PS1 - 1a

Friday, July 8, 2022 12:54 AM

(1) a) $\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{m}\sum_{i=1}^{m}\left(y^{(i)} - g(\theta^T x^{(i)})\right)x_j^{(i)}$

$\frac{\partial^2 J(\theta)}{\partial \theta_j \partial \theta_k} = -\frac{1}{m}\sum_{i=1}^{m} g(\theta^T x^{(i)})(1-(\theta^T x^{(i)}))x_j^{(i)} x_k^{(i)}$

For any z,

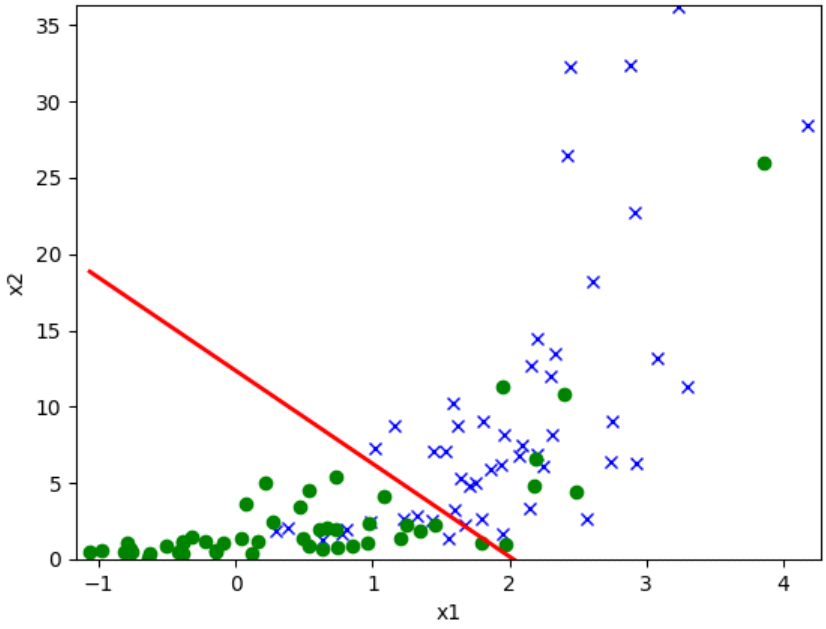
$z^T H z = \frac{1}{m}\sum_{i=1}^{m}\sum_{j,k=1} g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))\, x_j^{(i)} x_k^{(i)} z_j z_k$

$= \frac{1}{m}\sum_{i=1}^{m}\sum_{j,k=1} g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)}))\,(x^{(i)T} z)^2 \geq 0$

PS1 - 1b

Friday, July 8, 2022 7:48 PM

b)



PS1 - 1c

Friday, July 8, 2022 7:50 PM

c) $$p(y=1 \mid x, \phi, \mu_0, \mu_1, \Sigma) = \frac{p(x \mid y=1)\, p(y=1)}{p(x \mid y=1)\, p(y=1) + p(x \mid y=0)\, p(y=0)}$$

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

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

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

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

PS1 - 1d

Friday, July 8, 2022 7:57 PM

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

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

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

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

$\frac{\partial \ell}{\partial \phi} = \sum_{i=1}^{n} y^{(i)} \frac{1}{\phi} + (1-y^{(i)}) \frac{1}{1-\phi} = \frac{\sum_{i=1}^{n} 1(y^{(i)}=1)}{\phi} + \frac{m - \sum_{i=1}^{n} 1(y^{(i)}=1)}{1-\phi} = 0$

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

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

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

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

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

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

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

By symmetry, $\mu_1 = \frac{\sum_{i=1}^{n} (y^{(i)}=1) x^{(i)}}{\sum_{i=1}^{n} (y^{(i)}=1)}$

$\varepsilon = \sigma^2$

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

$n \quad (x^{(i)} \quad \mu_0 \quad 1 \quad (x^{(i)} - \mu_1)^2 \Big) \Big\} \Big]$

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

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

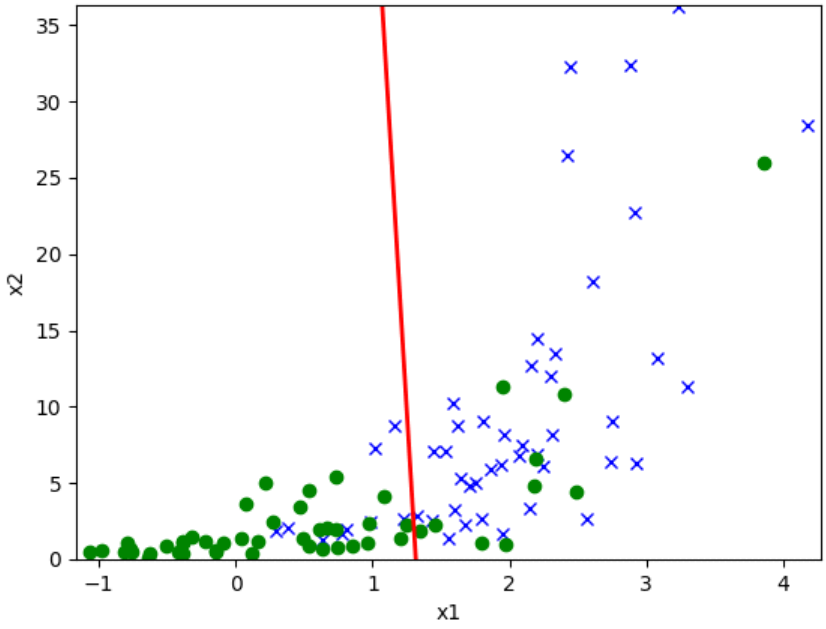
$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} 1\{y^{(i)}=0\}(x^{(i)}-\mu_0)^2 + \sum_{i=1}^{n} 1\{y^{(i)}=1\}(x^{(i)}-\mu_1)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n} (x^{(i)}-\mu_{y^{(i)}})^2$$

PS1 - 1e

Friday, July 8, 2022 8:11 PM

e)

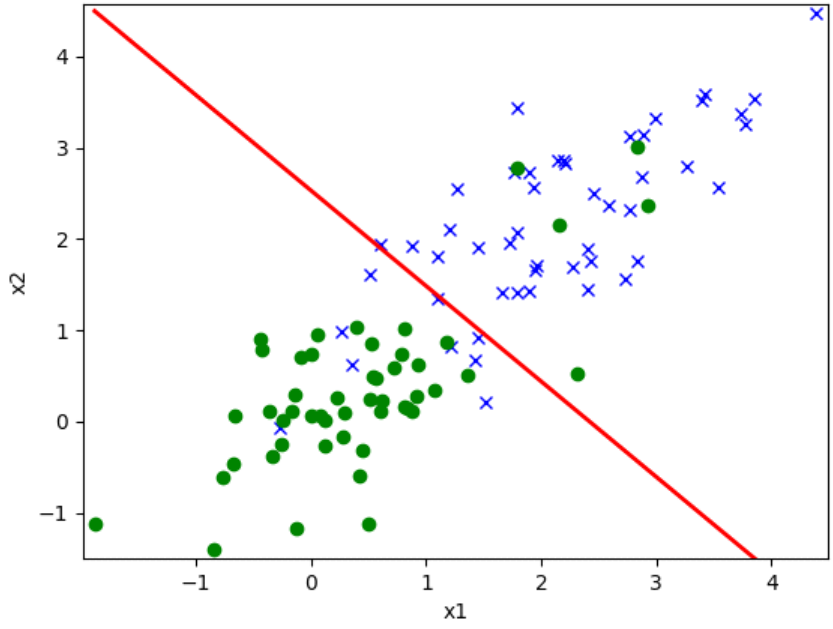


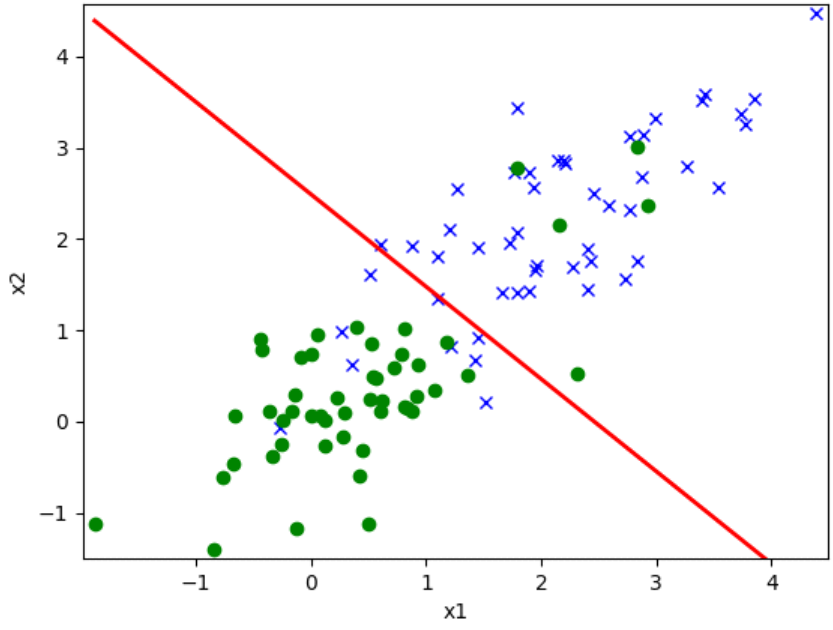
PS1 - 1fgh

Friday, July 8, 2022 8:11 PM

f) For dataset 1, the decision boundaries for logistic regression and GDA are similar, however for logistic regression the line is more diagonal whereas for GDA it is closer to vertical. GDA has slightly more green dots that are misclassified whereas logistic regression has a few more blue data points that are misclassified. Overall, logistic regression appears to perform slightly better

g) Logistic Regression





GDA

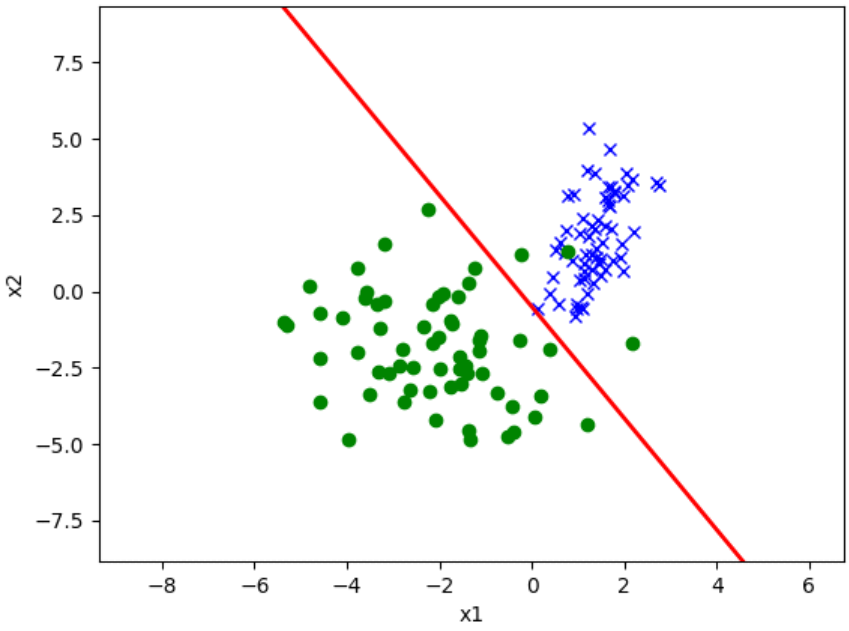
For dataset 2, the two decision boundaries are nearly identical. For dataset 1, GDA appears to perform slightly worse than logistic regression as there are a few more green dots that are misclassified. This may be because the distribution of the data is not Gaussian, which is one of the assumptions made by GDA.

h) Any transformation which causes the data to become closer to Gaussian distributed would work. An example may be taking the logarithm of the data.

PS1 - 2a

Friday, July 8, 2022 8:11 PM

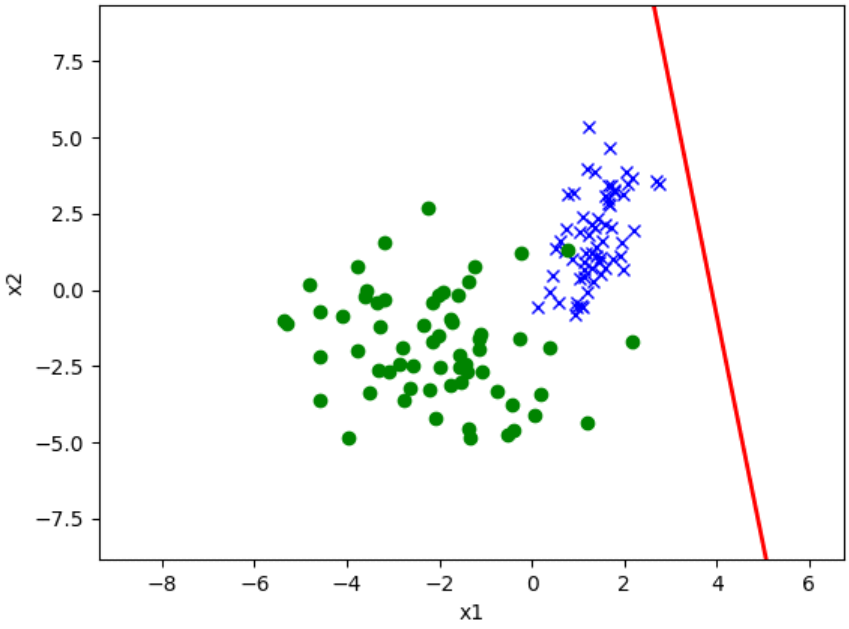
2. a)



PS1 - 2b

Friday, July 8, 2022 8:11 PM

b)



PS1 - 2c

Friday, July 8, 2022

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

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

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

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

$$= 1$$

PS1 - 2d

Saturday, July 9, 2022 1:55 AM

d) $p(y^{(i)}=1 \mid x^{(i)}) = p(y^{(i)}=1 \wedge t^{(i)}=1 \mid x^{(i)})$ from c)

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

$= p(t^{(i)}=1 \mid x^{(i)})\, c$

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

PS1 - 2e

Saturday, July 9, 2022 1:55 AM

e) Let $g(x) = p(t^{(i)} = 1 \mid x^{(i)})$

From d, $h(x^{(i)}) = a\, g(x^{(i)})$

$$E[h(x^{(i)}) \mid y^{(i)} = 1] = E(h(x^{(i)}) \mid y^{(i)} = 1]$$

$$= \frac{E[h(x^{(i)}) 1\{y^{(i)} = 1\})}{p(y^{(i)} = 1)}$$

$$= \frac{E\left[E[1\{y^{(i)} = 1\} \mid x^{(i)}] \cdot h(x^{(i)})\right]}{E[p(y^{(i)} = 1 \mid x^{(i)})]}$$

$$= \frac{E[h(x^{(i)}) \cdot h(x^{(i)})]}{E(h(x^{(i)})}$$

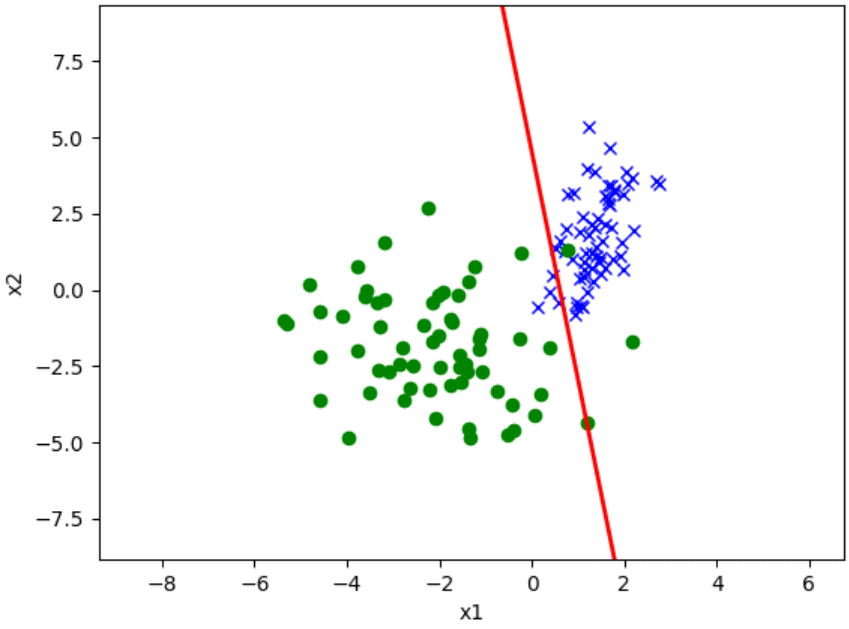
$$= a \frac{E[g(x^{(i)}) \cdot g(x^{(i)}))}{E[g(x^{(i)})]}$$

$$= a$$

PS1 - 2f

Saturday, July 9, 2022 1:56 AM

f)



PS1 - 3abc

Friday, July 8, 2022 9:06 PM

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

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

$\eta = \log\lambda$

$T(y) = t$

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

b) $g(\eta) = E[y;\eta] = \lambda = e^{\eta}$

c) $\ell(\theta) = \sum_{i=1}^{m} \log p(y^{(i)}|x^{(i)};\theta)$

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

$\frac{\partial \ell(\theta)}{\partial \theta_j} = \sum_{i=1}^{m} y^{(i)} x_j^{(i)} - x_j^{(i)} g(\theta^T x^{(i)})$

$= \sum_{i=1}^{m} (y^{(i)} - g(\theta^T x^{(i)})) x_j^{(i)}$

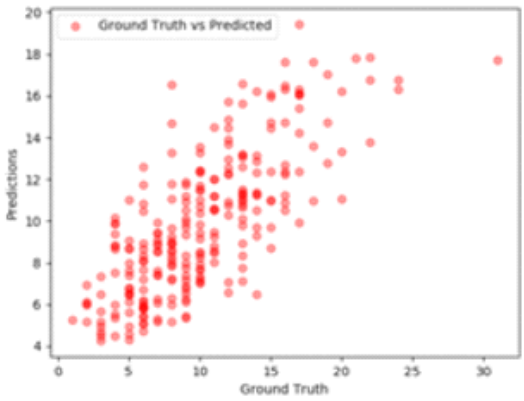
The update rule is $\theta_j := \theta_j + \alpha(y^{(i)} - g(\theta^T x^{(i)}))x_j^{(i)}$

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

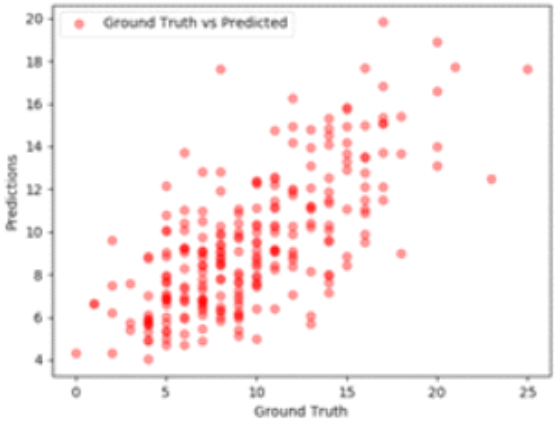
PS1 - 3d

Saturday, July 9, 2022 2:11 AM

d)



Test set



Validation set

PS1 - 4a

Saturday, July 9, 2022 2:13 AM

(4) a) $\frac{\partial}{\partial \eta} p(y;\eta) = p(y;\eta)\, y - \frac{\partial}{\partial \eta} a(\eta)$

$\int \frac{\partial}{\partial \eta} p(y;\eta)\, dy = \int p(y;\eta)\, y - \frac{\partial}{\partial \eta} a(\eta)\, dy$

$\frac{\partial}{\partial \eta} \int p(y;\eta)\, dy = \int y\, p(y;\eta)\, dy - \int p(y;\eta) \frac{\partial}{\partial \eta} a(\eta)\, dy$

$\frac{\partial}{\partial \eta} 1 = \mathbb{E}[Y;\eta] - \frac{\partial}{\partial \eta} a(\eta) \int p(y;\eta)\, dy$

$\mathbb{E}[Y;\eta] = \frac{\partial}{\partial \eta} a(\eta)$

PS1 - 4b

Saturday, July 9, 2022 3:39 PM

b) $\frac{\partial^2}{\partial \eta^2} a(\eta) = \frac{\partial}{\partial \eta} \mathbb{E}[Y|x;\theta] = \frac{\partial}{\partial \eta} \int y\, p(y;\eta)\, dy$

$= \int y\, p(y;\eta)\, y - \frac{\partial}{\partial \eta} a(\eta)\, dy$

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

$= \mathbb{E}[Y^2;\eta] - \mathbb{E}[Y;\eta]^2$

$= \mathrm{Var}(Y;\eta)$

PS1 - 4c

Saturday, July 9, 2022 3:40 PM

c) $l(\theta) = -\log[p(y;\eta)]$

$= -\log[b(y)\exp(\eta y - a(\eta))]$

$= a\eta - \eta y + C$

$= a(\theta^T x) - \theta^T x y + C$

$\nabla_\theta l(\theta) = \frac{\partial}{\partial \eta} a(\eta) \nabla_\theta \eta - yx$

$= \frac{\partial}{\partial \eta} a(\eta) x - yx$

$\nabla_\theta^2 l(\theta) = \nabla_\theta(\nabla_\theta l(\theta))$

$= \nabla_\theta \frac{\partial}{\partial \eta} a(\eta) x - yx$

$= x \nabla_\theta \frac{\partial}{\partial \eta} a(\eta)$

$= x \frac{\partial}{\partial \eta} \frac{\partial}{\partial \eta} a(\eta) \nabla_\theta \eta$

$= \frac{\partial^2}{\partial \eta^2} a(\eta) x x^T$

$= \mathrm{Var}(y;\eta) x x^T$

For any θ, $\mathrm{Var}(y;\eta)$ is positive, and xx^T is PSD, therefore the Hessian $\nabla_\theta^2 l(\theta) = \mathrm{Var}(y;\eta) x x^T$ of the GLM NLL loss is PSD and convex

PS1 - 5a

Saturday, July 9, 2022 3:46 PM

(5) a) $J(\theta) = \frac{1}{2} \sum_{i=1}^{N} (\theta^T \hat{x}^{(i)} - y^{(i)})^2$

$\nabla_\theta J(\theta) = \sum_{i=1}^{N} (\theta^T \hat{x}^{(i)} - y^{(i)})(\hat{x}^{(i)})$

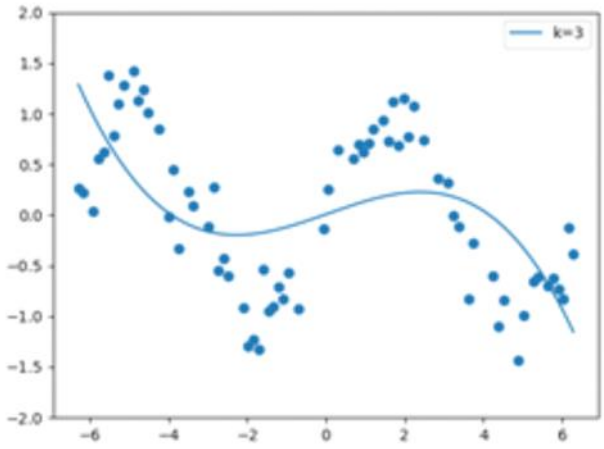
$\propto \lambda \sum_{i=1}^{N} (\theta^T \hat{x}^{(i)} - y^{(i)})(\hat{x}^{(i)})$

update rule: $\theta := \theta - \lambda \sum_{i=1}^{N} (\theta^T \hat{x}^{(i)} - y^{(i)})(\hat{x}^{(i)})$

PS1 - 5b

Saturday, July 9, 2022 3:48 PM

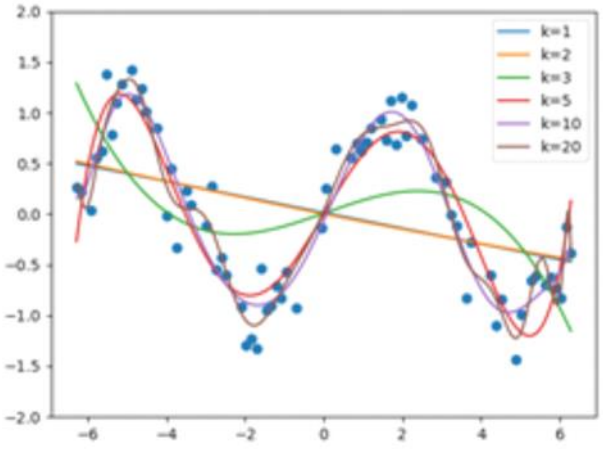
b)



PS1 - 5c

Saturday, July 9, 2022 3:49 PM

c)

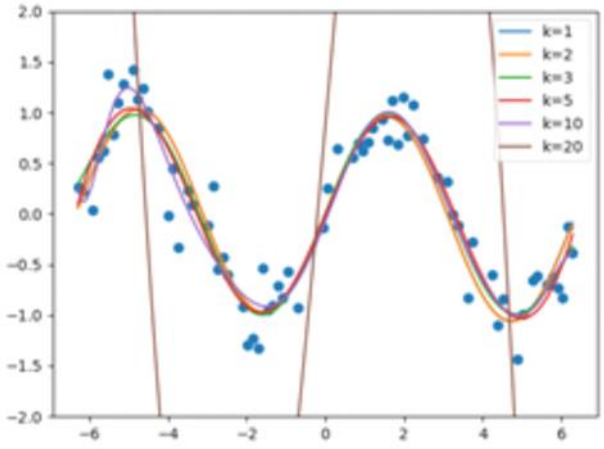


The higher degree of the polynomial, the closer the fit to the data. However, higher degree polynomials are less smooth and less simple by nature of having more parameters

PS1 - 5d

Saturday, July 9, 2022 3:51 PM

d)

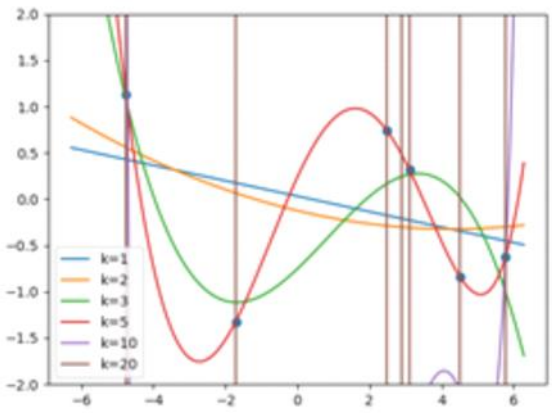


With the sin(x) feature, the polynomials seem to fit the data closer, with the exception of the k=20 polynomial

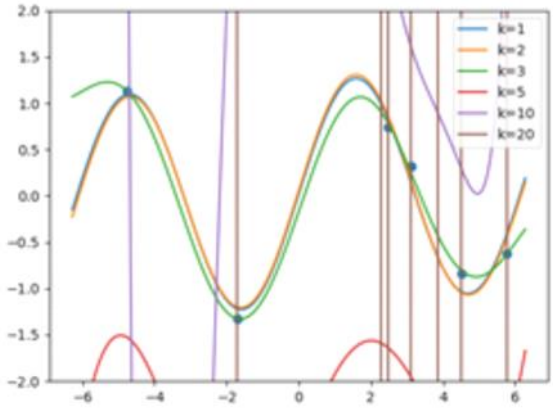
PS1 - 5e

Saturday, July 9, 2022 3:53 PM

e)



Polynomial



With sin(x) feature

With limited data, the polynomial approximations can easily pass through all the data points but vary greatly in their shape. The expanded feature version seems to fit better. However, in both cases as the degree becomes too large there is overfitting.