Maximum Likelihood Estimation of Parameters in GDA

Xiangxiang Xu
xuxx14@mails.tsinghua.edu.cn

October 11, 2019

Following the notations used in the slides, we have

$$p(y^{(i)};\phi) = \phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}} \tag{1}$$

and

$$p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x^{(i)}-\mu_{y^{(i)}})^{\mathrm{T}}\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}})\right). \tag{2}$$

Then, the log likelihood of the data can be written as

$$\begin{aligned}
l(\phi,\mu_0,\mu_1,\Sigma) &= \log\prod_{i=1}^{m} p(x^{(i)},y^{(i)};\phi,\mu_0,\mu_1,\Sigma) && (3)\\
&= \log\prod_{i=1}^{m} p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma)p(y^{(i)};\phi) && (4)\\
&= \sum_{i=1}^{m}\log p(x^{(i)}|y^{(i)};\mu_0,\mu_1,\Sigma) + \sum_{i=1}^{m}\log p(y^{(i)};\phi) && (5)\\
&= -\frac{mn}{2}\log 2\pi - \frac{m}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})^{\mathrm{T}}\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}}) && (6)\\
&\quad + \sum_{i=1}^{m}\left[y^{(i)}\log\phi + (1-y^{(i)})\log(1-\phi)\right]. && (7)
\end{aligned}$$

To find the optimal ϕ, we compute the derivative

$$\frac{\partial l(\phi,\mu_0,\mu_1,\Sigma)}{\partial\phi} = \frac{\partial}{\partial\phi}\sum_{i=1}^{m}\left[y^{(i)}\log\phi + (1-y^{(i)})\log(1-\phi)\right] \tag{8}$$

$$= \frac{1}{\phi}\sum_{i=1}^{m} y^{(i)} - \frac{1}{1-\phi}\sum_{i=1}^{m}\left[1 - y^{(i)}\right]. \tag{9}$$

Set the derivative to zero, yielding

$$\frac{1}{\phi}\sum_{i=1}^{m} y^{(i)} = \frac{1}{1-\phi}\sum_{i=1}^{m}\left[1 - y^{(i)}\right] = \frac{1}{\phi + (1-\phi)}\sum_{i=1}^{m}\left[y^{(i)} + (1 - y^{(i)})\right] = m, \tag{10}$$

which implies

$$\phi = \frac{1}{m}\sum_{i=1}^{m} y^{(i)}. \tag{11}$$

Similarly, the partial derivative with respect to μ_0 is

$$\begin{aligned}
\frac{\partial l(\phi, \mu_0, \mu_1, \Sigma)}{\partial \mu_0} &= -\frac{1}{2} \cdot \frac{\partial}{\partial \mu_0} \sum_{i=1}^{m} (x^{(i)} - \mu_{y^{(i)}})^{\mathrm{T}} \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) & (12) \\
&= -\frac{1}{2} \cdot \frac{\partial}{\partial \mu_0} \left[\sum_{i=1}^{m} (x^{(i)} - \mu_0)^{\mathrm{T}} \Sigma^{-1} (x^{(i)} - \mu_0) \cdot \mathbb{1}\{y^{(i)} = 0\} \right. \\
&\qquad \left. + \sum_{i=1}^{m} (x^{(i)} - \mu_1)^{\mathrm{T}} \Sigma^{-1} (x^{(i)} - \mu_1) \cdot \mathbb{1}\{y^{(i)} = 1\} \right] & (13) \\
&= -\frac{1}{2} \cdot \frac{\partial}{\partial \mu_0} \left[\sum_{i=1}^{m} (x^{(i)} - \mu_0)^{\mathrm{T}} \Sigma^{-1} (x^{(i)} - \mu_0) \cdot \mathbb{1}\{y^{(i)} = 0\} \right] & (14) \\
&= \sum_{i=1}^{m} \left[\Sigma^{-1} (x^{(i)} - \mu_0) \cdot \mathbb{1}\{y^{(i)} = 0\} \right] & (15) \\
&= \Sigma^{-1} \sum_{i=1}^{m} \left[(x^{(i)} - \mu_0) \cdot \mathbb{1}\{y^{(i)} = 0\} \right]. & (16)
\end{aligned}$$

Therefore, the optimal μ_0 satisfies

$$\sum_{i=1}^{m} \left[(x^{(i)} - \mu_0) \cdot \mathbb{1}\{y^{(i)} = 0\} \right] = 0, \tag{17}$$

which implies

$$\mu_0 = \frac{1}{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} = 0\}} \sum_{i=1}^{m} x^{(i)} \cdot \mathbb{1}\{y^{(i)} = 0\}. \tag{18}$$

The expression of μ_1 can be obtained similarly.

Finally, for Σ, we have

$$\begin{aligned}
\frac{\partial l(\phi, \mu_0, \mu_1, \Sigma)}{\partial \Sigma} &= -\frac{m}{2}\frac{\partial}{\partial \Sigma}\log|\Sigma| - \frac{1}{2}\frac{\partial}{\partial \Sigma}\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})^{\mathrm{T}}\Sigma^{-1}(x^{(i)} - \mu_{y^{(i)}}) \\
&= -\frac{m}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\left[\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^{\mathrm{T}}\right]\Sigma^{-1} \\
&= O,
\end{aligned}$$

which yields that

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^{\mathrm{T}}.$$

Through the derivations, we have used several facts of matrix derivatives [1]. In particular, let $A \in \mathbb{R}^{n\times n}$ be an invertible and symmetric matrix, and $v \in \mathbb{R}^n$ be a vector. Then, (15) follows from the fact that

$$\frac{\partial}{\partial v}v^{\mathrm{T}}Av = 2Av, \tag{22}$$

and (20) follows from

$$\frac{\partial}{\partial A}\log|A| = A^{-1} \tag{23}$$

and

$$\frac{\partial}{\partial A}v^{\mathrm{T}}A^{-1}v = -A^{-1}vv^{\mathrm{T}}A^{-1}. \tag{24}$$

References

[1] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.