}_d - \mu)}{\partial \mu} \\ &= \sum_{d \in D} 2(\hat{\alpha}_d - \mu)^\intercal \Sigma^{-1} \cdot (-I) = 0 \end{aligned}$$

Multiplying the above formula with Σ, we can get

$$\sum_{d \in D} 2(\hat{\alpha}_d - \mu)^\intercal = 0 \Rightarrow \mu = \frac{1}{|D|} \sum_{d \in D} \hat{\alpha}_d$$

Now let's turn to take the derivative w.r.t. Σ^{-1}. Before that, there are several results we should know.

1. $tr(x^\intercal A x) = tr(A x x^\intercal)$
2. $\frac{\partial \log |A|}{\partial A} = A^{-\intercal}$
3. $\frac{\partial tr(AB)}{\partial A} = \frac{\partial tr(BA)}{\partial A} = B^\intercal$

$$\begin{aligned} \frac{\partial L(D)}{\partial \Sigma^{-1}} &= \sum_{d \in D} \left[\Sigma^{-\intercal} (-(\Sigma^{-1})^{-2}) + \frac{\partial tr((\hat{\alpha}_d - \mu)^\intercal \Sigma^{-1} (\hat{\alpha}_d - \mu))}{\partial \Sigma^{-1}} \right] \\ &= \sum_{d \in D} \left[-\Sigma + \frac{\partial tr(\Sigma^{-1} (\hat{\alpha}_d - \mu)) (\hat{\alpha}_d - \mu)^\intercal}{\partial \Sigma^{-1}} \right] \\ &= \sum_{d \in D} \left[-\Sigma + (\hat{\alpha}_d - \mu)) (\hat{\alpha}_d - \mu)^\intercal \right] = 0 \end{aligned}$$

$$\Longrightarrow \Sigma = \frac{1}{|D|} \sum_{d \in D} (\hat{\alpha}_d - \mu)) (\hat{\alpha}_d - \mu)^\intercal$$