Question 1:

part a

We will start by taking the first derivative of $J(\theta)$

$$\begin{aligned}
\nabla_\theta J(\theta) &= -\frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)}\frac{g'(\theta^T x^{(i)})}{g(\theta^T x^{(i)})}\cdot x^{(i)} + (1-y^{(i)})\cdot\frac{-g'(\theta^T x^{(i)})}{1-g(\theta^T x^{(i)})}\cdot(1-g(\theta^T x^{(i)}))\cdot x^{(i)}\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)}\frac{g(\theta^T x^{(i)})}{g(\theta^T x^{(i)})}\cdot(1-g(\theta^T x^{(i)}))x^{(i)} + (1-y^{(i)})\cdot\frac{1-g(\theta^T x^{(i)})}{1-g(\theta^T x^{(i)})}\cdot g(\theta^T x^{(i)})\cdot x^{(i)}\right] \\
&= -\frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)} - y^{(i)}g(\theta^T x^{(i)}) - g(\theta^T x^{(i)}) + y^{(i)}g(\theta^T x^{(i)})\right]\cdot x^{(i)} \\
&= -\frac{1}{n}\sum_{i=1}^{n}\left[y^{(i)} - g(\theta^T x^{(i)})\right]\cdot x^{(i)}
\end{aligned}$$

Then we take the Hessian of the loss function

$$\begin{aligned}
H = \nabla_\theta^2 J(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\nabla_\theta g(\theta^T x^{(i)})\cdot\nabla_\theta(\theta^T x^{(i)})\cdot x^{(i)} \\
&= \frac{1}{n}\sum_{i=1}^{n} g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))x^{(i)}[x^{(i)}]^T
\end{aligned}$$

Given that $h_\theta(x) = g(\theta^T x)$, for any vector $z \in R^d$, define $x \in R^n$

$$\begin{aligned}
z^T(Hz) &= z^T\cdot\begin{pmatrix} z^{(1)}\cdot H \\ z^{(2)} \\ \cdot \\ \cdot \\ \cdot \\ z^{(d)}\cdot H \end{pmatrix} \\
&= \sum_{j=1}^{d}[z^{(j)}]^T z^{(j)}\cdot\frac{1}{n}\sum_{i=1}^{n}\left[x^{(i)}[x^{(i)}]^T g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))\right] \\
&= \sum_{j}[z^{(j)}]^T z^{(j)} x^T x\cdot\frac{1}{n}\sum_{i=1}^{n}\left[h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))\right]
\end{aligned}$$

Since $x^T x$ is a scalar, define $z \in R^d$, $\sum_j [z^{(j)}]^T z^{(j)} x^T x$ can be written as

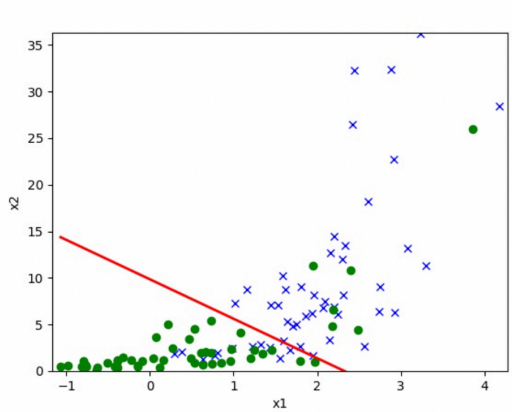
$$(x^T x)\cdot(z^T z) = z^T(x^T x)z \quad = (x^T z)^T(x^T z) = (x^T z)^2$$

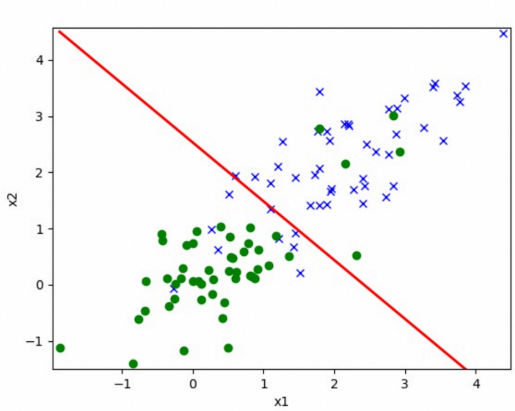
So, the quadratic form becomes

$$zH(\theta)z = (x^T z)^2\cdot\frac{1}{n}\sum_{i=1}^{n}\left[h_\theta(x^{(i)})(1-h_\theta(x^{(i)}))\right]$$

Observe that all the terms are positive because $h(\theta)$ is a possibility function and $(x^T z)^2 \geq 0$.
We conclude that matrix $H$ is positive semi-definite.

part b :




Dataset 1 Dataset 2

Logistic Regression Results given two datasets.

Part C.

simplify $P(y=1 \mid x; \phi, \mu_0, \mu_1, \Sigma)$

$$= \frac{P(x=x \mid y=1)\, P(y=1)}{P(x=x \mid y=1)\, P(y=1) + P(x=x \mid y=0)\, P(y=0)}$$

$$= \frac{1}{1 + \frac{P(x=x \mid y=0)\, P(y=0)}{P(x=x \mid y=1)\, P(y=1)}}$$

Take a closer look at the function in the denominator :

$$\frac{P(x=x \mid y=0)\, P(y=0)}{P(x=x \mid y=1)\, P(y=1)} = \frac{\exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0)\right)}{\exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\right)} \cdot \frac{1-\phi}{\phi}$$

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

as both the nominator and denominator share the common constant factor $\left[(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}\right]^{-1}$

As $x-\mu_0 = x-\mu_1+\mu_1-\mu_0$

The exponent term $\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1) - \frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)$ can be written as

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

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

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

Hence, the above fraction equals

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

if we let $\theta^T = (\mu_1-\mu_0)^T\Sigma^{-1} \Rightarrow \theta = \left[(\mu_1-\mu_0)^T\Sigma^{-1}\right]^T$

$$= \left(\Sigma^{-1}\right)^T(\mu_1-\mu_0)$$

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

If we set the threshold at 0.5, the decision boundary will have the equation $\theta^T x + \theta_0 = 0$

part d.

Log-likelihood function: $\ell(\phi, \mu_0, \mu_1, \Sigma) = \log \prod_{i=1}^{n} p(x^{(i)}|y^{(i)})\, p(y^{(i)})$

$$= \sum_{i=1}^{n}\left[\log p(x^{(i)}|y^{(i)}) + \log p(y^{(i)})\right]$$

$$= \sum_{i=1}^{n}\left[-\log(2\pi)^{d/2} - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x^{(i)}-\mu_{y^{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}}) + y^{(i)}\log\phi + (1-y^{(i)})\log(1-\phi)\right]$$

$\hat{\phi}$ is achieved by $\nabla_\phi \ell(\phi, \mu_0, \mu_1, \Sigma) = 0$.

$$\nabla_\phi \ell(\phi, \mu_0, \mu_1, \Sigma) = \sum_{i=1}^{n}\left[\frac{y^{(i)}}{\phi} - \frac{1-y^{(i)}}{1-\phi}\right] = 0.$$

$$\Rightarrow \sum_{i=1}^{n}\phi = \sum_{i=1}^{n} y^{(i)}$$

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

$$\nabla_{\mu_0} \ell(\theta) = \nabla_{\mu_0} \sum_{i=1}^{n} \left[ -\frac{1}{2} \{y^{(i)}=0\} (x^{(i)}-\mu_0)^T \Sigma^{-1} (x^{(i)}-\mu_0) \right]$$
$$= \sum_{i=1}^{n} \left[ -\Sigma^{-1}(x^{(i)}-\mu_0) \times (-1) \right] \qquad (\nabla X^T A X = 2AX)$$
$$= \Sigma^{-1} \sum_{i=1}^{n} \{y^{(i)}=0\} (x^{(i)}-\mu_0)$$

let $\nabla_{\mu_0} \ell(\theta) = 0$ we obtain

$$\Sigma^{-1} \cdot \sum_{i=1}^{n} x^{(i)} = \Sigma^{-1} \cdot \sum_{i=1}^{n} \hat{\mu}_0$$
$$\hat{\mu}_0 = \frac{\sum_{i=1}^{n} 1\{y^{(i)}=0\} x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)}=0\}}$$

Similarly, taking the gradients w.r.t. $\mu_1$, we obtain

$$\nabla_{\mu_1} \ell(\theta) = \Sigma^{-1} \sum_{i=1}^{n} \{y^{(i)}=1\} (x^{(i)}-\mu_1)$$

let $\nabla_{\mu_1} \ell(\theta) = 0$, there is

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} 1\{y^{(i)}=1\} x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)}=0\}}$$

Claim that for any vector $z \in \mathbb{R}^d$, $\nabla_\Sigma z^T \Sigma^{-1} z = -\Sigma^{-1} z z^T \Sigma^{-1}$

$$\frac{d}{d\Sigma}(z^T \Sigma^{-1} z) = \frac{d}{d\Sigma} Tr[z^T \Sigma^{-1} z] = Tr\left[z z^T \frac{d\Sigma^{-1}}{d\Sigma}\right] = Tr\left[-z z^T \Sigma^{-1} (\Sigma)' \Sigma^{-1}\right]$$
$$= Tr\left[-\Sigma^{-1} z z^T \Sigma^{-1} (\Sigma)'\right]$$

Given $\Sigma\Sigma^{-1} = I$, we have $\frac{d}{d\Sigma}(\Sigma^{-1}) = -\Sigma^{-1}(\Sigma)'\Sigma^{-1}$

Hence $\nabla_\Sigma (z^T \Sigma^{-1} z) = -\Sigma^{-1} z z^T \Sigma^{-1}$

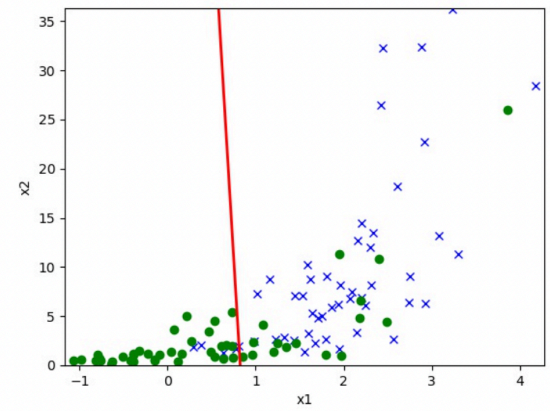
let $z = (x^{(i)} - \mu_{y^{(i)}})$ as $x^{(i)} - \mu_{y^{(i)}}$ is a vector

$$\nabla_\Sigma \ell(\theta) = \nabla_\Sigma \sum_{i=1}^{n} \left( -\frac{1}{2} \log|\Sigma| - \frac{1}{2} \boxed{(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}})} \right)$$
$$= \sum_{i=1}^{n} \left( -\frac{1}{2}\Sigma^{-1} + \frac{1}{2}(x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} \right) \qquad [\nabla_\Sigma \log|\Sigma| = (\Sigma^{-1})^T]$$
$$= 0.$$

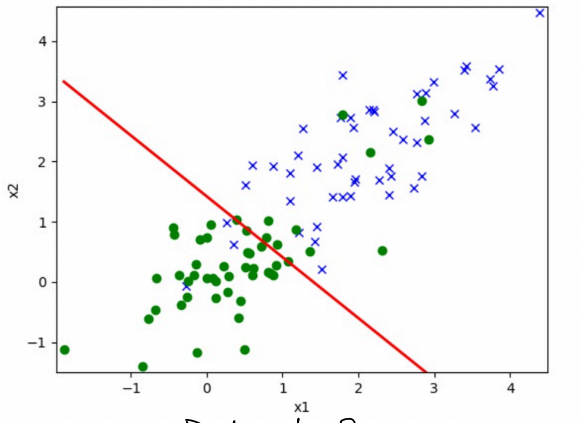
So, we have $n - \sum_{i=1}^{n} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} = 0$

$$\Rightarrow \hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$

part e:



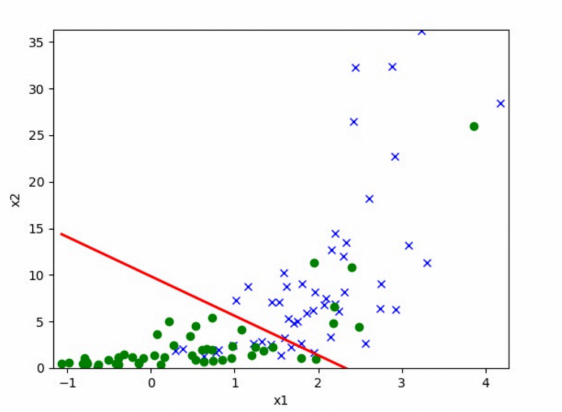

Dataset 1



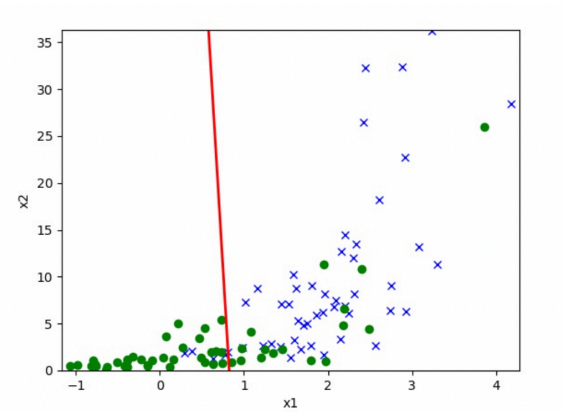

Dataset 2

GDA Results on two given datasets

part f:



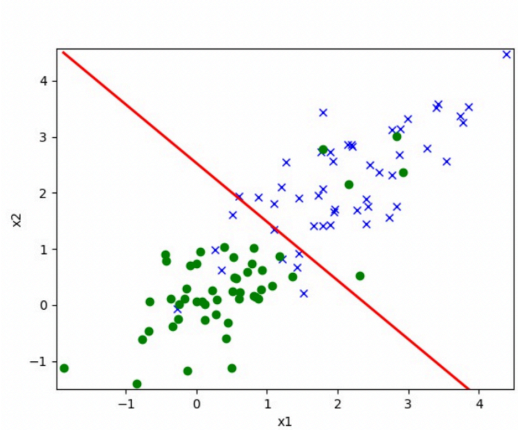

Logistic Regression



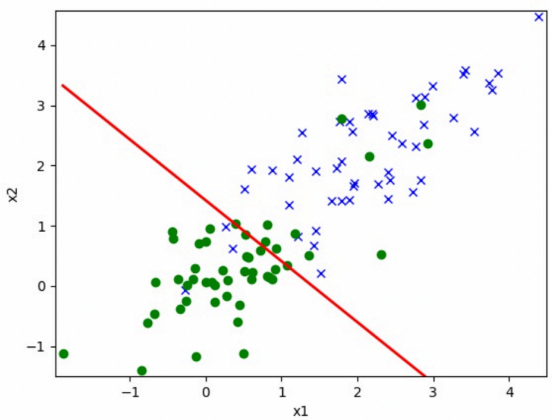

GDA.

for dataset 1, Logistic Regression obtains a better decision boundary than GDA, so the classification model using logistic regression yields more accurate results:

part g




Logistic Regression




GDA.

GDA performance in validation set 2 is similar to the logistic regression. But in dataset 1, GDA's performance is worse than Logistic Regression.

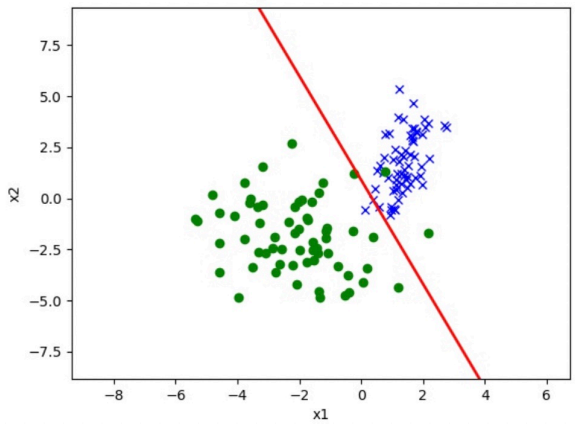
GDA makes more specific assumptions about the dataset than LR. [Eg. Gaussian DIST] when Dataset is gaussian dist, GDA better.

part h.

We can apply the power transform by taking the square root and the log. to $x^{(i)}$, so that the transformed data is gaussian distributed.

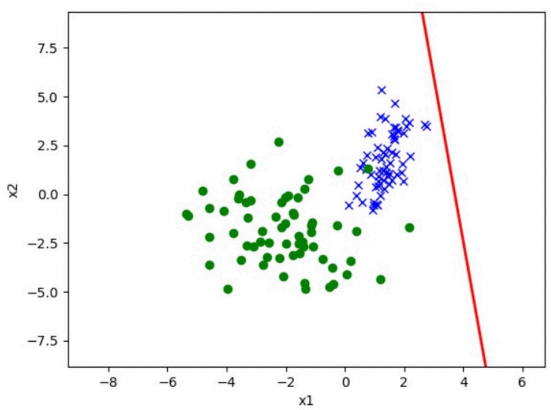
Question 2:

part a:




fully observed result on the test data

part b:




naive method on partial labels

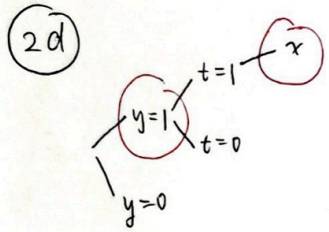
part c

since $\forall x^{(i)}$, $p(y^{(i)}=1 \mid t^{(i)}=0, x) = 0$

$$p(t^{(i)}=1 \mid y^{(i)}=1, x^{(i)}) = \frac{p(y^{(i)}=1 \mid t^{(i)}=1, x)\, p(t^{(i)}=1 \mid x)}{p(y^{(i)}=1 \mid t^{(i)}=1, x)\, p(t^{(i)}=1 \mid x) + p(y^{(i)}=1 \mid t^{(i)}=0, x)\, p(t^{(i)}=0 \mid x)}$$

$$= 1$$

2d




$$p(y^{(i)}=1 \mid x^{(i)}) = p((y^{(i)}=1, x^{(i)}, t^{(i)}=1) \text{ or } (y^{(i)}=1, x^{(i)}, t^{(i)}=0))$$

$$= p(y^{(i)}=1 \mid x^{(i)}, t^{(i)}=1)\, p(t^{(i)}=1 \mid x^{(i)}) + p(y^{(i)}=1 \mid t^{(i)}=0, x^{(i)})\, p(t^{(i)}=0 \mid x^{(i)})$$

$$= \alpha\, p(t^{(i)}=1 \mid x^{(i)}) + 0.$$

$$\therefore p(t^{(i)}=1 \mid x^{(i)}) = \frac{1}{\alpha}\, p(y^{(i)}=1 \mid x^{(i)})$$

Part e

we will show that $h(x^{(i)}) = \alpha$ when $y^{(i)}=1$, and $h(x^{(i)})=0$ when $y^{(i)}=0$.

$h(x^{(i)}) =$

$p(y^{(i)}=1 \mid x^{(i)})$ from 2d, we obtain that

$$p(y^{(i)}=1 \mid x^{(i)}) = \alpha \cdot p(t^{(i)}=1 \mid x^{(i)}) + 0 \cdot p(t^{(i)}=0 \mid x^{(i)})$$

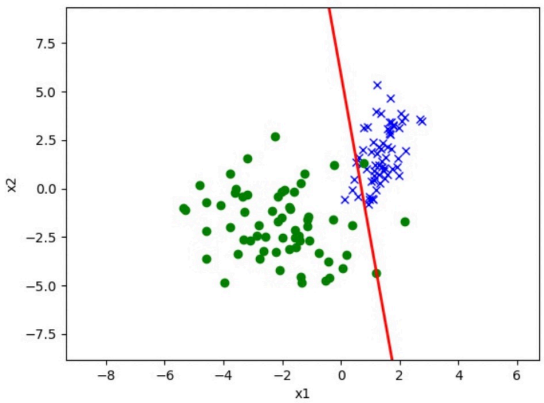
Given the assumption $p(t^{(i)}=1 \mid x^{(i)})=1$ and $p(t^{(i)}=0 \mid x^{(i)})=0$.

$$h(x^{(i)}) = \alpha \cdot 1$$

Hence, given $y^{(i)}=1$, $h(x^{(i)})$ always take the value of $\alpha$, ~~so~~ expected value $= \alpha$.

Therefore, $E(h(x^{(i)}) \mid y^{(i)}=1) = \alpha$

part f




Rescaled result on the test set.

Question 3

Part a:

we can write the Poisson Distribution in the form

$$p(y;\eta) = b(y)\exp(\eta T(y) - a(\eta)) \quad \text{s.t.}$$

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

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

part b.

Because $E[y;\eta] = \lambda = e^{\eta}$, so the cannonical response function is

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

part c

$\ell(\theta) = \log p(y^{(i)} | x^{(i)}; \theta)$, picking the $j$th $x^{(i)}$, $y^{(i)}$ and given that $\theta^T x = \eta$

$$\frac{\partial}{\partial\theta_j}\ell(\theta) = \frac{\partial}{\partial\theta_j}\log\left[\frac{1}{y!}\exp(\eta\cdot y - e^{\eta})\right] = \frac{\partial}{\partial\theta_j}\log\left[\frac{1}{y!}\exp(y\theta^T x^{(i)} - e^{\theta^T x^{(i)}})\right]$$

$$= \frac{\partial}{\partial\theta_j}\left[\log\frac{1}{y!} + y\cdot\theta^T x^{(i)} - e^{\theta^T x}\right]$$

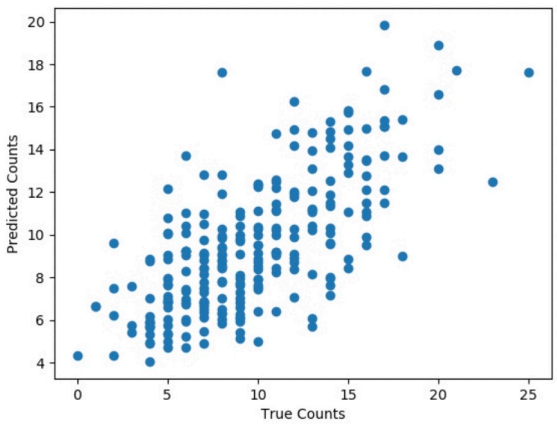
$$= y x^{(j)} - e^{\theta^T x^{(i)}}\cdot x^{(j)}$$

$$= \left(y - e^{\theta^T x^{(i)}}\right) x^{(j)}$$

Hence, the SGD update rule for the GLM is that

$$\theta_{j+1} = \theta_j + \alpha\left(y - \theta^T x^{(i)}\right) x^{(j)} \quad \text{for any } j$$

Part d.




Learned Poisson result on the validation set

4. Note that (1) $\int_{-\infty}^{\infty} p(y;\eta)\,dy = 1$ and (2) $\int_{-\infty}^{\infty} \frac{\partial^k}{\partial \eta^k} \cdot p(y;\eta)\,dy = 0$. for $k = 1, 2, \cdots$

Hence we have

$$\int p'(y;\eta)\,dy = 0 \quad \text{and} \quad \int p''(y;\eta)\,dy = 0$$

(3) Let $\ell = \log p(y;\eta)$, then

$$\ell' = \frac{p'}{p} \quad \text{and} \quad \ell'' = -\left(\frac{p'}{p}\right)^2$$

(4) $E(\ell') = \int \ell'\, p(y;\eta)\,dy = \int \frac{p'(y;\eta)}{p(y;\eta)} \cdot p(y;\eta)\,dy$ (By (3))

$= \int p'(y;\eta)\,dy$ (By (2))

$= 0$

(5) $E(\ell'') = \int p''(y;\eta)\,dy - E\left[\left(\frac{p'(y;\eta)}{p(y;\eta)}\right)^2\right]$

$= -E\left[(\ell')^2\right]$

$= -E\left[(\ell')^2\right] - \left[E(\ell')\right]^2$

$= -\mathrm{Var}(\ell')$

(4a)

Because $\eta = \eta^T$, $\eta \in \mathbb{R}$

$\ell(\eta) = \log p(y;\eta) = \log b(y) + \eta T(y) - a(\eta)$

$\ell'(\eta) = T(y) - a'(\eta)$.

$E(\ell') = E(T(y) - a'(\eta)) = E(T(y)) - a'(\eta)$

$= 0$

$\therefore E(T(y)) = a'(\eta)$

(4b) $\ell''(\eta) = -a''(\eta)$

$E(\ell'') = -\mathrm{Var}(\ell')$

So, $\mathrm{Var}(T(y) - a'(\eta)) = -E(-a''(\eta))$

$\mathrm{Var}(T(y)) = a''(\eta)$

($a(\eta)$ is constant w.r.t $T(y)$)

4C

The NLL of the distribution is written as :

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log p(y^{(i)};\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(\log(b(y^{(i)})) + y^{(i)}\theta^T x^{(i)} - a(\theta^T x^{(i)})\right)$$

Take the gradient w.r.t. $\theta$ :

$$\nabla_\theta J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)}x^{(i)} - a'(\theta^T x^{(i)})\cdot x^{(i)}\right)$$

Then take the hessian : $H_{J(\theta)}$

$$H_{J(\theta)} = -\frac{1}{n}\sum_{i=1}^{n}\left(-a''(\theta^T x^{(i)})\, x^{(i)}\cdot[x^{(i)}]^T\right)$$

we already know that $a''(\theta^T x^{(i)}) = a''(\eta) = \mathrm{Var}(y^{(i)}|x^{(i)})$ and $E(t''(\eta)) = \mathrm{Var}(t')$

$= \mathrm{Var}(Y)$ by part b.

$= \mathrm{Var}(Y|X;\theta)$

Hence, $H_{J(\theta)} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(y^{(i)}|x^{(i)})\, x^{(i)}[x^{(i)}]^T$

consider any vector $a \in \mathbb{R}^d$,

$$a^T H_{J(\theta)} a = a^T \frac{1}{n}\sum_{i=1}^{n}\mathrm{var}(y^{(i)}|x^{(i)})\, x^{(i)}[x^{(i)}]^T \cdot a$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathrm{var}(y^{(i)}|x^{(i)})\,[a^T x^{(i)}]^2$$

since $\mathrm{var}(y^{(i)}|x^{(i)}) \geq 0$ for $\forall i$ and $[a^T x^{(i)}]^2$ is always positive.

$a^T H_{J(\theta)} a \geq 0$ i.e. The hessian of the loss is positive semi-definite.

which concludes that NLL loss of GLM is convex.

Question 5

(5a) The objective function is written as

$$J(\theta) = \sum_{i=1}^{n} \left(\theta^T \hat{x}^{(i)} - y^{(i)}\right)^2$$

And the update gradient descent equation is

~~$\theta_{t+1} = \theta$~~

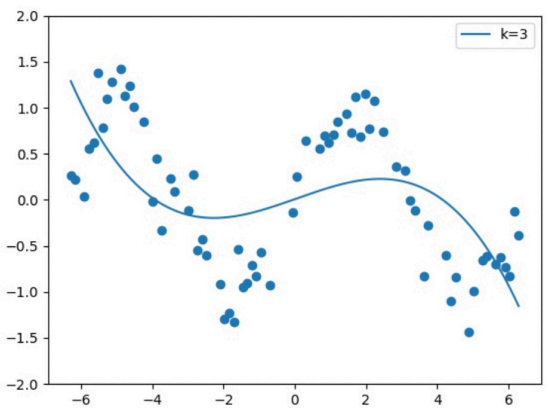
$$\theta^{t+1} = \theta^t - \alpha' \nabla_\theta J(\theta)$$

$$= \theta^t - \alpha' \sum_{i=1}^{n} 2\left(\theta^T \hat{x}^{(i)} - y^{(i)}\right) \hat{x}^{(i)}$$

for some $\alpha$, there is the update rule =

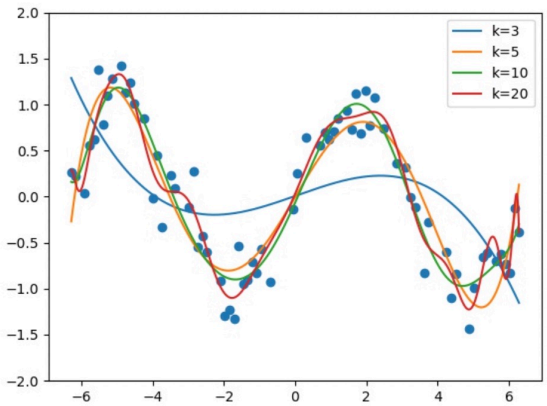
$$\theta^{t+1} = \theta^t - \alpha \sum_{i=1}^{n} \left(\theta^T \hat{x}^{(i)} - y^{(i)}\right) \hat{x}^{(i)}$$

part b:




degree-3 Polynomial Regression

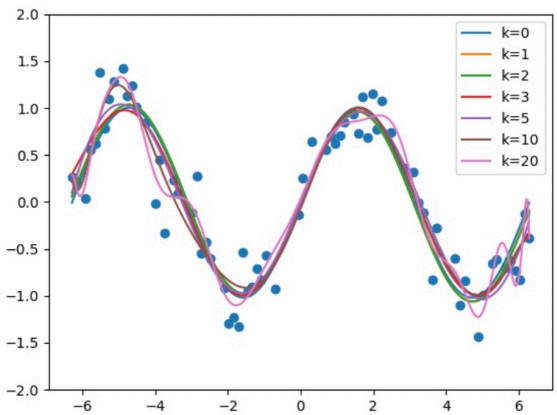
Part c:




Degree- k Polynomial Regression

The model fits the dataset better as k increases. However at k=20 onward, the model starts to overfit the data.

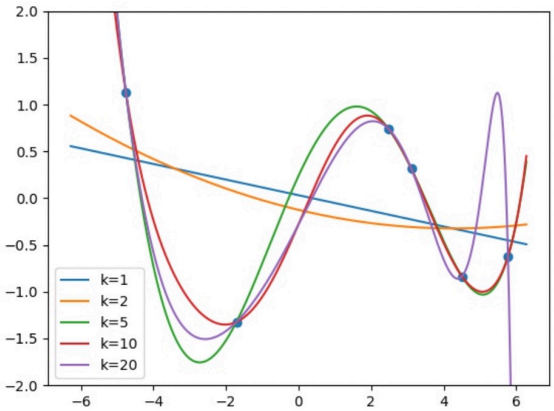
Part d:




Degree- k Polynomial Regression with Sine-feature-map.

The model fits the training data perfectly at lower number of k. As k increases, the model tends to overfit the dataset

Part e:




Overfitting with small data

Since there are only 6 sets of data in this case. The model fits the data into sine feature map when $k \geq 5$. However, due to the limited amount of training data, when the degree increases, the model does not confirm to the sine shape but some strange and asymmetric shape.
It shows the result of overfitting.