

2 [27 points] Softmax Regression via Gradient Ascent

(a) [13 points] Derive the gradient ascent update rule for the log-likelihood of the training data.

We have:

$$l(\mathbf{w}) = \sum_{i=1}^N \sum_{k=1}^K \log [p(y^{(i)} = k | \mathbf{x}^{(i)}, \mathbf{w})] \mathbf{I}(y^{(i)}=k)$$

Taking gradient with respect to \mathbf{w}_m :

$$\begin{aligned} \nabla_{\mathbf{w}_m} l(\mathbf{w}) &= \nabla_{\mathbf{w}_m} \sum_{i=1}^N \sum_{k=1}^K \log [p(y^{(i)} = k | \mathbf{x}^{(i)}, \mathbf{w})] \mathbf{I}(y^{(i)}=k) \\ &= \nabla_{\mathbf{w}_m} \sum_{i=1}^N \sum_{k=1}^K \mathbf{I}(y^{(i)} = k) \log [p(y^{(i)} = k | \mathbf{x}^{(i)}, \mathbf{w})] \\ &= \sum_{i=1}^N \nabla_{\mathbf{w}_m} \sum_{k=1}^K \mathbf{I}(y^{(i)} = k) \log [p(y^{(i)} = k | \mathbf{x}^{(i)}, \mathbf{w})] \\ &= \sum_{i=1}^N \nabla_{\mathbf{w}_m} \sum_{k=1}^K \mathbf{I}(y^{(i)} = k) \left[\log \left(\frac{\exp(\mathbf{w}_k^T \phi(\mathbf{x}^{(i)}))}{1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi(\mathbf{x}^{(i)}))} \right) \right] \\ &= \sum_{i=1}^N \nabla_{\mathbf{w}_m} \sum_{k=1}^K \mathbf{I}(y^{(i)} = k) \left[\log \left(\exp(\mathbf{w}_k^T \phi(\mathbf{x}^{(i)})) \right) - \log \left(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi(\mathbf{x}^{(i)})) \right) \right] \\ &= \sum_{i=1}^N \nabla_{\mathbf{w}_m} \sum_{k=1}^K \mathbf{I}(y^{(i)} = k) \left[\mathbf{w}_k^T \phi(\mathbf{x}^{(i)}) - \log \left(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi(\mathbf{x}^{(i)})) \right) \right] \\ &= \sum_{i=1}^N \nabla_{\mathbf{w}_m} \left(\left[\sum_{k=1}^K \mathbf{I}(y^{(i)} = k) \mathbf{w}_k^T \phi(\mathbf{x}^{(i)}) \right] - \left[\sum_{k=1}^K \mathbf{I}(y^{(i)} = k) \log \left(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi(\mathbf{x}^{(i)})) \right) \right] \right) \end{aligned}$$

As the log term on the right does not contain k , it can be taken out of the summation and since $\sum_{k=1}^K \mathbf{I}(y^{(i)} = k) = 1$, we obtain the following:

$$\begin{aligned} &= \sum_{i=1}^N \nabla_{\mathbf{w}_m} \left(\left[\sum_{k=1}^K \mathbf{I}(y^{(i)} = k) \mathbf{w}_k^T \phi(\mathbf{x}^{(i)}) \right] - \log \left(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi(\mathbf{x}^{(i)})) \right) \right) \\ &= \sum_{i=1}^N \nabla_{\mathbf{w}_m} \left[\sum_{k=1}^K \mathbf{I}(y^{(i)} = k) \mathbf{w}_k^T \phi(\mathbf{x}^{(i)}) \right] - \nabla_{\mathbf{w}_m} \log \left(1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi(\mathbf{x}^{(i)})) \right) \end{aligned}$$

The left term contains \mathbf{w}_m iff $m = t_i$. Second term contains \mathbf{w}_m (produced by summation)

$$\begin{aligned} &= \sum_{i=1}^N \mathbf{I}(y^{(i)} = m) \phi(\mathbf{x}^{(i)}) - \frac{\exp(\mathbf{w}_m^T \phi(\mathbf{x}^{(i)})) \phi(\mathbf{x}^{(i)})}{1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_j^T \phi(\mathbf{x}^{(i)}))} \\ &= \sum_{i=1}^N \phi(\mathbf{x}^{(i)}) [\mathbf{I}(y^{(i)} = m) - p(y^{(i)} = m | \mathbf{x}^{(i)})] \end{aligned}$$

3 [22 points] Gaussian Discriminate Analysis

(a) [8 points]

Note, parameters can be omitted.

$$p(y = 1 \mid \mathbf{x}; \phi, \Sigma, \mu_0, \mu_1) = p(y = 1 \mid \mathbf{x})$$

$$p(\mathbf{x} \mid y = 1; \phi, \Sigma, \mu_0, \mu_1) = p(\mathbf{x} \mid y = 1)$$

$$p(\mathbf{x} = 1; \phi, \Sigma, \mu_0, \mu_1) = p(\mathbf{x} = 1)$$

$$p(y = 1; \phi, \Sigma, \mu_0, \mu_1) = p(y = 1)$$

Now,

$$\begin{aligned} p(y = 1 \mid \mathbf{x}; \phi, \Sigma, \mu_0, \mu_1) &= p(y = 1 \mid \mathbf{x}) \\ &= \frac{p(\mathbf{x} \mid y = 1)p(y = 1)}{p(\mathbf{x})} \\ &= \frac{p(\mathbf{x} \mid y = 1)p(y = 1)}{p(\mathbf{x} \mid y = 1)p(y = 1) + p(\mathbf{x} \mid y = 0)p(y = 0)} \\ &= \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \mu_1)^T \Sigma^{-1}(\mathbf{x} - \mu_1)\right) \phi}{\exp\left(-\frac{1}{2}(\mathbf{x} - \mu_1)^T \Sigma^{-1}(\mathbf{x} - \mu_1)\right) \phi + \exp\left(-\frac{1}{2}(\mathbf{x} - \mu_0)^T \Sigma^{-1}(\mathbf{x} - \mu_0)\right) (1 - \phi)} \\ &= \frac{1}{1 + \exp\left(\log\left(\frac{1-\phi}{\phi}\right) - \frac{1}{2}(\mathbf{x} - \mu_0)^T \Sigma^{-1}(\mathbf{x} - \mu_0) + \frac{1}{2}(\mathbf{x} - \mu_1)^T \Sigma^{-1}(\mathbf{x} - \mu_1)\right)} \\ &= \frac{1}{1 + \exp\left(\log\left(\frac{1-\phi}{\phi}\right) + \mathbf{x}^T \Sigma^{-1} \mu_0 - \mathbf{x}^T \Sigma^{-1} \mu_1 - \frac{1}{2} \mu_0^T \Sigma^{-1} \mu_0 + \frac{1}{2} \mu_1^T \Sigma^{-1} \mu_1\right)} \\ &= \frac{1}{1 + \exp\left(\log\left(\frac{1-\phi}{\phi}\right) + \mathbf{x}^T \Sigma^{-1}(\mu_0 - \mu_1) - \frac{1}{2} \mu_0^T \Sigma^{-1} \mu_0 + \mu_1^T \Sigma^{-1} \mu_1\right)} \end{aligned}$$

By setting

$$\mathbf{w}_0 = \frac{1}{2}(\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1) - \log \frac{1-\phi}{\phi},$$

$$\mathbf{w} = -\Sigma^{-1}(\mu_1 - \mu_0),$$

A constant intercept term $\mathbf{x}_0 = 1$,

$$\text{We get: } p(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{w}^T \mathbf{x} + \mathbf{w}_0 \mathbf{x}_0))} = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x})}.$$

(b) [8 points] Question (b) is the special case of (c) with $M = 1$. Let us derive the general case directly:

$$\begin{aligned}
\ell(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^N p(\mathbf{x}^{(i)} \mid y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi) \\
&= \sum_{i=1}^N \log p(\mathbf{x}^{(i)} \mid y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) + \sum_{i=1}^N \log p(y^{(i)}; \phi) \\
&= \sum_{i=1}^N \left[\log \frac{1}{(2\pi)^{\frac{M}{2}} |\Sigma|^{\frac{1}{2}}} - \frac{1}{2} \left(\mathbf{x}^{(i)} - \mu_{y^{(i)}} \right)^T \Sigma^{-1} \left(\mathbf{x}^{(i)} - \mu_{y^{(i)}} \right) + \log \phi^{y^{(i)}} + \log (1 - \phi)^{(1-y^{(i)})} \right] \\
&\simeq \sum_{i=1}^N \left[-\frac{1}{2} \log |\Sigma| - \frac{1}{2} \left(\mathbf{x}^{(i)} - \mu_{y^{(i)}} \right)^T \Sigma^{-1} \left(\mathbf{x}^{(i)} - \mu_{y^{(i)}} \right) + y^{(i)} \log \phi + (1 - y^{(i)}) \log (1 - \phi) \right]
\end{aligned}$$

(the constant term is independent of the parameters, thus removed.)

Then, the likelihood is maximized by setting the derivative with respect to each parameter to zero:

(1) with respect to ϕ :

$$\begin{aligned}
\frac{\partial \ell}{\partial \phi} &= \sum_{i=1}^N \left(\frac{y^{(i)}}{\phi} - \frac{1 - y^{(i)}}{1 - \phi} \right) \\
&= \sum_{i=1}^N \frac{1(y^{(i)} = 1)}{\phi} + \frac{N - \sum_{i=1}^N 1(y^{(i)} = 1)}{1 - \phi}
\end{aligned}$$

Therefore, $\phi = \frac{1}{N} \sum_{i=1}^N 1(y^{(i)} = 1)$, i.e. the percentage of the training examples such that $y^{(i)} = 1$.

(2) with respect to μ_0 :

$$\begin{aligned}
\nabla_{\mu_0} \ell &= -\frac{1}{2} \sum_{i: y^{(i)}=0} \nabla_{\mu_0} \left(\mathbf{x}^{(i)} - \mu_0 \right)^T \Sigma^{-1} \left(\mathbf{x}^{(i)} - \mu_0 \right) \\
&= -\frac{1}{2} \sum_{i: y^{(i)}=0} \nabla_{\mu_0} \left[-2\mu_0^T \Sigma^{-1} \mathbf{x}^{(i)} + \mu_0^T \Sigma^{-1} \mu_0 \right] \\
&= -\frac{1}{2} \sum_{i: y^{(i)}=0} \left[-2\Sigma^{-1} \mathbf{x}^{(i)} + 2\Sigma^{-1} \mu_0 \right]
\end{aligned}$$

By setting the gradient to zero,

$$\begin{aligned}
&\sum_{i: y^{(i)}=0} \left[\Sigma^{-1} \mathbf{x}^{(i)} - \Sigma^{-1} \mu_0 \right] = 0 \\
&\sum_{i=1}^N 1 \{y^{(i)} = 0\} \Sigma^{-1} \mathbf{x}^{(i)} - \sum_{i=1}^N 1 \{y^{(i)} = 0\} \Sigma^{-1} \mu_0 = 0
\end{aligned}$$

Thus we obtain $\mu_0 = \frac{\sum_{i=1}^N 1\{y^{(i)}=0\} \mathbf{x}^{(i)}}{\sum_{i=1}^N 1\{y^{(i)}=0\}}$

(3) with respect to μ_1 :

The calculations are similar for μ_1 . The resulting maximum likelihood estimate is: $\mu_1 = \frac{\sum_{i=1}^N 1\{y^{(i)}=1\} \mathbf{x}^{(i)}}{\sum_{i=1}^N 1\{y^{(i)}=1\}}$

(4) with respect to Σ :

The last step is to calculate the gradient with respect to Σ . Here, we assume $M = 1$, i.e., $|\Sigma| = \sigma^2$.

The log-likelihood of the data then can be written:

$$\begin{aligned} \ell(\phi, \mu_0, \mu_1, \Sigma) &\simeq \sum_{i=1}^N \left[-\frac{1}{2} \log |\Sigma| - \frac{1}{2} \left(\mathbf{x}^{(i)} - \mu_{y^{(i)}} \right)^T \Sigma^{-1} \left(\mathbf{x}^{(i)} - \mu_{y^{(i)}} \right) + y^{(i)} \log \phi + (1 - y^{(i)}) \log (1 - \phi) \right] \\ &= \sum_{i=1}^N \left[-\log \sigma - \frac{1}{2\sigma^2} \left(\mathbf{x}^{(i)} - \mu_{y^{(i)}} \right)^T \left(\mathbf{x}^{(i)} - \mu_{y^{(i)}} \right) + y^{(i)} \log \phi + (1 - y^{(i)}) \log (1 - \phi) \right] \end{aligned}$$

By taking derivative with respect to σ and set it to zero:

$$\nabla_{\sigma} \ell = \sum_{i=1}^N \left[-\frac{1}{\sigma} + \frac{1}{\sigma^3} \left(\mathbf{x}^{(i)} - \mu_{y^{(i)}} \right)^T \left(\mathbf{x}^{(i)} - \mu_{y^{(i)}} \right) \right] = 0$$

You obtain: $\Sigma = \frac{1}{N} \sum_{i=1}^N \left(\mathbf{x}^{(i)} - \mu_{y^{(i)}} \right)^T \left(\mathbf{x}^{(i)} - \mu_{y^{(i)}} \right)$

(c) **[6 points]** Elaborated above.