XCS299i Problem Set #2

JUAN RICARDO PEDRAZA ESCOBAR

Machine Learning

Stanford Center for Professional Development

August,2021

1.a

Since $gt(z) = g(z)(1 - g(z))$ and $h(x) = g(\theta T\, x)$, it follows $\partial h(x)/\partial\theta_k = h_\theta(x^{(i)})\left(1 - h_\theta(x^{(i)})\right)$.

Letting $h_\theta(x^{(i)}) = g(\theta T\, x(i)) = 1/(1 + exp(-\theta T\, x(i)))$, we have

$$\frac{\partial \log h_\theta(x^{(i)})}{\partial \theta_k} = \frac{1}{h_\theta(x^{(i)})} \frac{\partial h_\theta(x^{(i)})}{\partial \theta_k}$$

Replacing $\partial h(x)/\partial\theta_k$,

$$\frac{\partial \log h_\theta(x^{(i)})}{\partial \theta_k} = \frac{1}{h_\theta(x^{(i)})} h_\theta(x^{(i)})\left(1 - h_\theta(x^{(i)})\right) x_k = \left(1 - h_\theta(x^{(i)})\right) x_k$$

$$\frac{\partial \log\left(1 - h_\theta(x^{(i)})\right)}{\partial \theta_k} = \frac{1}{1 - h_\theta(x^{(i)})} \frac{\partial\left(1 - h_\theta(x^{(i)})\right)}{\partial \theta_k} = h_\theta(x^{(i)})\, x_k$$

Substituting into our equation for $J(\theta)$, we have

$$\frac{\partial J(\theta)}{\partial \theta_k} = -\frac{1}{n}\sum_{i=1}^{n} y^{(i)}\left(1 - h_\theta(x^{(i)})\right) x_k - \left(1 - y^{(i)}\right) h_\theta(x^{(i)})\, x_k$$

$$\frac{\partial J(\theta)}{\partial \theta_k} = -\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - h_\theta(x^{(i)})\right) x_k$$

Consequently, the (k, l) entry of the Hessian is given by,

$$H_{kl} = \frac{\partial^2 J(\theta)}{\partial \theta_k\, \partial \theta_l}$$

$$\frac{\partial^2 J(\theta)}{\partial \theta_k\, \partial \theta_l} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial J(\theta)}{\partial \theta_l}\left(y^{(i)} - h_\theta(x^{(i)})\right) x_k$$

$$= \frac{1}{n}\sum_{i=1}^{n} \frac{\partial J(\theta)}{\partial \theta_l}\left(y^{(i)} - h_\theta(x^{(i)})\right) x_k$$

By deriving whit respect to θl,

$$\frac{\partial^2 J(\theta)}{\partial \theta_k\, \partial \theta_l} = \frac{1}{n}\sum_{i=1}^{n} h_\theta(x^{(i)})\left(1 - h_\theta(x^{(i)})\right) x_k^{(i)}\, x_l^{(i)}$$

Using the fact that Xij = xixj if and only if X = xxT , we have

$$H = \frac{1}{n}\sum_{i=1}^{n} h_\theta\left(x^{(i)}\right)\left(1 - h_\theta\left(x^{(i)}\right)\right) x^{(i)}x^{(i)T}$$

To prove that H is positive semi-definite, show $z^T H\, z \geq 0$ for all z ∈ Rd.

$$z^T H\, z = \frac{1}{n}\, z^T \left(\sum_{i=1}^{n} h_\theta\left(x^{(i)}\right)\left(1 - h_\theta\left(x^{(i)}\right)\right) x^{(i)}x^{(i)T}\right) z$$

$$z^T H\, z = \frac{1}{n}\sum_{i=1}^{n} h_\theta\left(x^{(i)}\right)\left(1 - h_\theta\left(x^{(i)}\right)\right) (z^T x^{(i)})(x^{(i)T} z)$$

$$z^T H\, z = \frac{1}{n}\sum_{i=1}^{n} h_\theta\left(x^{(i)}\right)\left(1 - h_\theta\left(x^{(i)}\right)\right) (z^T x^{(i)})^2$$

- Note that the function $h_\theta\left(x^{(i)}\right)$ follows $0 \leq h_\theta\left(x^{(i)}\right) \geq 1$ and the square term $(z^T x^{(i)})^2$ is always positive so ,

$$\frac{1}{n}\sum_{i=1}^{n} h_\theta\left(x^{(i)}\right)\left(1 - h_\theta\left(x^{(i)}\right)\right) (z^T x^{(i)})^2 \quad \geq \quad 0$$

$$z^T H\, z \quad \geq \quad 0$$

1.c

For shorthand, we let $H = \{\phi, \Sigma, \mu 0, \mu 1\}$ denote the parameters for the problem. Since the given formulae are conditioned on y, use Bayes rule to get:

$$p(y = 1|x; H) = \frac{p(x|y = 1; H)p(y = 1; H)}{p(x; H)}$$

$$= \frac{p(x|y = 1; H)p(y = 1; H)}{p(x|y = 1; H)p(y = 1; H) + p(x|y = 0; H)p(y = 0; H)}$$

We replace the equations that we use to model the distribution of (x, y),

$$p(y = 1|x; H) = \frac{exp\left(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\right)\ \phi}{exp\left(-\frac{1}{2}(x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\right)\phi + exp\left(-\frac{1}{2}(x - \mu_0)^T \Sigma^{-1}(x - \mu_0)\right)\ (1 - \phi)}$$

$$= \frac{1}{1 + exp\left(log\left(\frac{1-\phi}{\phi}\right) - \frac{1}{2}\ (x - \mu_0)^T \Sigma^{-1}(x - \mu_0) + \frac{1}{2}\ (x - \mu_1)^T \Sigma^{-1}(x - \mu_1)\right)}$$

We take the quadratic term located in the previous expression denominator and apply properties of distribution,

$$(x - \mu_0)^T \Sigma^{-1}(x - \mu_0) + (x - \mu_1)^T \Sigma^{-1}(x - \mu_1)$$

$$= x^T \Sigma^{-1} x - \mu_0^T \Sigma^{-1} x - x^T \Sigma^{-1} \mu_0 + \mu_0^T \Sigma^{-1} \mu_0 - x^T \Sigma^{-1} x + \mu_1^T \Sigma^{-1} x + x^T \Sigma^{-1} \mu_1 - \mu_1^T \Sigma^{-1} \mu_1$$

Simplifying and reorganizing the term,

$$= 2(\mu_1 - \mu_0)^T \Sigma^{-1} x + \mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1$$

Therefore,

$$p(y = 1|x; H) = \frac{1}{1 + exp\left(log\left(\frac{1-\phi}{\phi}\right) + (\mu_1 - \mu_0)^T \Sigma^{-1} x + \frac{1}{2}\ \mu_1^T \Sigma^{-1} \mu_1 - \frac{1}{2}\ \mu_0^T \Sigma^{-1} \mu_0\right)}$$

Where,

- $\theta = -(\mu_0 - \mu_1)\Sigma^{-1}$
- $\theta_0 = \frac{1}{2}\ \mu_0^T \Sigma^{-1} \mu_0 - \frac{1}{2}\ \mu_1^T \Sigma^{-1} \mu_1 - log\left(\frac{1-\phi}{\phi}\right)$

Finally gives,

$$p(y = 1|x; \phi, \Sigma, \mu_0, \mu_1) = p(y = 1|x; H) = \frac{1}{1 + exp(-(\theta^T x + \theta_0))}$$

1.d

First, derive the expression for the log-likelihood of the training data:

$$\ell\,(\phi, \mu_0\,, \mu_1, \Sigma) \;=\; log \prod_{i=1}^{n} p\big(x^{(i)}\,\big|y^{(i)}\,;\;\mu_0\,, \mu_1, \Sigma\big)\;\; p(y^{(i)}\,;\;\phi)$$

$$= \sum_{i=1}^{n} log\, p\big(x^{(i)}\,\big|y^{(i)}\,;\;\mu_0\,, \mu_1, \Sigma\big)\; + \sum_{i=1}^{n} log\;\; p(y^{(i)}\,;\;\phi)$$

Using the equations for this model,

$$= \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\;\; \exp\left(-\frac{\big(x^{(i)} - \big(1 - y^{(i)}\big)\mu_0 - y^{(i)}\mu_1\big)^2}{2\sigma^2}\right) + \sum_{i=1}^{n} log\;\phi^{\,y^{(i)}}(1-\phi)^{1-y^{(i)}}$$

$$= -m\,log\big(\sqrt{2\pi}\sigma\big) + \sum_{i=1}^{n}\left[y^{(i)}\,log\,\phi + \big(1 - y^{(i)}\,\big)(log(1-\phi)) - \frac{\big(x^{(i)} - \big(1 - y^{(i)}\big)\mu_0 - y^{(i)}\mu_1\big)^2}{2\sigma^2}\right]$$

Now, the likelihood is maximized by setting the derivative (or gradient) with respect to each of the parameters to zero.

- For ϕ :

$$\frac{\partial \ell}{\partial \phi} = \sum_{i=1}^{n} y^{(i)}\frac{1}{\phi} + (1 - y^{(i)}\,)\frac{1}{1-\phi}$$

$$\frac{\partial \ell}{\partial \phi} = \sum_{i=1}^{n} y^{(i)}\frac{1}{\phi} + (1 - y^{(i)}\,)\frac{1}{1-\phi}$$

$$= \frac{\Sigma_{i=1}^{n}\; 1\,(y^{(i)} = 1)}{\phi} + \frac{m - \Sigma_{i=1}^{n}\; 1\,(y^{(i)} = 1)}{1-\phi}$$

Setting this equal to zero and solving for ϕ gives the maximum likelihood estimate.

$$0 = \frac{\Sigma_{i=1}^{n}\; 1\,(y^{(i)} = 1)}{\phi} + \frac{m - \Sigma_{i=1}^{n}\; 1\,(y^{(i)} = 1)}{1-\phi}$$

$$0 = \sum_{i=1}^{n}\; 1\,(y^{(i)} = 1)\;\,(1-\phi) + \left(m - \sum_{i=1}^{n}\; 1\,(y^{(i)} = 1)\right)\phi$$

$$\phi = \frac{\sum_{i=1}^{n} 1\left(y^{(i)} = 1\right)}{n\ \phi}$$

$$\phi = \frac{1}{n}\sum_{i=1}^{n} 1\left(y^{(i)} = 1\right)$$

- For μ0:

Hint: Remember that Σ (and thus Σ^{-1}) is symmetric.

$$\nabla_{\mu_0}\ell = \frac{\partial \ell}{\partial \mu_0}\sum_{i=1}^{n} -\frac{\left(x^{(i)} - \left(1 - y^{(i)}\right)\mu_0 - y^{(i)}\mu_1\right)^2}{2\sigma^2}$$

$$= \sum_{i=1}^{n} -\left(\frac{\left(x^{(i)} - (1 - y^{(i)})\mu_0 - y^{(i)}\mu_1\right)}{\sigma^2}\right)\frac{\partial \ell}{\partial \mu_0}\left(x^{(i)} - (1 - y^{(i)})\mu_0 - y^{(i)}\mu_1\right)$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x^{(i)} - \mu_0\right)\quad 1\left(y^{(i)} = 0\right)$$

Setting this gradient to zero gives the maximum likelihood estimate for μ0.

$$0 = \sum_{i=1}^{n}\left(x^{(i)} - \mu_0\right)\quad 1\left(y^{(i)} = 0\right)$$

$$0 = \sum_{i=1}^{n} 1\left(y^{(i)} = 0\right)x^{(i)} - \sum_{i=1}^{n} 1\left(y^{(i)} = 0\right)\mu_0$$

$$\mu_0 = \frac{\sum_{i=1}^{n} 1\left(y^{(i)} = 0\right)x^{(i)}}{\sum_{i=1}^{n} 1\left(y^{(i)} = 0\right)\mu_0}$$

- For μ1:

Hint: Remember that Σ (and thus Σ−1) is symmetric.

$$\nabla_{\mu_1}\ell = \frac{\partial \ell}{\partial \mu_1}\sum_{i=1}^{n} -\frac{\left(x^{(i)} - \left(1 - y^{(i)}\right)\mu_0 - y^{(i)}\mu_1\right)^2}{2\sigma^2}$$

$$= \sum_{i=1}^{n} -\left(\frac{\left(x^{(i)} - (1 - y^{(i)})\mu_0 - y^{(i)}\mu_1\right)}{\sigma^2}\right)\frac{\partial \ell}{\partial \mu_1}\left(x^{(i)} - (1 - y^{(i)})\mu_0 - y^{(i)}\mu_1\right)$$

$$= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x^{(i)} - \mu_1\right)\quad 1\left(y^{(i)} = 1\right)$$

Setting this gradient to zero gives the maximum likelihood estimate for μ1.

$$0 = \sum_{i=1}^{n} \left(x^{(i)} - \mu_0\right) \quad 1\left(y^{(i)} = 1\right)$$

$$0 = \sum_{i=1}^{n} 1\left(y^{(i)} = 1\right)x^{(i)} - \sum_{i=1}^{n} 1\left(y^{(i)} = 0\right)\mu_1$$

$$\mu_1 = \frac{\sum_{i=1}^{n} 1\left(y^{(i)} = 1\right)x^{(i)}}{\sum_{i=1}^{n} 1\left(y^{(i)} = 1\right)\mu_0}$$

- For Σ, note that $\Sigma = \sigma^2$,

$$\frac{\partial \ell}{\partial \sigma^2} = \frac{\partial \ell}{\partial \sigma^2} - m\,log\left(\sqrt{2\pi}\sigma\right) + \sum_{i=1}^{n}\left[y^{(i)}\,log\,\phi + \left(1 - y^{(i)}\right)(log(1-\phi)) - \frac{\left(x^{(i)} - \left(1 - y^{(i)}\right)\mu_0 - y^{(i)}\mu_1\right)^2}{2\sigma^2}\right]$$

$$= \frac{\partial \ell}{\partial \sigma^2}\left[-m\,log\left(\sqrt{2\pi}\sigma\right) - \frac{\left(x^{(i)} - \left(1 - y^{(i)}\right)\mu_0 - y^{(i)}\mu_1\right)^2}{2\sigma^2}\right]$$

$$= -\frac{m}{2\sigma^2} + \sum_{i=1}^{n} 1\left(y^{(i)} = 0\right)\frac{\left(x^{(i)} - \mu_0\right)^2}{\sigma^4} + \sum_{i=1}^{n} 1\left(y^{(i)} = 1\right)\frac{\left(x^{(i)} - \mu_1\right)^2}{\sigma^4}$$

Setting this gradient to zero gives the maximum likelihood estimate for σ^2

$$\sigma^2 = \frac{1}{n}\left(\sum_{i=1}^{n} 1\left(y^{(i)} = 0\right)\left(x^{(i)} - \mu_0\right)^2 + \sum_{i=1}^{n} 1\left(y^{(i)} = 1\right)\left(x^{(i)} - \mu_1\right)^2\right)$$

$$\Sigma = \sigma^2 = \frac{1}{n}\left(\sum_{i=1}^{n}\left(x^{(i)} - y_{(i)}\right)\left(x^{(i)} - y_{(i)}\right)^T\right)$$

1.f

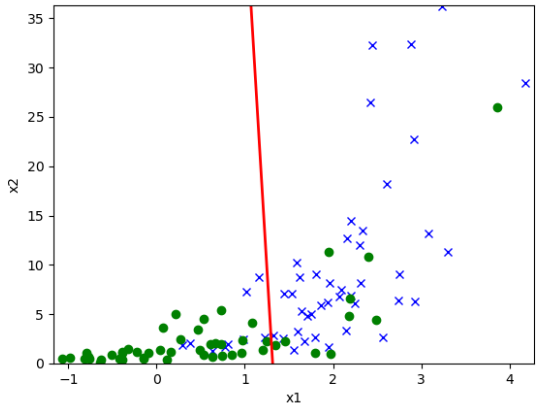

Figure 1 . GDA decision boundary on data set 1

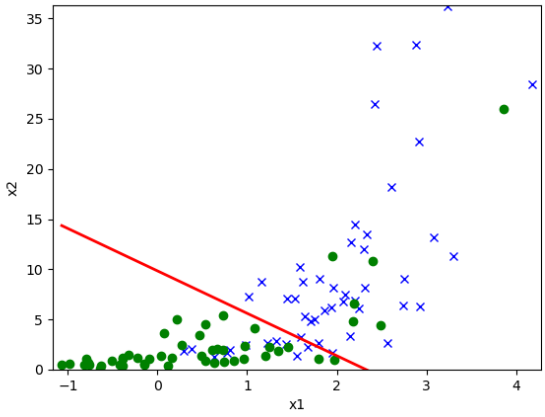

Figure 2. Logistic regression decision boundary dataset 1.

- Comment

The decision boundary from logistic regression seems more reasonable on Dataset1. The logistic decision boundary puts weight on both x1 and x2. However, with the GDA decision boundary the accuracy is lower on Dataset 1.

1.g

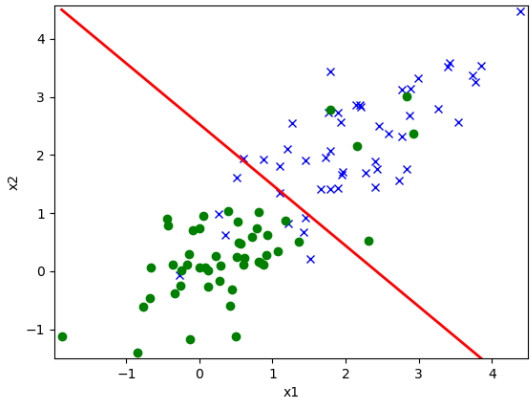

Figure 3. GDA Decision Boundary on Dataset 2.

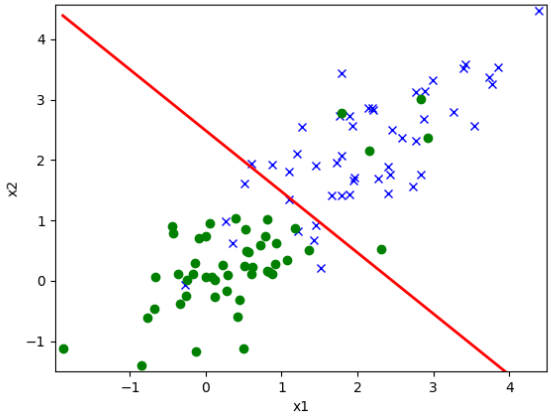

Figure 4. Logistic regression Decision Boundary on Dataset 2.

- Comment

Decision boundaries from logistic regression and GDA are nearly identical for Dataset 2. However, in the first dataset (Dataset 1) GDA seem to perform worse than logistic regression. This probably relates to the fact that the training data(xi) are not Gaussian and make the GDA model performs worse. Also, as we already know logistic regression is more powerful when the data set is not drawn from a multivariate Gaussian.

1.h

- Comment.

By setting,

$$x_2^{(i)} := log\ x_2^{(i)}$$

all $x_2^{(i)}$ in the Dataset 2 become Gaussian, therefore the GDA model performs significantly better.

2.a

$$p(y\,;\,\lambda) = \frac{e^{-\lambda}\, e^{y \log \lambda}}{y\,!}$$

$$p(y\,;\,\lambda) = \frac{1}{y!}\exp(y \log \lambda - \lambda)$$

Using the standard form for the exponential family to compare,

$$\eta = \log \lambda$$

$$b(y) = \frac{1}{y!}$$

$$T(y) = y$$

$$\alpha(\eta) = e^{\,\eta}$$

2.b

For GML model the canonical response function will be,

$$g(\eta) = E[Y; \eta]$$

$$g(\eta) = \lambda$$

$$g(\eta) = e^{\eta}$$

2.c

The log-likelihood of an example $(x(i), y(i))$ is defined as $\ell(\theta) = log\ p(y(i)\ x(i);\ \theta)$. To derive the stochastic gradient ascent rule, use the results in part (a) and the standard GLM assumption that $\eta = \theta T\ x$.

$$\frac{\partial \ell\ (\theta)}{\partial \theta_j} = \frac{\partial \log(\ p\ \ y^{(i)}\ |x^{(i)};\ \theta\ )}{\partial \theta_j}$$

$$= \frac{\partial \log\ \left(\frac{1}{y!}\exp(\ \eta^T\ y^{(i)} - e^{\ \eta})\right)}{\partial \theta_j}$$

$$= \frac{\partial \log\ \left(\exp(\ \eta^T\ y^{(i)} - e^{\ \eta})\right)}{\partial \theta_j} + \frac{\partial \log\ \left(\frac{1}{y!}\right)}{\partial \theta_j}$$

$$= \frac{\partial \left(\left(\sum_k \theta_k x_k^{(i)}\right) y^{(i)} - e^{\theta_k x_k^{(i)}}\right)}{\partial \theta_j}$$

$$= \left(y^{(i)} - e^{\theta_k x_k^{(i)}}\right) x_j^{(i)}$$

Thus, the stochastic gradient ascent update rule should be:

$$\theta j := \theta j\ +\ \alpha\ \frac{\partial \ell\ (\theta)}{\partial \theta_j}$$

which reduces here to:

$$\theta j := \theta j\ +\ \alpha\ \left(y^{(i)} - e^{\theta_k x_k^{(i)}}\right) x_j^{(i)}$$