Indian Institute of Technology, Kanpur
Probabilistic Machine Learning, 18–19

Instructors: Piyush Rai
Authors: Gurpreet Singh
Email: guggu@iitk.ac.in

SCRIBE

1

Approximate Inference and Sampling Methods

1. Locally Conjugate Models

Inference with multiple unknows (parameters or hyperparameters) becomes tricky, and usually have intractable posteriors as well. Special methods such as EM or MLE-II have to be used.

We can approximate these posteriors using the ideo of **local** or **conditional conjugacy**. Essentially, the routine is to update one parameter keeping others fixed (similar to EM).

1.1 Local Conjugacy

Often is the case that the overall posterior $\mathbb{P}\left[\,\mathbf{\Theta}\;\middle|\;\mathbf{X}\,\right] = \frac{\mathbb{P}\left[\,\mathbf{x}\;\middle|\;\mathbf{\Theta}\,\right]\mathbb{P}\left[\,\mathbf{\Theta}\,\right]}{\mathbb{P}\left[\,\mathbf{X}\,\right]}$ is intractable. We define the conditional probability of a parameter Θ_k as follows

$$\mathbb{P}\left[\,\Theta_k\;\middle|\;\mathbf{X}_k, \mathbf{\Theta}_{-k}\,\right] = \frac{\mathbb{P}\left[\,\mathbf{X}_k\;\middle|\;\Theta_k, \mathbf{\Theta}_{-k}\,\right]\mathbb{P}\left[\,\Theta_k\,\right]}{\int \mathbb{P}\left[\,\mathbf{X}_k\;\middle|\;\Theta_k, \mathbf{\Theta}_{-k}\,\right]\mathbb{P}\left[\,\Theta_k\,\right] d\Theta_k}$$

Note. Here, $\mathbf{\Theta}_k$ is the set of all parameters excluding Θ_k and $\mathbf{X}_k$ is the data that is dependent on Θ_k

Suppose the conditional posteriors (CP) admit **local conjugacy** *i.e.* the above term is tractable for all parameters. Such models are called locally conjugate models.

1.2 Bayesian Matrix Factorization

We assume the data to be modelled as a low rank matrix with some noise term. We need to factorize this $N \times M$ matrix ($\mathbf{R}$) into two matrices (interpreted as users and items) $\mathbf{U}$ and $\mathbf{V}$ of sizes $N \times K$ and $K \times M$ respectively such that $K \ll N, M$ and $\mathbf{R} \approx \mathbf{U} \times \mathbf{V}$

We assume the error term, and hence the likelihood to be gaussian

$$\begin{aligned} \mathbf{R} &= \mathbf{U}\mathbf{V} + \boldsymbol{\epsilon} \\ r_{ij} &= \mathbf{u}_i^T\mathbf{v}_j + \epsilon_{ij}\mathbb{P}\left[\,r_{ij}\,\right] \qquad\qquad &= r_{ij}\;\middle|\;\mathbf{u}_i^T\mathbf{v}_j, \beta^{-1} \end{aligned}$$

We assume Gaussian priors on the user and item latent features

$$\begin{aligned} \mathbb{P}\left[\,\mathbf{u}_i\,\right] &= \mathcal{N}\left(\mathbf{u}_i\;\middle|\;\mathbf{0}, \lambda_u^{-1}\mathbf{I}_K\right) \\ \mathbb{P}\left[\,\mathbf{v}_j\,\right] &= \mathcal{N}\left(\mathbf{v}_j\;\middle|\;\mathbf{0}, \lambda_v^{-1}\mathbf{I}_K\right) \end{aligned}$$

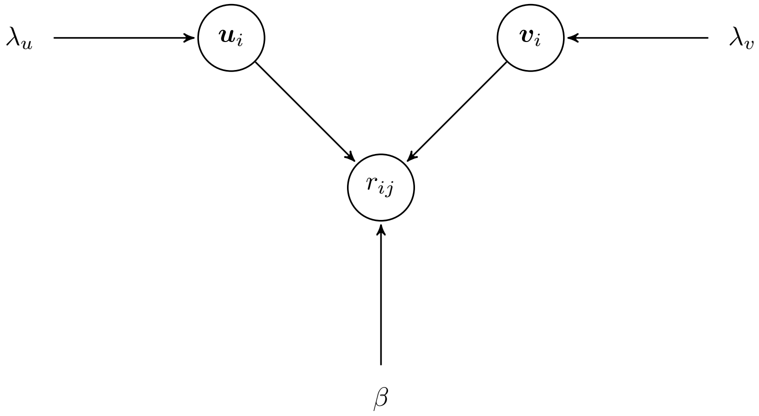

Figure 1: Matrix Factorization as Latent Factor Model

The BMF model with Gaussian likelihood and Gaussian priors has local conjugacy. We can write the posteriors for the latent variables as follows

$$\begin{aligned}
\mathbb{P}\left[\mathbf{u}_i \,\middle|\, \mathbf{R}, \mathbf{V}, \mathbf{U}_{-i}\right] &\propto \prod_{j:r_{ij}\neq 0} \mathbb{P}\left[r_{ij} \,\middle|\, \mathbf{u}_i, \mathbf{v}_j\right] \mathbb{P}\left[\mathbf{u}_i\right] \\
\mathbb{P}\left[\mathbf{u}_j \,\middle|\, \mathbf{R}, \mathbf{V}, \mathbf{U}_{-j}\right] &\propto \prod_{i:r_{ij}\neq 0} \mathbb{P}\left[r_{ij} \,\middle|\, \mathbf{u}_i, \mathbf{v}_j\right] \mathbb{P}\left[\mathbf{v}_j\right]
\end{aligned}$$

This is very similar to bayesian linear regression and we can find compute the posterior which is of the form

For user latent variables

$$\begin{aligned}
\mathbb{P}\left[\mathbf{u}_i \,\middle|\, \mathbf{R}, \mathbf{V}\right] &= \mathcal{N}\left(\mathbf{u}_i \,\middle|\, \boldsymbol{\mu}_{\mathbf{u}_i}, \boldsymbol{\Sigma}_{\mathbf{u}_i}\right) \\
\boldsymbol{\Sigma}_{\mathbf{u}_i} &= \left(\lambda_u \mathbf{I}_K + \beta \sum_{j:r_{ij}\neq 0} \mathbf{v}_j \mathbf{v}_j^T\right)^{-1} \\
\boldsymbol{\mu}_{\mathbf{u}_i} &= \boldsymbol{\Sigma}_{\mathbf{u}_i}\left(\beta \sum_{j:r_{ij}\neq 0} r_{ij}\mathbf{v}_j\right)
\end{aligned}$$

For item latent variables

$$\begin{aligned}
\mathbb{P}\left[\mathbf{v}_j \,\middle|\, \mathbf{R}, \mathbf{U}\right] &= \mathcal{N}\left(\mathbf{v}_j \,\middle|\, \boldsymbol{\mu}_{\mathbf{v}_j}, \boldsymbol{\Sigma}_{\mathbf{v}_j}\right) \\
\boldsymbol{\Sigma}_{\mathbf{v}_j} &= \left(\lambda_v \mathbf{I}_K + \beta \sum_{i:r_{ij}\neq 0} \mathbf{u}_i \mathbf{u}_i^T\right)^{-1} \\
i\boldsymbol{\mu}_{\mathbf{u}_i} &= \boldsymbol{\Sigma}_{\mathbf{v}_j}\left(\beta \sum_{i:r_{ij}\neq 0} r_{ij}\mathbf{u}_i\right)
\end{aligned}$$

This idea is similar in spirit to alternating optimization methods (such as EM).

2. Sampling from Distributions

We can approximate a distribution using a set of randomly drawn samples from it. However, it is not possible to directly sample *difficult* distributions. Hence, we need sampling techniques that allow us to sample from such distributions *e.g.* we sometimes might need to approximately infer / sample intractable posteriors. These samples can be used to approximate anything that depends on these distributions *e.g.* expectations or posteriors.

2.0.1 Emperical Distribution

We sample using L points $\left\{\mathbf{z}^{(l)}\right\}_{l=1}^{L}$ and the emperical distribution is defined as

$$\underset{L}{\mathbb{P}}\left[\,L\,\right] \quad = \quad \sum_{l=1}^{L} w_l \, \delta_{\mathbf{z}^{(l)}}\left(A\right)$$

where δ is the Dirac Delta function

$$\delta_{\mathbf{z}}\left(A\right) \quad = \quad \begin{cases} 0 & \text{if } \mathbf{z} \in A \\ 1 & \text{if } \mathbf{z} \notin A \end{cases}$$

w_l is the weight of a point. This distribution can be viewed as a histogram and the weight as the height of the histogram bar. This method can be used to sample from both simple as well as difficult distributions.

2.1 Transformation Methods

Transformation methods are used to sample (mostly) generic or standard distributions, using simple distributions which can be easily sampled (such as uniform distribution). We essentially transform

2.1.1 Inverse CDF Method

This method uses the change of variable rule to transform a simple distribution to more complex distributions. Suppose we want to sample the random variable z using a simpler distribution x, then we can write

$$\mathbb{P}\left[\,z\,\right] \quad = \quad \mathbb{P}\left[\,x\,\right]\left|\frac{\delta\left(x\right)}{\delta\left(z\right)}\right|$$

if x is a uniform distribution

$$x \quad = \quad \int_{-\infty}^{\hat{z}} \mathbb{P}\left[\,z\,\right] dz = \, F\left(\hat{z}\right)$$

Thus, we can first draw a random sample x from $Unif\left(0,1\right)$ and transform into z using $\hat{z} = h^{-1}\left(u\right)$

2.1.2 Box-Muller Method

This is a transformation technique to generate samples from two-dimensional Gaussian Distribution. The transformation is from the distribution $Unif^2\left(0,1\right)$ to the desired one.

Algorithm 1: Box-Muller Transformation

1. Assume two samples u, v drawn from $U \sim Unif\left(0,1\right)$ and $V \sim Unif\left(0,1\right)$
2. Transform these random variables into R, Θ such that $R = \sqrt{-2\log\left(U\right)}$ and $\Theta = 2\pi V$
3. Since Θ is a linear transformation, we know $\Theta \sim Unif\left(0, 2\pi\right)$
4. For R

$$\begin{aligned} \mathbb{P}\left[\,R \leq r\,\right] &= \mathbb{P}\left[\,\sqrt{-2\log\left(U\right)} \leq r\,\right] \\ &= 1 - \mathbb{P}\left[\,U < e^{-\frac{r^2}{2}}\,\right] \\ &= 1 - e^{-\frac{r^2}{2}} \\ \\ \implies f_R\left(r\right) &= re^{-\frac{r^2}{2}} \end{aligned}$$

5. Since U and V are independent, R and Θ will also be independent. Hence $f_{R,\Theta}\left(r,\theta\right) = \frac{1}{2\pi} re^{-\frac{r^2}{2}}$
6. We further transform these into two random variables Z_1 and Z_2

$$Z_1 = R\cos 2\pi\Theta \qquad \text{and} \qquad Z_2 = R\sin 2\pi\Theta$$

7. The join distribution of Z_1, Z_2 is given by

$$\begin{aligned} f_{Z_1,Z_2}\left(z_1,z_2\right)\frac{\delta\left(z_1,z_2\right)}{\delta\left(r,\theta\right)} &= f_{R,\Theta}\left(r,\theta\right) \\ &= \left[\frac{1}{\sqrt{2\pi}}e^{-\frac{z_1^2}{2}}\right]\left[\frac{1}{\sqrt{2\pi}}e^{-\frac{z_2^2}{2}}\right] \\ &= \mathcal{N}\left(z_1, z_2 \,\middle|\, \mathbf{0}, \mathbf{I}_2\right) \end{aligned}$$

It is also possible to transform Normal distrubition to any other gaussian distribution. Suppose a random variable $X \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_k\right)$. Transform this into $Z \sim \boldsymbol{\mu}, \boldsymbol{\Sigma}$ as follows

$$Z \quad = \quad \boldsymbol{\mu} + LX$$

where L is the Cholesky decomposition of $\boldsymbol{\Sigma}$

Note. Cholesky decomposition of $\boldsymbol{\Sigma}$ will always exist as it is a positive definite matrix

2.2 Rejection Sampling

Suppose we have a random variable with an intractable normalization factor, we can still sample it using rejection sampling. Consider a distribution $p(z)$ such that

$$\mathbb{P}\left[\, z \,\right] \quad = \quad \frac{\widetilde{\mathbb{P}}\left[\, z \,\right]}{Z_p}$$

where Z_p is not computable.

In order to sample, we need a *proposal distribution* $q\left(z\right)$ such that

$$M\, q\left(z\right) > \widetilde{\mathbb{P}}\left[\, z \,\right] \qquad\qquad \forall\, z \quad \text{where M is a constant}$$

The steps to generate sample are as follows

```
Algorithm 2: Rejection Sampling

1. Sample a random variable $z_*$ from $q\left(z\right)$
2. Sample a uniform random variable $u \sim Unif\left(0, M\, q\left(z_*\right)\right)$
3. If $u \leq \widetilde{\mathbb{P}}\left[\, z_* \,\right]$, then accept else reject
```

The proof of the algorithm can be seen as follows

$$\begin{aligned}
\mathbb{P}\left[\text{accept } \,\middle|\, z\right] &= \frac{\widetilde{\mathbb{P}}\left[\, z \,\right]}{M\, q\left(z\right)} \\
\mathbb{P}\left[\, z, \text{accept } \right] &= q\left(z\right) \mathbb{P}\left[\text{accept } \,\middle|\, z\right] && = \frac{\widetilde{\mathbb{P}}\left[\, z \,\right]}{M} \\
\mathbb{P}\left[\text{accept } \right] &= \int \frac{\widetilde{\mathbb{P}}\left[\, z \,\right]}{M} dz && = \frac{Z_p}{M} \\
\mathbb{P}\left[\, z \,\middle|\, \text{accept } \right] &= \frac{\widetilde{\mathbb{P}}\left[\, z \,\right]}{Z_p} && = \mathbb{P}\left[\, z \,\right]
\end{aligned}$$

3. Expections via Monte Carlo Sampling

$\mathcal{Z}$

<++>