EECS 545: Homework #2

Mingliang Duanmu
duanmuml@umich.edu

February 8, 2022

1 Logistic regression

a

For each element in $\mathbf{H}$, we have

$$H_{ij} = \frac{\partial^2 E}{\partial w_i \partial w_j}$$

Since $H_{ij} = 0$ when $i \neq j$, we have

$$\begin{aligned} H_{jj} =& \frac{\partial^2}{\partial w_j^2} \sum_{i=1}^{N} -(y^{(i)} \log \sigma(\mathbf{w}^T \mathbf{x}^{(i)}) + (1 - y^{(i)}) \log(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}))) \\ =& \frac{\partial}{\partial w_j} \sum_{i=1}^{N} x_j^{(i)} (\sigma(w_j x_j^{(i)}) - y^{(i)}) \\ =& -\sum_{i=1}^{N} ((x_j^{(i)})^2 \sigma(w_j x_j^{(i)})(1 - \sigma(w_j x_j^{(i)})) \end{aligned}$$

Therefore, we can represent hessian $\mathbf{H}$ as

$$\mathbf{H} = -\mathbf{x}^T \mathbf{S} \mathbf{x}$$

where

$$S_{ii} = \sigma(\mathbf{w}^T \mathbf{x}^{(i)})(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)}))$$
$$S_{ij} = 0 (i \neq j)$$

Therefore,

$$\mathbf{z}^T \mathbf{H} \mathbf{z} = -\mathbf{z}^T \mathbf{x}^T \mathbf{S} \mathbf{x} \mathbf{z} = -\Sigma_i \Sigma_j z_i x_i x_j z_j \sigma(\mathbf{w}^T \mathbf{x}^{(i)})(1 - \sigma(\mathbf{w}^T \mathbf{x}^{(i)})) = -(\mathbf{x}^T \mathbf{z})^2 p(1 - p) \leq 0$$

where $0 \leq p \leq 1$, so the hessian is negative semi-definite and thus ℓ is concave and has no local maxima other than the global one.

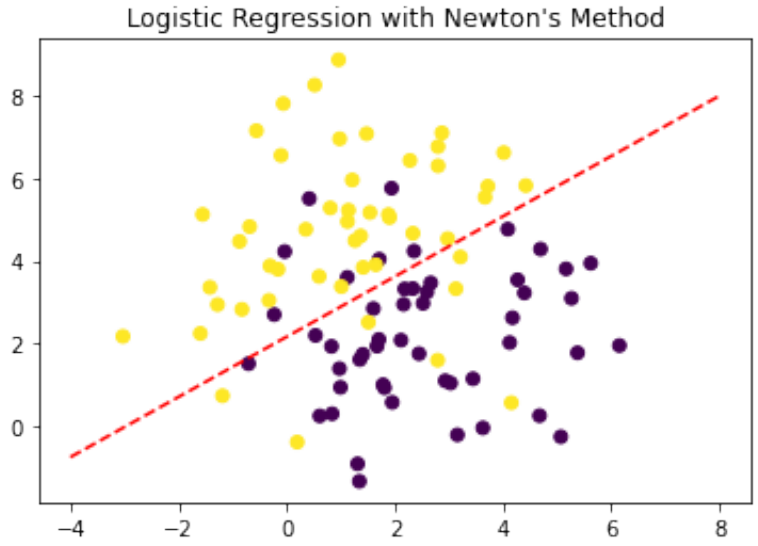
b

The update rule implied by Newton's method is

$$\mathbf{w}_{n+1} = \mathbf{w}_n - \mathbf{H}_{\ell(\mathbf{w})}^{-1} \nabla \ell(\mathbf{w})$$

The fitted coefficients are [-1.84922892 -0.62814188 0.85846843].

c



2 Softmax Regression via Gradient Ascent

a

For $y^{(i)} = k \neq m$, we have

$$\begin{aligned}
&\nabla_{w_m}(\mathbf{I}(y^{(i)} = k)p(y^{(i)} = k \mid \mathbf{x}^{(i)}, \mathbf{w})) \\
=&\mathbf{I}(y^{(i)} = k)\nabla_{w_m}(\frac{\exp(\mathbf{w}_k^T\phi(\mathbf{x}^{(i)}))}{1 + \sum_{j=1}^{K-1}\exp(\mathbf{w}_j^T\phi(\mathbf{x}^{(i)}))}) \\
=&\mathbf{I}(y^{(i)} = k)\phi(\mathbf{x}^{(i)})(-\frac{\exp(\mathbf{w}_k^T\phi(\mathbf{x}^{(i)}))}{1 + \sum_{j=1}^{K-1}\exp(\mathbf{w}_j^T\phi(\mathbf{x}^{(i)}))})
\end{aligned}$$

For $y^{(i)} = k = m$, we have

$$\begin{aligned}
&\nabla_{w_m}(\mathbf{I}(y^{(i)} = k)p(y^{(i)} = k \mid \mathbf{x}^{(i)}, \mathbf{w})) \\
=&\mathbf{I}(y^{(i)} = k)\nabla_{w_m}(\frac{\exp(\mathbf{w}_k^T\phi(\mathbf{x}^{(i)}))}{1 + \sum_{j=1}^{K-1}\exp(\mathbf{w}_j^T\phi(\mathbf{x}^{(i)}))}) \\
=&\mathbf{I}(y^{(i)} = k)\phi(\mathbf{x}^{(i)})(1 - \frac{\exp(\mathbf{w}_k^T\phi(\mathbf{x}^{(i)}))}{1 + \sum_{j=1}^{K-1}\exp(\mathbf{w}_j^T\phi(\mathbf{x}^{(i)}))})
\end{aligned}$$

By summing up all the classes, we have

$$\begin{aligned}
\nabla_{\mathbf{w}_m} l(\mathbf{w}) &= \sum_{i=1}^{N}\phi(\mathbf{x}^{(i)})\sum_{k=1}^{K}(\mathbf{I}(y^{(i)} = k)p(y^{(i)} = k \mid \mathbf{x}^{(i)}, \mathbf{w})) \\
&= \sum_{i=1}^{N}\phi(\mathbf{x}^{(i)})\left[\mathbf{I}(y^{(i)} = m) - p(y^{(i)} = m \mid \mathbf{x}^{(i)}, \mathbf{w})\right]
\end{aligned}$$

b

The accuracy of the softmax regression implementation is 94%, which is greater than the `sklearn` logistic regression model.

3 Gaussian Discriminate Analysis

a

According to Bayes Rule,

$$\begin{aligned} p(y=1 \mid \mathbf{x}; \phi, \Sigma, \mu_0, \mu_1) &= \frac{p(\mathbf{x} \mid y=1)p(y=1)}{p(\mathbf{x} \mid y=0)p(y=0) + p(\mathbf{x} \mid y=1)p(y=1)} \\ &= \frac{1}{1 + \frac{p(\mathbf{x}|y=0)p(y=0)}{p(\mathbf{x}|y=1)p(y=1)}} \end{aligned}$$

where

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

where

$$\mathbf{w} = \left[\frac{1}{2}(\mu_1^T\Sigma^{-1}\mu_1 - \mu_0^T\Sigma^{-1}\mu_0) + \log\frac{p(y=0)}{p(y=1)}, (\mu_0-\mu_1)\Sigma^{-1}\right]^T$$

and $\mathbf{x}$ has $M+1$ dimensions with $\mathbf{x}_0 = 1$.

b

$$\begin{aligned} \ell(\phi, \mu_0, \mu_1, \Sigma) &= \log\prod_{i=1}^{N} p(\mathbf{x}^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) \\ &= \sum_{i=1}^{N} \log(p(\mathbf{x}^{(i)} \mid y^{(i)}; \phi, \mu_0, \mu_1, \Sigma)p(y^{(i)}; \phi)) \end{aligned}$$

To maximize $\ell(\phi, \mu_0, \mu_1, \Sigma)$, we take the partial derivative to find ϕ

$$\begin{aligned} &\frac{\partial}{\partial\phi}\ell(\phi, \mu_0, \mu_1, \Sigma) \\ =&\frac{\partial}{\partial\phi}\sum_{i=1}^{N}(-\frac{M}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{x}^{(i)}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mu) + y^{(i)}\log\phi + (1-y^{(i)})\log(1-\phi)) \\ =&\sum_{i=1}^{N}(\frac{y^{(i)}}{\phi} - \frac{1-y^{(i)}}{1-\phi}) = 0 \end{aligned}$$

Therefore,

$$\phi = \frac{1}{N}\sum_{i=1}^{N} 1\{y^{(i)} = 1\}$$

Similarly, we find μ_0 by

$$\begin{aligned} &\frac{\partial}{\partial\mu_0}\ell(\phi, \mu_0, \mu_1, \Sigma) \\ =&\frac{\partial}{\partial\mu_0}\sum_{i=1}^{N}(-\frac{M}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{x}^{(i)}-\mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)}-\mu) + y^{(i)}\log\phi + (1-y^{(i)})\log(1-\phi)) \\ =&\sum_{i=1}^{N}(-1\{y^{(i)}=0\}\mathbf{\Sigma}^{-1}\mu_0 + 1\{y^{(i)}=0\}\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)})^T) = 0 \end{aligned}$$

We get

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

Similarly, we get

$$\mu_1 = \frac{\sum_{i=1}^{N} 1\{y^{(i)} = 1\}\mathbf{x}^{(i)}}{\sum_{i=1}^{N} 1\{y^{(i)} = 1\}}$$

Similarly, we find σ by

$$\begin{aligned}
&\frac{\partial}{\partial\sigma}\ell(\phi, \mu_0, \mu_1, \Sigma) \\
=&\frac{\partial}{\partial\sigma}\sum_{i=1}^{N}(-\frac{M}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{x}^{(i)} - \mu)^T\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)} - \mu) + y^{(i)}\log\phi + (1 - y^{(i)})\log(1 - \phi)) \\
=&\sum_{i=1}^{N}(-\frac{1}{2\Sigma} + \frac{1}{2\Sigma^2}(x^{(i)} - \mu_{y^{(i)}})^2) = 0
\end{aligned}$$

We get

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

c

The proof is exactly the same as the case when $M = 1$

$$\ell(\phi, \mu_0, \mu_1, \Sigma) = \sum_{i=1}^{N}\log(p(\mathbf{x}^{(i)} \mid y^{(i)}; \phi, \mu_0, \mu_1, \Sigma)p(y^{(i)}; \phi))$$

To maximize $\ell(\phi, \mu_0, \mu_1, \Sigma)$, we take the partial derivative to find ϕ

$$\frac{\partial}{\partial\phi}\ell(\phi, \mu_0, \mu_1, \Sigma) = \sum_{i=1}^{N}(\frac{y^{(i)}}{\phi} - \frac{1 - y^{(i)}}{1 - \phi}) = 0$$

Therefore,

$$\phi = \frac{1}{N}\sum_{i=1}^{N} 1\{y^{(i)} = 1\}$$

Similarly, we find μ_0 by

$$\frac{\partial}{\partial\mu_0}\ell(\phi, \mu_0, \mu_1, \Sigma) = \sum_{i=1}^{N}(-1\{y^{(i)} = 0\}\mathbf{\Sigma}^{-1}\mu_0 + 1\{y^{(i)} = 0\}\mathbf{\Sigma}^{-1}(\mathbf{x}^{(i)})^T) = 0$$

We get

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

Similarly, we get

$$\mu_1 = \frac{\sum_{i=1}^{N} 1\{y^{(i)} = 1\}\mathbf{x}^{(i)}}{\sum_{i=1}^{N} 1\{y^{(i)} = 1\}}$$

4 Naive Bayes for classifying SPAM

a

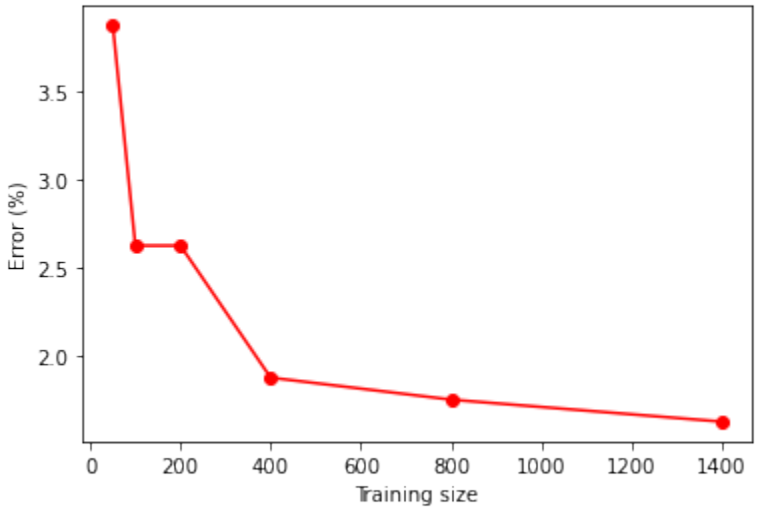
The error is 1.6250%.

b

The top 5 tokens that are most indicative of the SPAM class are [’httpaddr’, ’spam’, ’unsubscrib’, ’ebai’, ’valet’].

c

Training size: 50, Error: 3.8750%
Training size: 100, Error: 2.6250%
Training size: 200, Error: 2.6250%
Training size: 400, Error: 1.8750%
Training size: 800, Error: 1.7500%
Training size: 1400, Error: 1.6250%



The training set size of 1400 gives the best classification error.