Stanford CS229 ps1 sol

Shuifan Wang

June 2025

1 Linear Classi ers (logistic regression and GDA)

(a)

$$
\begin{aligned}
\frac{\partial^2}{\partial\theta_j\theta_k}J(\theta) &= -\frac{1}{m}\sum_{i=1}^{m}\frac{\partial}{\partial\theta_j}[x_k^{(i)}(y^{(i)}-g(\theta^T x^{(i)}))] \\
&= \sum_{i=1}^{m}\frac{x_k^{(i)}x_j^{(i)}}{m}[g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))],
\end{aligned}
$$

$$
\begin{aligned}
z^T H z = \sum_k\sum_j z_k H_{kj} z_j &= \sum_k\sum_j\sum_{i=1}^{m}\frac{z_k x_k^{(i)} z_j x_j^{(i)}}{m}[g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))] \\
&= \sum_{i=1}^{m}\frac{[g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))]}{m}\sum_k\sum_j z_k x_k^{(i)} z_j x_j^{(i)} \\
&= \sum_{i=1}^{m}\frac{[g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))]}{m}((x^{(i)})^T z)^2 \geq 0.
\end{aligned}
$$

(b)

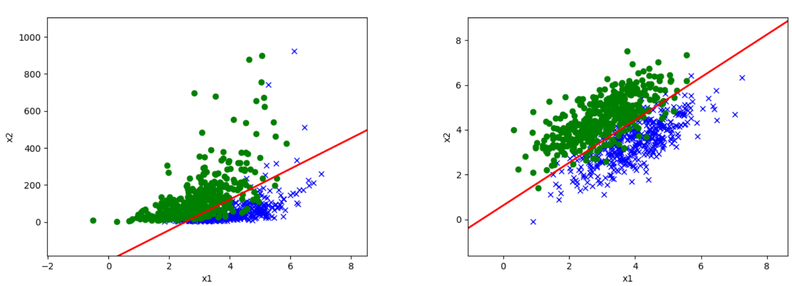

Figure 1: Problem 1(b)

(c)

$$
\begin{aligned}
p(y=1|x;\phi,\mu_0,\mu_1,\Sigma) &= \frac{p(y=1;\phi)p(x|y=1;\mu_0,\mu_1,\Sigma)}{p(y=0;\phi)p(x|y=0;\mu_0,\mu_1,\Sigma)+p(y=1;\phi)p(x|y=1;\mu_0,\mu_1,\Sigma)} \\
&= \frac{\phi\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))}{(1-\phi)\exp(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0))+\phi\exp(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))} \\
&= \frac{1}{1+\exp(\log(1-\phi)-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)-\log\phi+\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1))} \\
&= \frac{1}{1+\exp(-(\theta^Tx+\theta_0))},
\end{aligned}
$$

where

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

(d)

$$
\begin{aligned}
l(\phi,\mu_0,\mu_1,\Sigma) =& \log\prod_{i=1}^m \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp(-\frac{y^{(i)}}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)-\frac{1-y^{(i)}}{2}(x^{(i)} \\
&-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0))\phi^{y^{(i)}}(1-\phi)^{1-y^{(i)}} \\
=& -\frac{nm}{2}\log(2\pi)-\frac{m}{2}\log|\Sigma|+\sum_{i=1}^m(-y^{(i)}\log\phi+(1-y^{(i)})\log(1-\phi) \\
&-\frac{y^{(i)}}{2}(x^{(i)}-\mu_1)^T\Sigma^{-1}(x^{(i)}-\mu_1)-\frac{1-y^{(i)}}{2}(x^{(i)}-\mu_0)^T\Sigma^{-1}(x^{(i)}-\mu_0))
\end{aligned}
$$

$$
\frac{\partial l}{\partial\phi}=0\Rightarrow\frac{\sum_{i=1}^m y^{(i)}}{\phi}=\frac{m-\sum_{i=1}^m y^{(i)}}{1-\phi}\Rightarrow\phi=\frac{\sum_{i=1}^m 1\{y^{(i)}=1\}}{m}.
$$

$$
\frac{\partial l}{\partial\mu_0}=0\Rightarrow\Sigma^{-1}\sum_{i=1}^m(1-y^{(i)})(x^{(i)}-\mu_0)=0\Rightarrow\mu_0=\frac{\sum_{i=1}^m 1\{y^{(i)}=0\}x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)}=0\}}.
$$

Similarly,

$$
\mu_1=\frac{\sum_{i=1}^m 1\{y^{(i)}=1\}x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)}=1\}}.
$$

$$
\frac{\partial l}{\partial\Sigma^{-1}}=0\Rightarrow\frac{m}{2}\Sigma=\frac{1}{2}(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T\Rightarrow\Sigma=\frac{\sum_{i=1}^m(x^{(i)}-\mu_{y^{(i)}})(x^{(i)}-\mu_{y^{(i)}})^T}{m},
$$

Since $\nabla_A\log|A|=A^{-1}$.

(e)

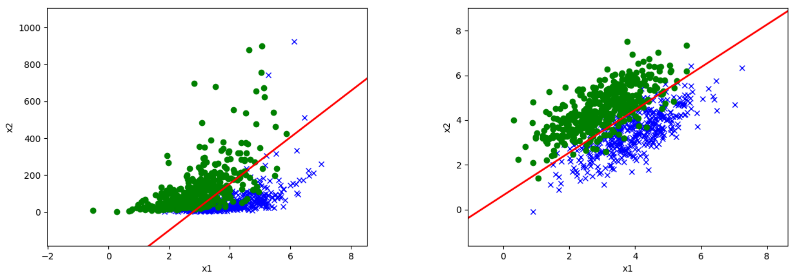

Figure 2: Problem 1(e)

(f) and (g)

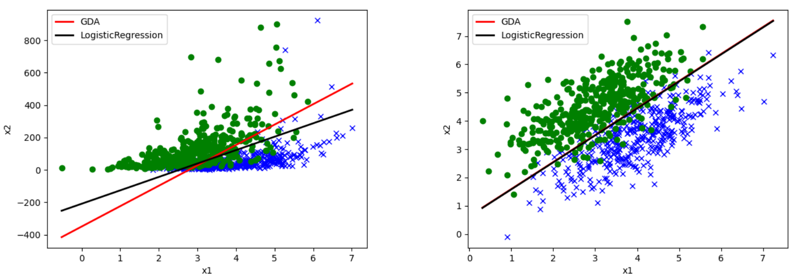

Figure 3: Problem 1(f) and (g)

As for Dataset 2, they have the same performance, since $p(x|y)$ may be Gaussian distribution.

(h) The **Box–Cox transformation** is a class of parameterized power transformations, commonly used to "stretch" skewed data with unequal variances into an approximately normal distribution and stabilize the variance.

p.s. I just run training database.

2 Incomplete, Positive-Only Labels

(a) Since $p(y^{(i)} = 1, t^{(i)} = 1, x^{(i)}) = p(y^{(i)} = 1|t^{(i)} = 1, x^{(i)})p(t^{(i)} = 1|x^{(i)})p(x^{(i)})$, $p(y^{(i)} = 1, t^{(i)} = 1, x^{(i)} = 1) = p(t^{(i)} = 1|y^{(i)} = 1, x^{(i)})p(y^{(i)} = 1|x^{(i)})p(x^{(i)})$,

$$p(y^{(i)} = 1|t^{(i)} = 1, x^{(i)})p(t^{(i)} = 1|x^{(i)}) = p(t^{(i)} = 1|y^{(i)} = 1, x^{(i)})p(y^{(i)} = 1|x^{(i)})$$

We have $p(y^{(i)} = 1|t^{(i)} = 1, x^{(i)}) = p(y^{(i)} = 1|t^{(i)} = 1)$, $p(t^{(i)} = 1|y^{(i)} = 1, x^{(i)}) = 1$. So $p(t^{(i)} = 1|x^{(i)}) = p(y^{(i)} = 1|x^{(i)})/\alpha$, where $\alpha = p(y^{(i)} = 1|t^{(i)} = 1)$.

(b)
$$h(x^{(i)}) \approx p(y^{(i)} = 1|x^{(i)}) = \alpha p(t^{(i)} = 1|x^{(i)}) \approx \alpha.$$

(c) The accuracy on testing set is: 0.9838709677419355.

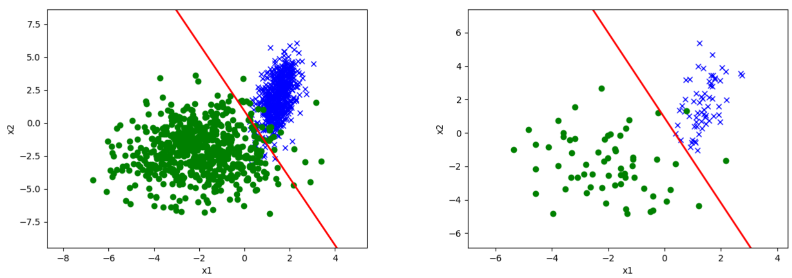

Figure 4: Problem 2(c)

(d) The accuracy on testing set is: 0.5.

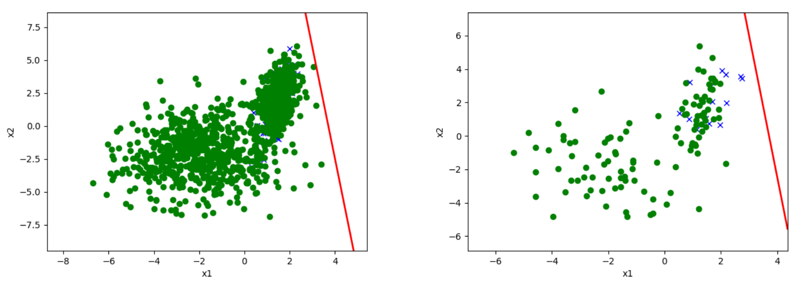

Figure 5: Problem 2(d)

(e) The accuracy on testing set is: 0.967741935483871

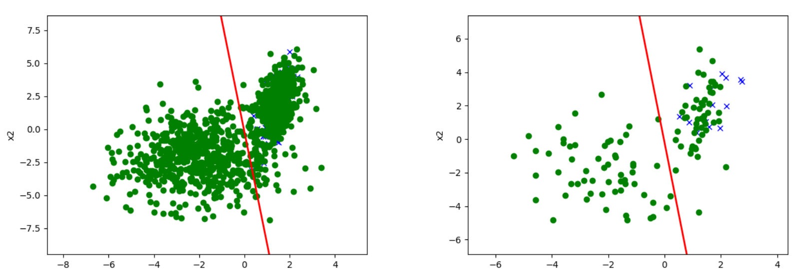

Figure 6: Problem 2(e)

3 Poisson Regression

(a)

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

where $b(y) = 1/y!$, $\eta = \log\lambda$, $T(y) = y$, $a(\eta) = e^\eta$.
(b) $g(\eta) = \mathbb{E}[y;\eta] = e^\eta$.
(c) $h_\theta(x^{(i)}) = \exp(\theta^T x^{(i)})$, therefore $\theta_j := \theta_j + \alpha(y^{(i)} - \exp(\theta^T x^{(i)}))x_j^{(i)}$.
(d)

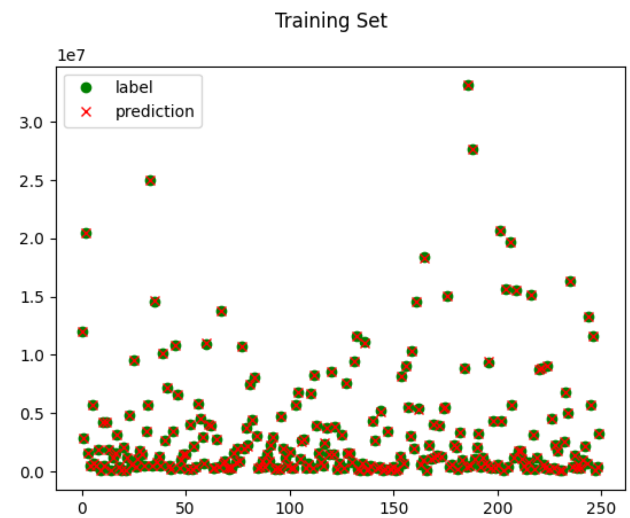

Figure 7: Problem 3(d)

4 Convexity of Generalized Linear Models

(a)

$$\begin{aligned}\frac{\partial}{\partial\eta}\int p(y;\eta)dy &= \int \frac{\partial}{\partial\eta}p(y;\eta)dy \\ &= \int (y - \frac{\partial a(\eta)}{\partial\eta})p(y;\eta)dy \\ &= 0,\end{aligned}$$

$$\mathbb{E}(Y|X;\theta) = \int yp(y;\eta)dy = \int \frac{\partial a(\eta)}{\partial\eta}p(y;\eta)dy = \frac{\partial a(\eta)}{\partial\eta}.$$

(b)

$$
\begin{aligned}
\frac{\partial^2}{\partial \eta^2}\int p(y;\eta)dy &= \frac{\partial}{\partial \eta}\int \frac{\partial}{\partial \eta}p(y;\eta)dy \\
&= \frac{\partial}{\partial \eta}\int (y - \frac{\partial a(\eta)}{\partial \eta})p(y;\eta)dy \\
&= \int ((y - \frac{\partial a(\eta)}{\partial \eta})^2 - \frac{\partial^2 a(\eta)}{\partial \eta^2})p(y;\eta)dy,
\end{aligned}
$$

$$
\mathrm{Var}(Y|X;\theta) = \int (y - \frac{\partial a(\eta)}{\partial \eta})^2 p(y;\eta)dy = \int \frac{\partial^2 a(\eta)}{\partial \eta^2}p(y;\eta)dy = \frac{\partial^2 a(\eta)}{\partial \eta^2}.
$$

(c)

$$
l(\theta) = -\log \prod_{i=1}^{m} p(y^{(i)}|x^{(i)};\theta) = \sum_{i=1}^{m}(a(\theta^T x^{(i)}) - \eta y^{(i)} - \log b(y^{(i)})),
$$

$$
\frac{\partial^2 l(\theta)}{\partial \theta_j \partial \theta_k} = \sum_{i=1}^{m} a''(\theta^T x^{(i)})x_k^{(i)} x_j^{(i)},
$$

$$
z^T H z = \sum_k \sum_j \sum_{i=1}^{m} a''(\theta^T x^{(i)})x_k^{(i)} x_j^{(i)} z_k z_j = \sum_{i=1}^{m} \mathrm{Var}(Y|X = x^{(i)};\theta)((x^{(i)})^T z)^2 \geq 0,
$$

therefore H is PSD.

5 Locally weighted linear regression

(a) i. Let

$$
\begin{aligned}
W &= \frac{1}{2}\mathrm{diag}(w^{(1)}, \dots, w^{(n)}), \\
J(\theta) &= (X\theta - y)^T W (X\theta - y).
\end{aligned}
$$

ii.

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta(\theta^T X^T W X\theta - y^T W X\theta - \theta^T X^T W y + y^T W y) \\
&= \nabla_\theta \mathrm{tr}(\theta^T X^T W X\theta - y^T W X\theta - \theta^T X^T W y + y^T W y) \\
&= 2X^T W^T X\theta - 2X^T W^T y \\
&= 0.
\end{aligned}
$$

So in this weighted setting, $\theta = (X^T W X)^{-1} X^T W y$, since $W^T = W$.
iii.

$$
\begin{aligned}
l(\theta) &= \log \prod_{i=1}^{m} (\frac{1}{\sqrt{2\pi}\sigma^{(i)}} \exp(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2})) \\
&= -\frac{m}{2}\log(2\pi) - \sum_{i=1}^{m}(\log \sigma^{(i)} + \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2}),
\end{aligned}
$$

$$w^{(i)} = \frac{1}{(\sigma^{(i)})^2}.$$

(b) MSE = 0.3305312682137523. The model seems to be underfitting.

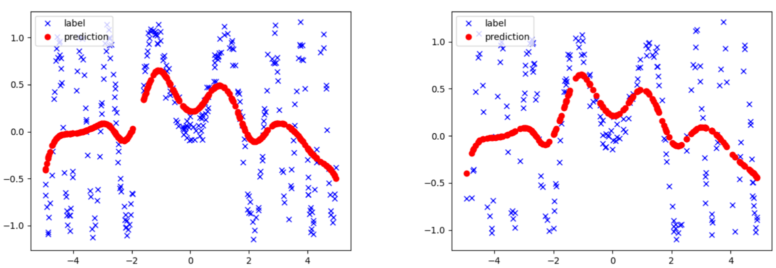

Figure 8: Problem 5(b)

(c) $\tau = 0.05$ achieves the lowest MSE = 0.01240007615046403 on the validation set. MSE = 0.01699014338687814 on the testing set.

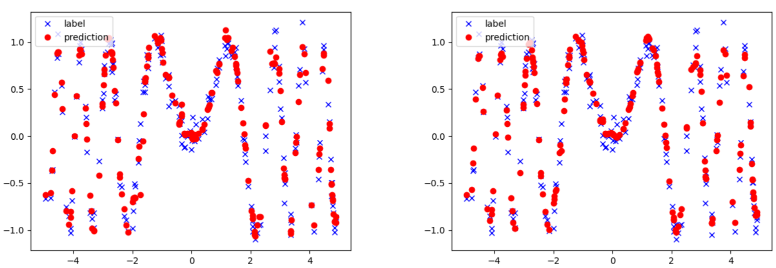

Figure 9: Problem 5(c) $\tau = 0.03$ or 0.05

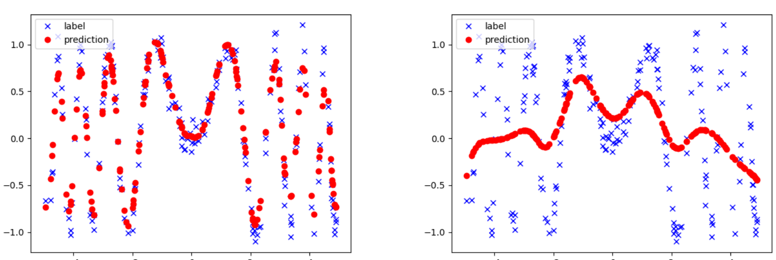

Figure 10: Problem 5(c) $\tau = 0.1$ or 0.5

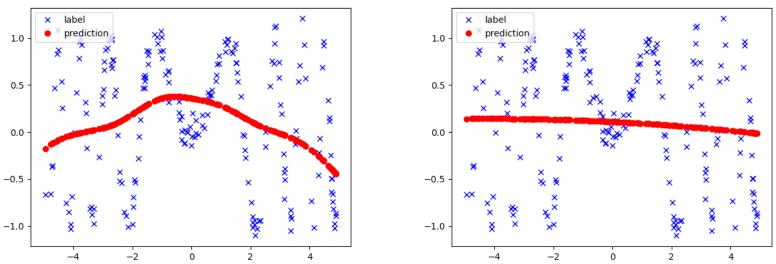

Figure 11: Problem 5(c) $\tau = 1.0$ or 10.0