Note on Minka "Estimating a Dirichlet distribution"

Tomoya Sasaki

September 12, 2018

1 Estimating a Dirichlet distribution

1.1 Basic setting

Let $\mathbf{p}$ denote a random vector each of which elements sum to 1 and $\boldsymbol{\alpha}$ a parameter vector of the Dirichlet distribution.

$$P(\mathbf{p}|\boldsymbol{\alpha}) \sim \mathcal{D}(\alpha_1, \cdots, \alpha_K) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} p_k^{\alpha_k - 1} \tag{1}$$

$$\text{where} \quad \sum_{k=1}^{n} p_k = 1 \quad \text{and} \quad p_k > 0 \tag{2}$$

We maximize $p(D|\boldsymbol{\alpha}) = \prod_{i=1}^{N} p(\mathbf{p}_i|\boldsymbol{\alpha})$ where $D = \{\mathbf{p}_1, \cdots, \mathbf{p}_N\}$.

$$\begin{aligned}
\log p(D|\boldsymbol{\alpha}) =& \log\left(\prod_{i=1}^{N} p(\mathbf{p}_i|\boldsymbol{\alpha})\right) & (3)\\
=& \sum_{i=1}^{N} \log\left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} p_{ik}^{\alpha_k - 1}\right) & (4)\\
=& \sum_{i=1}^{N} \left\{\log\left(\Gamma(\sum_{k=1}^{K} \alpha_k)\right) - \log\left(\prod_{k=1}^{K} \Gamma(\alpha_k)\right) + \log\left(\prod_{k=1}^{K} p_{ik}^{\alpha_k - 1}\right)\right\} & (5)\\
=& N \log\left(\Gamma(\sum_{k=1}^{K} \alpha_k)\right) - N \sum_{k=1}^{K} \log\left(\Gamma(\alpha_k)\right) + \sum_{i=1}^{N} \sum_{k=1}^{K} (\alpha_k - 1) \log p_{ik} & (6)\\
=& N \log\left(\Gamma(\sum_{k=1}^{K} \alpha_k)\right) - N \sum_{k=1}^{K} \log\left(\Gamma(\alpha_k)\right) + N \sum_{k=1}^{K} (\alpha_k - 1) \log \bar{p_k} & (7)
\end{aligned}$$

$$\text{where} \quad \log \bar{p_k} = \frac{1}{N}\sum_{i=1}^{N} \log p_{ik} \tag{8}$$

The objective function $p(D|\boldsymbol{\alpha})$ is convex in $\boldsymbol{\alpha}$ since the Dirichlet distribution is in the exponential family. This implies that the likelihood function is unimodal and the maximum can be found by a simple search.

1.2 Estimate α by MLE (1): A fixed-point iteration

The gradient of the log-likelihood with respect to each α_k can be written as follows.

$$g_k = \frac{\partial \log p(D|\boldsymbol{\alpha})}{\partial \alpha_k} \tag{9}$$

$$= N\frac{\partial}{\partial \alpha_k} \log\left(\Gamma(\sum_{k=1}^{K} \alpha_k)\right) - N\frac{\partial}{\partial \alpha_k}\sum_{k=1}^{K} \log\left(\Gamma(\alpha_k)\right) + N\frac{\partial}{\partial \alpha_k}\sum_{k=1}^{K}(\alpha_k - 1)\log \bar{p_k} \tag{10}$$

$$= N\left\{\Psi(\sum_{k=1}^{K}\alpha_k) - \Psi(\alpha_k) + \log \bar{p_k}\right\} \tag{11}$$

$$\text{where} \quad \Psi(x) = \frac{\partial \log \Gamma(x)}{\partial x} \tag{12}$$

Recall that $\frac{\partial}{\partial x_i}\sum_{i=1}^{N} x_i = x_i$ since all factors in $\sum_{i=1}^{N} x_i$ other than x_i disappear when you take derivative with respect to x_i.

We want to show $\Psi(\alpha_k^{new}) = \Psi(\sum_{k=1}^{K}\alpha_k^{old}) + \log \bar{p_k}$. Using the inequality in Appendix A in Minka's techinical report to the first factor of (7), we obtain following inequality. Note that here we regard α_k^{old} constant here.

$$\frac{\log p(D|\boldsymbol{\alpha})}{N} \tag{13}$$

$$= \log\left[\Gamma\left(\sum_{k=1}^{K}\alpha_k^{old}\right)\exp\left\{\sum_{k=1}^{K}(\alpha_k - \alpha_k^{old})\Psi\left(\sum_{k=1}^{K}\alpha_k^{old}\right)\right\}\right] - \sum_{k=1}^{K}\log\Gamma(\alpha_k) + \sum_{k=1}^{K}(\alpha_k - 1)\log \bar{p_k} \tag{14}$$

$$= \left(\sum_{k=1}^{K}\alpha_k\right)\Psi\left(\sum_{k=1}^{K}\alpha_k^{old}\right) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) + \sum_{k=1}^{K}(\alpha_k - 1)\log \bar{p_k} + \mathrm{const}(\alpha_k^{old}) \tag{15}$$

By taking derivative respect to α_k and set to zero to obtain the equation above.

$$\frac{\partial}{\partial \alpha_k}\left[\left(\sum_{k=1}^{K}\alpha_k\right)\Psi\left(\sum_{k=1}^{K}\alpha_k^{old}\right) - \sum_{k=1}^{K}\log\Gamma(\alpha_k) + \sum_{k=1}^{K}(\alpha_k - 1)\log \bar{p_k} + \mathrm{const}(\alpha_k^{old})\right] \tag{16}$$

$$= \Psi\left(\sum_{k=1}^{K}\alpha_k^{old}\right) - \Psi(\alpha_k) + \log \bar{p_k} = 0 \tag{17}$$

$$\therefore \quad \Psi(\alpha_k^{new}) = \Psi(\sum_{k=1}^{K}\alpha_k^{old}) + \log \bar{p_k} \tag{18}$$

$$\Leftrightarrow \alpha_k^{new} = \Psi^{-1}\left(\Psi(\sum_{k=1}^{K}\alpha_k^{old}) + \log \bar{p_k}\right) \tag{19}$$

1.3 Estimate α by MLE (2): Newton iteration

Newton itertion is a method to obtain x which satisfies $f(x) = 0$. The second order of Taylor expansion of $f(x)$ at x^* is

$$f(x) = f(x^*) + \frac{\partial f(x^*)}{\partial x}(x - x^*) + O((x - x^*)^2). \tag{20}$$

Ignoring the second order degree and higher, and inputing $f(x) = 0$, we obtain following equation, which is the update equation for Newton iteration method.

$$0 = f(x^*) + \frac{\partial f(x^*)}{\partial x}(x - x^*) \tag{21}$$

$$\Leftrightarrow x = x^* - \left(\frac{\partial f(x^*)}{\partial x}\right)^{-1} f(x^*) \tag{22}$$

In the following procedure $f(x)$ is equivalent to $\frac{\partial \log p(D|\boldsymbol{\alpha})}{\partial \alpha}$.

First, we take the second-derivatives(Hessian matrix) of the loglikelihood (which is equivalent to $\frac{\partial f(x)}{\partial x}$) since our target function is $\frac{\partial \log p(D|\boldsymbol{\alpha})}{\partial \alpha}$ and we want to obtain α with $\frac{\partial \log p(D|\boldsymbol{\alpha})}{\partial \alpha} = 0$.

$$\frac{\partial^2 \log p(D|\boldsymbol{\alpha})}{\partial^2 \alpha_k} = N\left\{\Psi'\left(\sum_{k=1}^{K}\alpha_k\right) - \Psi'(\alpha_k)\right\} \tag{23}$$

$$\frac{\partial^2 \log p(D|\boldsymbol{\alpha})}{\partial \alpha_k \partial \alpha_j} = N\Psi'\left(\sum_{k=1}^{K}\alpha_k\right) \quad \text{where } k \neq j \tag{24}$$

or

$$\frac{\partial^2 \log p(D|\boldsymbol{\alpha})}{\partial \alpha_k \partial \alpha_j} = \underbrace{N\Psi'\left(\sum_{k=1}^{K}\alpha_k\right)}_{(a)} \underbrace{-N\Psi'(\alpha_k)\,\delta(j-k)}_{(b)} \tag{25}$$

Note that Ψ' is known as the trigamma function and δ as the indicator function. We define each component of the Hessian as $h_{kj} = \frac{\partial^2 \log p(D|\boldsymbol{\alpha})}{\partial \alpha_k \partial \alpha_j}$ and $g_k = \frac{\partial \log p(D|\boldsymbol{\alpha})}{\partial \alpha_k}$. The Hessian can be written in matrix form as follows.

$$\mathbf{H} = \mathbf{Q} + \mathbf{1}\mathbf{1}^\top z \tag{26}$$

$$q_{jk} = -N\Psi'(\alpha_k)\delta(j-k) \qquad (b) \text{ in the equation (25)} \tag{27}$$

$$z = N\Psi'(\sum_{k=1}^{K}\alpha_k) \qquad (a) \text{ in the equation (25)} \tag{28}$$

Using the equation (22),

$$\boldsymbol{\alpha}^{new} = \boldsymbol{\alpha}^{old} - \mathbf{H}^{-1}\mathbf{g} \tag{29}$$

We examine second part in the right side of the equation (29). Note that $\mathbf{H}$ and $\mathbf{g}$ are function of α^{old}.

$$\mathbf{H}^{-1}\mathbf{g} = (\mathbf{Q} + \mathbf{1}\mathbf{1}^\top z)^{-1}\mathbf{g} \tag{30}$$

$$= \mathbf{Q}^{-1}\mathbf{g} - \frac{\mathbf{Q}^{-1}\mathbf{1}\mathbf{1}^\top\mathbf{Q}^{-1}\mathbf{g}}{\frac{1}{z} + \mathbf{1}^\top\mathbf{Q}^{-1}\mathbf{1}} \tag{31}$$

$$= \mathbf{Q}^{-1}\mathbf{g} - \mathbf{Q}^{-1}\mathbf{1}\left(\frac{\mathbf{1}^\top\mathbf{Q}^{-1}\mathbf{g}}{\frac{1}{z} + \mathbf{1}^\top\mathbf{Q}^{-1}\mathbf{1}}\right) \tag{32}$$

$$= \mathbf{Q}^{-1}(\mathbf{g} - \mathbf{1}b) \tag{33}$$

where

$$b = \frac{\mathbf{1}\mathbf{Q}^{-1}\mathbf{g}}{\frac{1}{z} + \mathbf{1}^\top\mathbf{Q}^{-1}\mathbf{1}} = \frac{\sum_{j=1}^{J}\frac{g_j}{q_{jj}}}{\frac{1}{z} + \sum_{j=1}^{J}\frac{1}{q_{jj}}} \tag{34}$$

From (30) to (31), we utilize following matrix property,

$$(\mathbf{A}+\mathbf{B}\mathbf{C}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{B}\mathbf{C}^\top\mathbf{A}^{-1}}{1+\mathbf{C}^\top\mathbf{A}^{-1}\mathbf{B}} \tag{35}$$

where $\mathbf{A}$ as a $p \times p$ regular matrix and $\mathbf{B}$ and $\mathbf{C}$ as $p \times 1$ vectors which compose a regular matrix $\mathbf{A}+\mathbf{B}^\top\mathbf{C}$.

Each component of $\mathbf{H}^{-1}\mathbf{g}$ can be written as follows.

$$[\mathbf{H}^{-1}\mathbf{g}]_k = [\mathbf{Q}^{-1}(\mathbf{g}-\mathbf{1}b)]_k \tag{36}$$

$$= \sum_{j=1}^{J} q_{kj}^{-1}(g_k - b) \tag{37}$$

$$= \sum_{j=1}^{J} \frac{\delta(j-k)}{-N\Psi'(\alpha_k)}(g_k - b) \tag{38}$$

$$= \frac{g_k - b}{-N\Psi'(\alpha_k)} \tag{39}$$

2 Appendix

2.1 Useful property of inverse of digamma function

To compute a high-accuracy solution to $\Psi(x) = y$, use following formula.

$$\Psi^{-1}(y) \approx \begin{cases} \exp(y) + \frac{1}{2} & \text{if } y \geq -2.22 \\ -\frac{1}{y-\Psi(1)} & \text{if } y \leq -2.22 \end{cases} \tag{40}$$

Then, you implement the Newton method using the following equation.

$$x^{new} = x^{old} = \frac{\Psi(x) - y}{\Psi'(x)} \tag{41}$$

Note that Ψ' is known as a trigamma function.

3 Useful reference

- `https://endymecy.gitbooks.io/spark-ml-source-analysis/content/%E8%81%9A%E7%B1%BB/LDA/docs/dirichlet.pdf`
- `https://qiita.com/research-PORT-INC/items/9e83a49f9b07eaccef6b` (Japanese)