CS229 Fall 2017
Problem Set #1 Solutions: Supervised Learning

Author: LFhase rimemosa@163.com

Logistic regression

(a) Given that

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} log(1 + e^{-y^{(i)}\theta^T x^{(i)}})$$

we can get

$$\frac{\partial J(\theta)}{\partial \theta_i} = \frac{1}{m}\sum_{k=1}^{m} \frac{-y^{(k)}x_i^{(k)}}{1 + e^{y^{(k)}\theta^T x^{(k)}}}$$

then

$$\frac{\partial J(\theta)}{\partial \theta_i \partial \theta_j} = \frac{1}{m}\sum_{k=1}^{m} \frac{x_i^{(k)}x_j^{(k)}e^{y^{(k)}\theta^T x^{(k)}}}{(1 + e^{y^{(k)}\theta^T x^{(k)}})^2}$$

which is H_{ij}
so

$$\begin{aligned} Z^T HZ &= \sum_{i=1}^{n}\sum_{j=1}^{n} z_i H_{ij} z_j \\ &= \frac{1}{m}\sum_{k=1}^{m} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} z_i x_i^{(k)} x_j^{(k)} z_j e^{y^{(k)}\theta^T x^{(k)}}}{(1 + e^{y^{(k)}\theta^T x^{(k)}})^2} \end{aligned}$$

known that

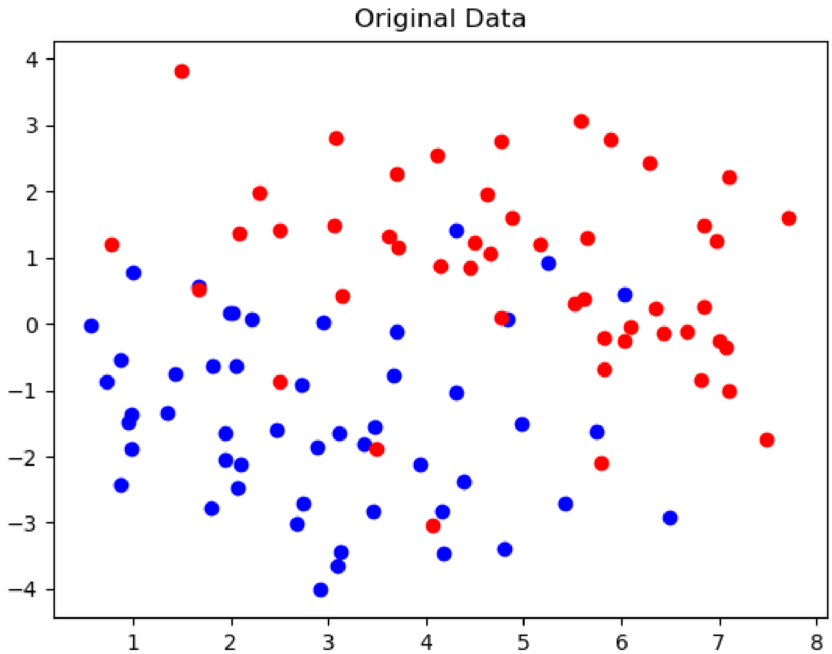
$$\sum_{i=1}^{n}\sum_{j=1}^{n} z_i x_i^{(k)} x_j^{(k)} z_j = (X^T Z)^2 \geq 0$$

$$\frac{e^{y^{(k)}\theta^T x^{(k)}}}{(1 + e^{y^{(k)}\theta^T x^{(k)}})^2} > 0$$

we can easily get

$$Z^T HZ \geq 0$$

(b) Firstly, we plot the original data



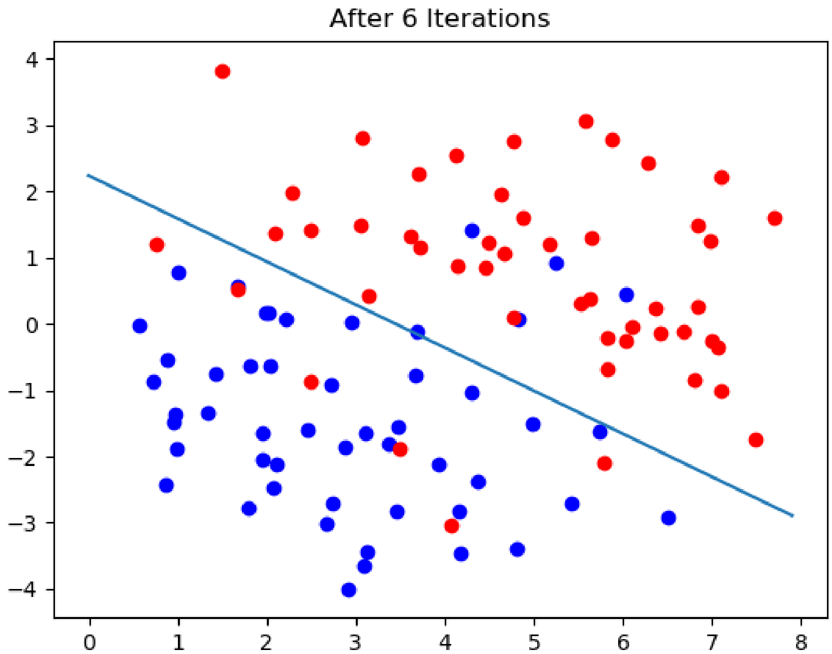
To implement Newton's Method, we calculate the value of

$$\frac{1}{m}\sum_{k=1}^{m}\frac{-y^{(k)}x_i^{(k)}}{1+e^{y^{(k)}\theta^T x^{(k)}}}$$

and **Hessian**
then using the **update rule**

$$\theta := \theta - H^{-1}\bigtriangledown_\theta\, l(\theta)$$

(c) Through 6 iterations, we finally get the result



and the θ is [-2.6205116, 0.76037154, 1.17194674].

Poisson regression and the exponential family

(a) Firstly, we get

$$\begin{aligned} p(y;\lambda) &= \frac{e^{-\lambda}\lambda^y}{y!} \\ &= \frac{1}{y!}exp(log\lambda^y - loge^\lambda) \\ &= \frac{1}{y!}exp(log\lambda y - \lambda) \end{aligned}$$

It's easy to know that

$$\begin{aligned} log\lambda = \eta, \lambda = e^\eta \\ T(y) = y \\ a(\eta) = -e^\eta \\ b(y) = \frac{1}{y!} \end{aligned}$$

and Poisson distribution is in the exponential family.

(b) Known that if $Y \sim P(\lambda)$, then $E[Y] = \lambda$,

$$\begin{aligned} h_\theta(x) &= E[y|x;\theta] \\ &= \lambda \\ &= e^\eta = e^{\theta^T x} \end{aligned}$$

so the response function for the family is

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

.

(c) Firstly, we can get likelihood function $L(\theta)$

$$\begin{aligned} L(\theta) &= \Pi_{k=1}^m P(y^{(k)}|x^{(k)};\lambda) \\ &= \Pi_{k=1}^m \frac{e^{-\lambda}\lambda^{y^{(k)}}}{y^{(k)}!} \end{aligned}$$

Then, the log-likelihood function $l(\theta)$

$$\begin{aligned} l(\theta) &= \sum_{k=1}^m (-logy^{(k)}!) + (-\lambda) + y^{(k)}log\lambda \\ &= \sum_{k=1}^m (-logy^{(k)}!) + (-e^{\theta^T x^{(k)}}) + y^{(k)}\theta^T x^{(k)} \end{aligned}$$

Then, we can get the derivative of the log-likelihood with respect to θ_i

$$\frac{\partial l(\theta)}{\partial \theta_i} = \sum_{k=1}^{m} (y^{(k)} - e^{\theta^T x^{(k)}}) x_i^{(k)}$$

So, the stochastic gradient ascent learning rule is

$$\theta_i = \theta_i + \alpha(y^{(k)} - e^{\theta^T x^{(k)}}) x_i^{(k)}$$

(d) Firstly, we can know that

$$p(y|x;\theta) = b(y)exp(\eta^T y - a(\eta))$$

then, for a simple training data (x, y) in stochastic gradient ascent,

$$\frac{\partial p(y|x;\theta)}{\partial \theta_i} = x_i(y - \frac{\partial a(\eta)}{\partial \eta})$$

known that,

$$\int_y p(y|x;\eta) = 1$$

we can imply derivation in both sides, and get

$$\int_y p(y|x;\eta)(y - \frac{\partial a(\eta)}{\partial \eta}) = 0$$

finally, we get

$$\frac{\partial a(\eta)}{\partial \eta} = \int_y p(y|x;\eta) y = h_\theta(x)$$

so the gradient we get is

$$\frac{\partial p(y|x;\theta)}{\partial \theta_i} = x_i(y - h_\theta(x))$$

the update rule is

$$\theta_i = \theta_i - \alpha(h_\theta(x) - y) x_i$$

Gaussian discriminant analysis

(a) Firstly, we have

$$p(y=1|x;\phi,\Sigma,\mu_1,\mu_-1) = \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1)+p(x|y=-1)p(y=-1)}$$
$$= \frac{1}{1+\frac{p(x|y=-1)p(y=-1)}{p(x|y=1)p(y=1)}}$$

apply such calculattion, we can get

$$p(y=1|x;\phi,\Sigma,\mu_1,\mu_-1) = \frac{1}{1+exp(-y(ln\frac{\phi}{1-\phi}+\frac{1}{2}((x-\mu_{-1})^T\Sigma^{-1}(x-\mu_{-1})-(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))}$$
$$= \frac{1}{1+exp(-y(ln\frac{\phi}{1-\phi}+(\mu_1-\mu_{-1})\Sigma^{-1}x+\frac{1}{2}(\mu_{-1}\Sigma^{-1}\mu_{-1}-\mu_1\Sigma^{-1}\mu_1)))}$$

It's the same when $y=-1$. They can be written in the form of a logistic function where

$$\theta = (\mu_1-\mu_{-1})\Sigma^{-1}$$

$$\theta_0 = ln\frac{\phi}{1-\phi}+1/2(\mu_{-1}\Sigma^{-1}\mu_{-1}-\mu_1\Sigma^{-1}\mu_1)$$

(b) Let m_1 be the number of samples with label $y=1$, and m_{-1} be the number of samples with label $y=-1$. The log-likelihood function is

$$l(\theta,\mu_{-1},\mu_1,\Sigma) = mlog(\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}})+m_1log\phi+m_{-1}log(1-\phi)+\sum_{i=1}^{m}(-\frac{1}{2})(x^{(i)}-\mu_{y^{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}})$$

so we can derive the partial derivate of each parameter,

$$\frac{\partial l}{\partial \phi} = \frac{m_1-m\phi}{\phi(1-\phi)}$$

$$\frac{\partial l}{\partial \mu_1} = \frac{\sum_{i=1}^{m_1}(x^{(i)}-\mu_1)}{\Sigma}$$

$$\frac{\partial l}{\partial \mu_{-1}} = \frac{\sum_{j=1}^{m_1}(x^{(j)}-\mu_{-1})}{\Sigma}$$

$$\frac{\partial l}{\partial \Sigma} = \frac{-\frac{1}{2}(m\Sigma-\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})^T(x^{(i)}-\mu_{y^{(i)}}))}{\Sigma^2}$$

let these partial derivates to be zero, we can get the maximum likelihood estimates the same as problem statement.

(c) As shown in (b)

Linear invariance of optimization algorithms

(a) According to Newton's Method, we know that

$$x^{(i+1)} = x^{(i)} - \frac{f(x^{(i)})}{f'(x^i)}$$

So

$$\begin{aligned}
z^{(i+1)} &= z^{(i)} - \frac{g(z^{(i)})}{g'(z^{(i)})} \\
&= z^{(i)} - \frac{f(Az^{(i)})}{f'(Az^{(i)})} \\
&= A^{(-1)}x^{(i)} - A^{(-1)}(x^{(i)} - x^{(i+1)}) \\
&= A^{(-1)}x^{(i+1)}
\end{aligned}$$

so that Newton's Method is invariant to linear reparameterizations.

(b) In gradient descent algorithm

$$\begin{aligned}
z^{(i+1)} &= z^{(i)} - \alpha g'(z^{(i)}) \\
&= A^{(-1)}x^{(i)} - A(x^{(i)} - x^{(i+1)}) \\
&\neq A^{(-1)}x^{(i+1)}
\end{aligned}$$

Regression for denoising quasar spectra

(a) (i) We know that the form can be written in

$$\sum_{i,j} W_{ij}(\Theta^T x^{(i)} - y^i)(\Theta^T x^{(j)} - y^j)$$

to let them fit, just let $W_{ii} = \frac{1}{2}w^{(i)}$

(ii) Firstly, we get the derivative

$$\bigtriangledown_\theta J(\theta) = 2(W(X\theta - \vec{y})^T X)$$

let the derivative be zero, we get

$$\theta = (X^T W X)^{-1} X^T W \vec{y}$$

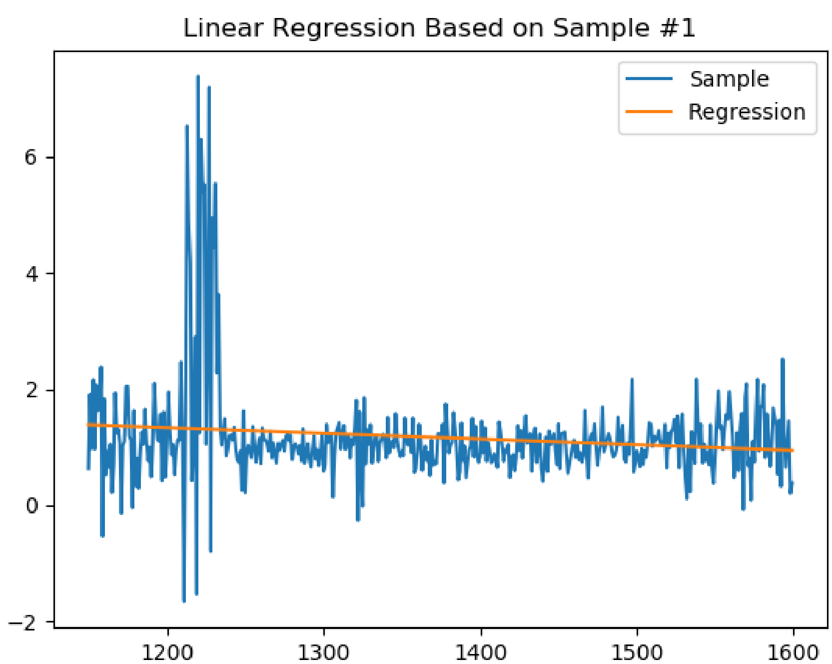
since W is normal, θ is what we want.

(iii) Write the log-likelihood function

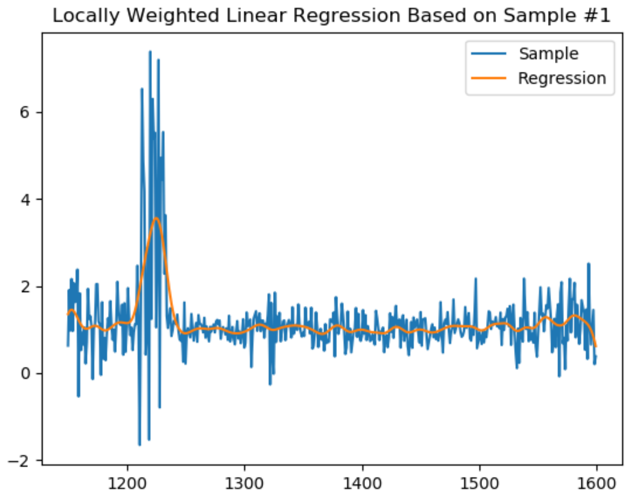
$$l = \sum_{i=1}^{m} log \frac{1}{\sqrt{2\pi}\sigma^{(i)}} - \sum_{i=1}^{m} \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2}$$

The left part is a constant, and the right part need to be minimized. Let $w_i = \frac{1}{2(\sigma^{(i)})^2}$, we get a form same as the weight linear regression.

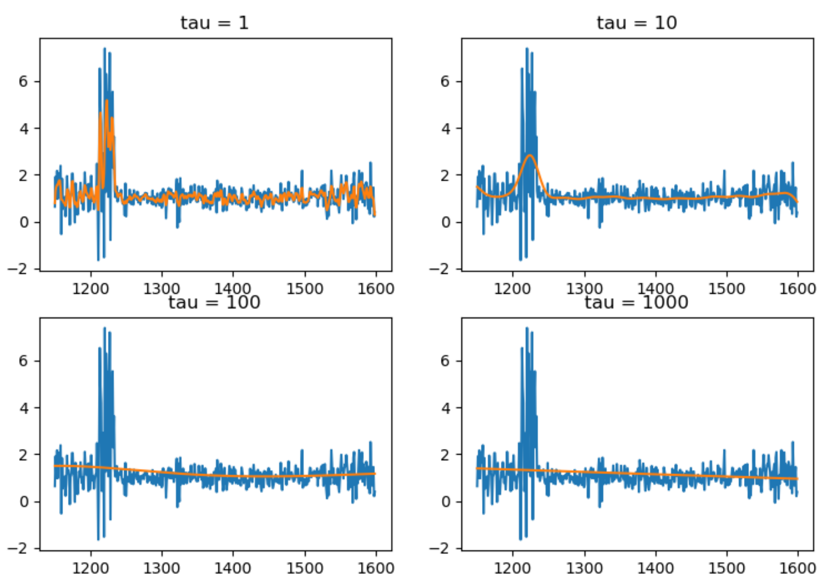
(b) (i) Let $\theta = (X^T X)^{-1} X^T \vec{y}$, we can get the result



(ii) Let $\theta = (X^T W X)^{-1} X^T W \vec{y}$, we can get the result



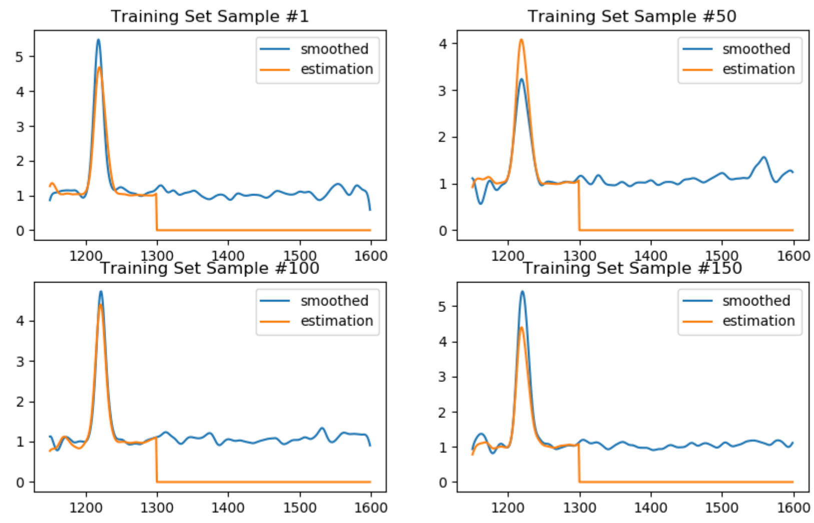
(iii) let τ be different value, we can get different fitting result.



From the figure above, we know that, when $\tau = 1$, we get a perfect fitting result, but it's likely to be overfitting.
As τ grows, the fitting gets worse and worse. So if we want to get a good result, we'd better trade off and select $\tau = 10$ (just an example) to get a balanced performance.

(c) (i) Just apply locally weighted linear regression to all 200 samples

(ii) Through the algorithm described in problem statement, we get the estimation of left part in training set.
The average error is 7.511903.
Here are some examples.[1, 50, 100, 150]



(iii) Through the algorithm described in problem statement, we get the estimation of left part in test set.
Here are some examples.[1, 6, 10, 15]