Stanford CS229: Problem Set #1, math parts

1) Linear Classifiers (Logistic Regression, GDA)

(a) The cost function for logistic regression was

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

recalling $h_\theta(x) = \frac{1}{1+\exp(-\theta^T x)}$ & that $(x^{(i)}, y^{(i)})$ denote the training examples.

- Find the Hessian of J, H.
- Show H is positive semidefinite.
  - ↳ This implies that J is convex ⇒ local extrema are global extrema!

Note that H is already symmetric.

Start w/ gradient. (Hint: it's the same as update rule for lin reg.)

$$\frac{\partial J}{\partial \theta_k} = \frac{-1}{m}\sum_{i=1}^{m} \frac{\partial J}{\partial \theta_k}\left[y^{(i)} \log(h_\theta(x^{(i)}))\right] + \frac{\partial J}{\partial \theta_k}\left[(1-y^{(i)}) \log(1-h_\theta(x^{(i)}))\right]$$

Break into pieces.

- $$y^{(i)} \cdot \frac{\partial J}{\partial \theta_k}\left[\log(h_\theta(x^{(i)}))\right] = y^{(i)}\left(\frac{1}{h_\theta(x^{(i)})}\right) \cdot \frac{\partial h_\theta(x^{(i)})}{\partial \theta_k}$$

- $$\frac{\partial h_\theta(x^{(i)})}{\partial \theta_k} = \frac{\partial}{\partial \theta_k}\left[\frac{1}{1+\exp(-\theta^T x)}\right]$$

$$= \frac{-\frac{\partial}{\partial \theta_k}\left[1+e^{-\theta^T x^{(i)}}\right]}{(1+e^{-\theta^T x^{(i)}})^2} = \frac{-e^{-\theta^T x^{(i)}} \cdot \frac{\partial}{\partial \theta_k}(-\theta^T x^{(i)})}{(1+e^{-\theta^T x^{(i)}})^2}$$

$$= \frac{e^{-\theta^T x^{(i)}} \cdot x_k^{(i)}}{(1+e^{-\theta^T x^{(i)}})^2}$$

$$\cdots = y^{(i)} \left(\frac{1+e^{-\theta^T x^{(i)}}}{1}\right) \cdot \left[\frac{e^{-\theta^T x^{(i)}} x_k^{(i)}}{(1+e^{-\theta^T x^{(i)}})^2}\right]$$

$$= y^{(i)} \frac{e^{-\theta^T x^{(i)}} x_k^{(i)}}{(1+\exp(-\theta^T x^{(i)}))}$$

- $(1-y^{(i)}) \frac{\partial J}{\partial \theta_k} \left[\log\left(1-h_\theta(x^{(i)})\right)\right]$

$$= (-)\left(\frac{1}{1-h_\theta(x^{(i)})}\right) \cdot \frac{\partial}{\partial \theta_k}\left[1-h_\theta(x^{(i)})\right]$$

$$= (-)\frac{1}{1-h_\theta(x^{(i)})} \cdot \frac{\partial}{\partial \theta_k}\left[-h_\theta(x^{(i)})\right]$$

also $1-h_\theta(x^{(i)}) = \frac{e^{-\theta^T x^{(i)}}}{1+e^{-\theta^T x^{(i)}}}$

$$= -(-)\left(\frac{1+e^{-\theta^T x^{(i)}}}{e^{-\theta^T x^{(i)}}}\right) \frac{e^{-\theta^T x^{(i)}} \cdot x_k^{(i)}}{(1+e^{-\theta^T x^{(i)}})^2} = -(1-y^{(i)})\, h_\theta(x^{(i)}) \cdot x_k^{(i)}$$

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

$$= \frac{-1}{m} \sum_{i=1}^{m} y^{(i)} (1-h_\theta(x^{(i)})) x_k^{(i)} - (1-y^{(i)}) h_\theta(x^{(i)}) x_k^{(i)}$$

$$x_k^{(i)} \left[y^{(i)} - y^{(i)} h_\theta(x^{(i)}) - h_\theta(x^{(i)}) + y^{(i)} h_\theta(x^{(i)}) \right]$$

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

2nd derivatives

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

$$- \frac{\partial}{\partial \theta_k}\left[h_\theta(x^{(i)})\right]$$

(11)

cancel $-$

$$= \frac{1}{m} \sum_{i=1}^{m} \frac{e^{-\theta^T x^{(i)}} \cdot (x_k^{(i)})^2}{(1+e^{-\theta^T x^{(i)}})^2} = \frac{1}{m} \sum_{i=1}^{m} h_\theta(x^{(i)}) (1-h_\theta(x^{(i)})) (x_k^{(i)})^2$$

Similarly,

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

So the Hessian:

$$H_{kj} = \begin{cases} \frac{1}{m} \sum_{i=1}^{m} h_\theta(x^{(i)}) \left(1 - h_\theta(x^{(i)})\right) (x_k^{(i)})^2 & \text{when } k = j \\ \frac{1}{m} \sum_{i=1}^{m} h_\theta(x^{(i)}) \left(1 - h_\theta(x^{(i)})\right) \left(x_k^{(i)} x_j^{(i)}\right) & \text{when } k \neq j \end{cases}$$

Or even more succinctly, recalling

$$g'(z) = g(z)\left(1 - g(z)\right),$$

$$H_{kj} = \frac{1}{m} \sum_{i=1}^{m} g'\left(\theta^T x^{(i)}\right) x_k^{(i)} x_j^{(i)} \qquad \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \begin{bmatrix} x_1 \dots x_n \end{bmatrix}$$

$$H = \frac{1}{m} \sum_{i=0}^{m} g'\left(\theta^T x^{(i)}\right) x^{(i)} x^{(i)T} \qquad (?\,?\,?)$$

(Express H as sum of outer products?)

Now, show H is PSD

↳ H is already symmetric

→ WTS: $z^T H z \geq 0 \quad \forall z \in \mathbb{R}^n$

$$z^T H z = \frac{1}{m} z^T \left(\sum_{i=0}^{m} g'(\theta^T x^{(i)}) \, x^{(i)} (x^{(i)})^T \right) z$$

(matrix mult distributes over addition)

$$= \frac{1}{m} \left(\sum_{i=0}^{m} g'(\theta^T x^{(i)}) \, z^T x^{(i)} (x^{(i)})^T z \right)$$

$g'(z) = g(z)(1 - g(z))$

↳ since in $(0, 1)$, $g'(z)$ can never be negative.

$$= \frac{1}{m} \left(\sum_{i=0}^{m} g'(\theta^T x^{(i)}) (x^{(i)})^T z \right)^T (x^{(i)})^T z \right)$$

$$= \frac{1}{m} \left(\sum_{i=0}^{m} g'(\theta^T x^{(i)}) \left((x^{(i)})^T z\right)^2 \right)$$

squared fn always pos.

The typical way to write the Hessian

$$X = \begin{bmatrix} — & x^{(1)} & — \\ & \vdots & \\ — & x^{(m)} & — \end{bmatrix}$$

$$\boxed{H = X^T D X}$$

where $D = \text{diag}(d)$

& $d = \begin{bmatrix} g'(\theta^T x^{(1)}) \\ \vdots \\ g'(\theta^T x^{(m)}) \end{bmatrix} = \begin{bmatrix} h_\theta(x^{(1)})(1-h_\theta(x^{(1)}) \\ \vdots \\ h_\theta(x^{(m)})(1-h_\theta(x^{(m)}) \end{bmatrix}$

Let $d^{(i)} = g'(\theta^T x^{(i)})$ (a scalar)

Todo: show this coincides with earlier duration.

- Let's get an expression for H_{ij}.

$$X^T D = \begin{bmatrix} | & & | \\ x^{(1)} & \cdots & x^{(m)} \\ | & & | \end{bmatrix} \begin{bmatrix} d^{(1)} & & \\ & \ddots & \\ & & d^{(m)} \end{bmatrix}$$

$$= \begin{bmatrix} | & & | \\ d^{(1)} x^{(1)} & \cdots & d^{(m)} x^{(m)} \\ | & & | \end{bmatrix}$$

$$(X^T D)_{ij} = x_i^{(j)} \cdot d^{(j)}.$$

$$X^T D X = \begin{bmatrix} | & & | \\ d^{(1)} x^{(1)} & \cdots & d^{(m)} x^{(m)} \\ | & & | \end{bmatrix} \begin{bmatrix} — & x^{(1)} & — \\ & \vdots & \\ — & x^{(m)} & — \end{bmatrix}$$

$$= \begin{bmatrix} \vdots & & \vdots \\ d^{(1)} x_k^{(1)} & \cdots & d^{(m)} x_k^{(m)} \\ \vdots & & \vdots \end{bmatrix} \begin{bmatrix} \cdots & x_j^{(1)} & \cdots \\ & \vdots & \\ & x^{(m)} & \end{bmatrix}$$

$$H_{kj} = \sum_{i=1}^{m} d^{(i)} x_k^{(i)} x_j^{(i)} = \sum_{i=1}^{m} g'(\theta^T x^{(i)}) x_k^{(i)} x_j^{(i)}$$

which is exactly the expression we derived earlier.

Now, WTS: $z^T H z \geq 0$ for all $z \in \mathbb{R}^m$

what does an element of that matrix look like?

$H \in \mathbb{R}^{m \times m}$?

$$\begin{bmatrix} z_1 & \cdots & z_m \end{bmatrix} \begin{bmatrix} — & h_1 & — \\ & \vdots & \\ — & h_m & — \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_m \end{bmatrix}$$

$$\left(\sum_{i=1}^{m} z_i \vec{h_i} \right) = \begin{bmatrix} \sum_{i=1}^{m} z_i H_{i1} & \cdots & \sum_{i=1}^{m} z_i H_{im} \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_m \end{bmatrix}$$

$$\sum_{i=1}^{m} z_i H_{i1} z_1 + \cdots + \sum_{i=1}^{m} z_i H_{im} z_m$$

$$z^T H z = \sum_{j=1}^{m} \sum_{i=1}^{m} z_i H_{ij} z_j$$

$$= \sum_{j=1}^{m} \sum_{i=1}^{m} z_i \left(\sum_{k=1}^{m} g'(\theta^T x^{(k)}) x_i^{(k)} x_j^{(k)} \right) z_j$$

$$= \sum_{k=1}^{m} g'(\theta^T x^{(k)}) \sum_{j=1}^{m} \sum_{i=1}^{m} z_i x_i^{(k)} x_j^{(k)} z_j$$

$$= (x^{(k)})^T z)^2$$

$$= \sum_{k=1}^{m} g(z)(1 - g(z)) ((x^{(k)})^T z)^2$$

$g \in (0,1)$. squared, always ≥ 0

That is remarkably similar to what I got earlier. Nice.

Need to verify: $\sum_i \sum_j z_i x_i x_j z_j = (x^T z)^2$

$$(x^T z) \cdot (x^T z) = \left(\sum_i x_i z_i \right) \left(\sum_j x_j z_j \right)$$

$$= \sum_i \sum_j x_i z_i x_j z_j$$

for all possible i & j pairings in the mult.

(b) implementing logistic regression using Newton's method

Vector valued version:

$$\theta^{(t+1)} := \theta^{(t)} + H^{-1} \nabla_\theta \ell(\theta)$$

$$X\theta = \begin{bmatrix} — & x^{(1)} & — \\ & \vdots & \\ — & x^{(m)} & — \end{bmatrix} \begin{bmatrix} \theta_0 \\ \vdots \\ \theta_n \end{bmatrix}$$

Plotting decision boundary

$$\frac{1}{1 + \exp(-\theta^T x)} = 0.5$$

$$1 = 0.5\,(1 + \exp(-\theta^T x))$$

$$1 = \tfrac{1}{2} + \tfrac{1}{2}\exp(-\theta^T x)$$

$$\tfrac{1}{2} = \tfrac{1}{2}\exp(-\theta^T x)$$

$$1 = \exp(-\theta^T x)$$

$$0 = -\theta^T x$$

$$\theta^T x = 0$$

$$\theta_0 + \theta_1 x_1 + \theta_2 x_2 = 0$$

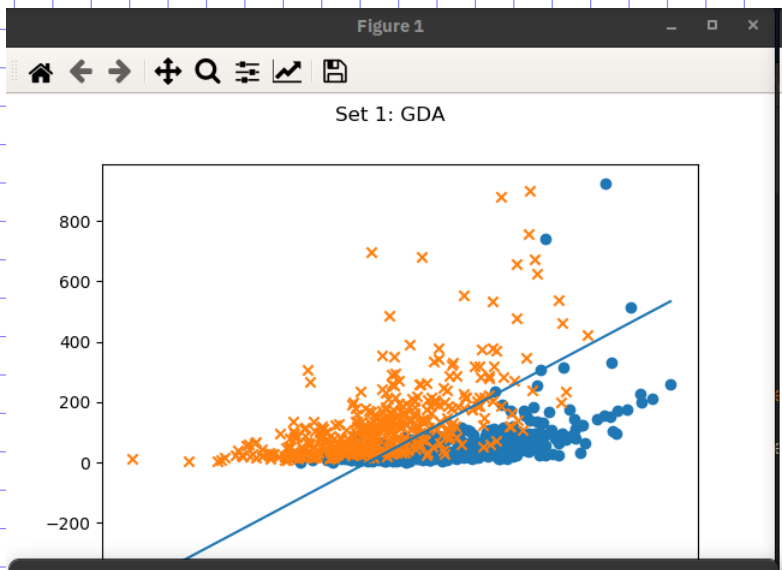
$$\theta_2 x_2 = -\theta_0 - \theta_1 x_1$$

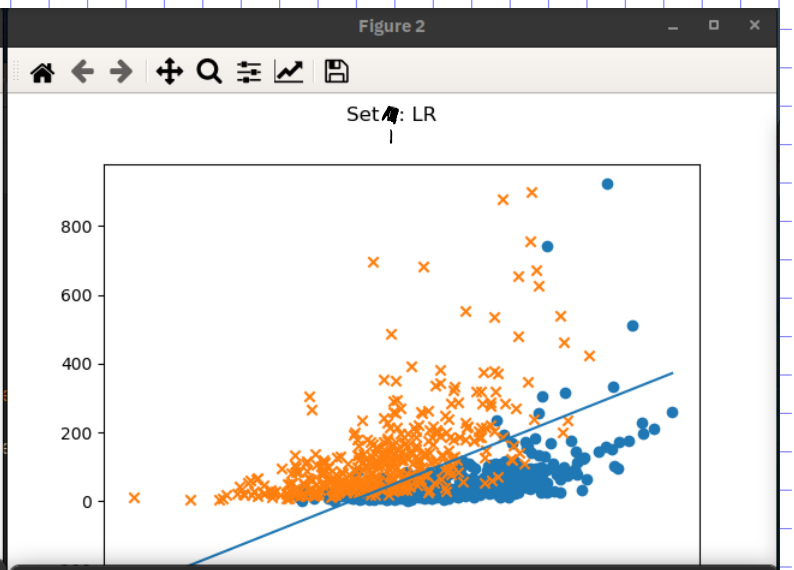
$$x_2 = \frac{-\theta_0 - \theta_1 x_1}{\theta_2}$$

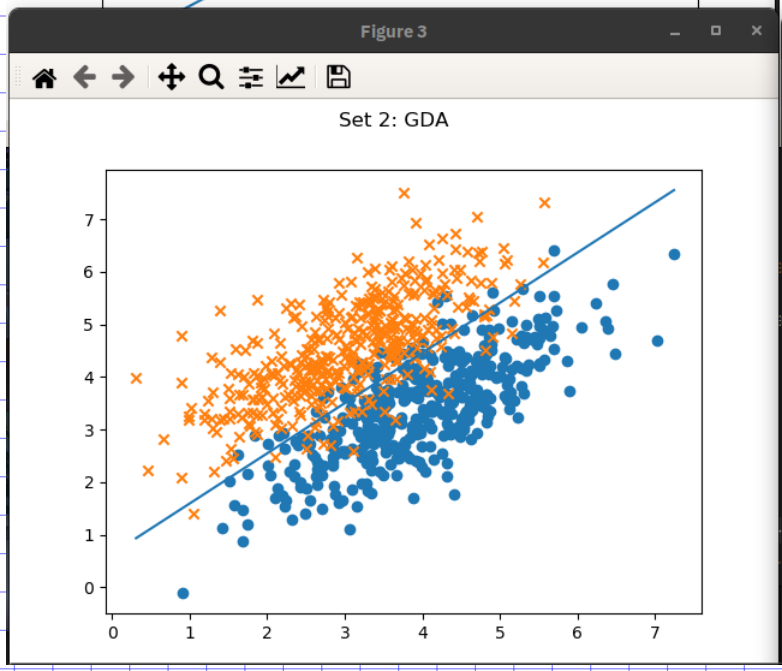
$$0 = -\theta^T x - \theta_0$$

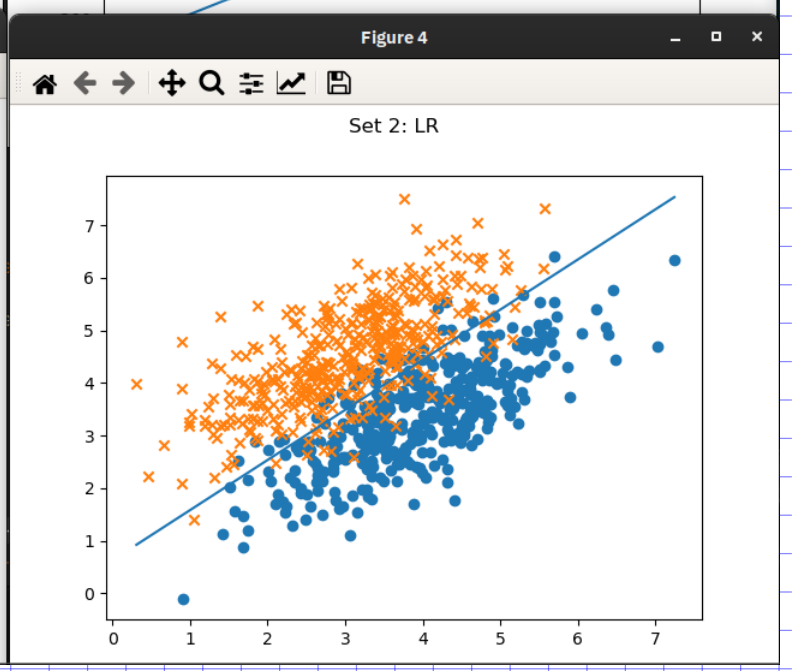
$$\theta_0 = -\theta^T x$$

$$\theta_0 = -\theta^1 x_1 - \theta^2 x_2$$










```
In [120]: %cpaste -q
print("set 2")
print("LR\n",confusion_mat(df_2_train["y"],LR_2_preds))
print("GDA\n",confusion_mat(df_2_train["y"],GDA_2_preds))
--
set 2
LR
 [[367  33]
 [ 35 365]]
GDA
 [[368  32]
 [ 37 363]]
```

Confusion matrices

```
--
set 1
LR
 [[358  42]
 [ 52 348]]
GDA
 [[351  49]
 [ 58 342]]
```

GDA/LR performance is about the same on set 2, w/ GDA performing slightly better.

GDA does worse on set 1.

↳ The data does not appear to be Gaussian.

(c) Recall GDA

I think I'm really off the rule here...

$x|y=0 \sim \mathcal{N}(\mu_0, \Sigma)$

$x|y=1 \sim \mathcal{N}(\mu_1, \Sigma)$

$y \sim \text{Bernoulli}(\phi)$

Show GDA has a liner decision boundary, namely;

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

where $\theta \in \mathbb{R}^n$ & $\theta_0 \in \mathbb{R}$ are functions of the given parameters.

Strictly speaking, density fn...

$$P(y=1|x) = \frac{P(x|y=1)\,P(y=1)}{P(x)}$$

(Attempt 1: incorrect)

$$= \frac{P(x|y=1)\,P(y=1) = \phi}{P(y=1)\cdot P(x|y=1) + P(y=0)\,P(x|y=0)}$$

$$= \frac{\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)\cdot\phi}{\phi\cdot\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}\exp\left(\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right) + (1-\phi)\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)}$$

(μ_1 in the numerator; μ_1 and μ_0 in the denominator)

$$= \frac{\mu_1\phi}{\phi\mu_1-(1-\phi)\mu_0} = \frac{\mu_1\phi}{\phi\mu_1-\mu_0+\phi\mu_0} = \frac{\mu_1\phi}{\phi(\mu_1+\mu_0)-\mu_0}$$

$$= \frac{1}{\frac{\phi(\mu_1+\mu_0)}{\mu_1\phi}-\frac{\mu_0}{\mu_1\phi}} = \frac{1}{1+\frac{\mu_0}{\mu_1}-\frac{\mu_0}{\mu_1}\cdot\frac{1}{\phi}} = \frac{1}{1+\left(1-\frac{1}{\phi}\right)\frac{\mu_0}{\mu_1}}$$

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

Can it be combined into one term?

doesn't seem correct / need to identify θ, θ_0

(Attempt 2)

$$= \frac{P(x|y=1)\,P(y=1)}{P(y=1)\cdot P(x|y=1) + P(y=0)\,P(x|y=0)} \qquad \phi$$

$$= \frac{\phi\, P(x|y=1)}{\phi\, P(x|y=1) + (1-\phi)\, P(x|y=0)}$$

$$= \frac{\phi M_1}{\phi M_1 + (1-\phi) M_0} = \frac{1}{\frac{\phi M_1}{\phi M_1} + \frac{1-\phi}{\phi}\frac{M_0}{M_1}}$$

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

$\left(\frac{1-\phi}{\phi}\right) = 1$

$\exp\left(\ln\left(\frac{1-\phi}{\phi}\right)\right)$

(distribute into ...

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

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

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

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

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

- Inverse of symm. matrix is symm $\Rightarrow \Sigma^{-1}$ symm.
- $x^TAy = y^TAx$ for any symm matrix A.

("Symmetric Bilinear Form")

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

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

Another fact you didn't know:

$$(\mu_0+\mu_1)^T\Sigma^{-1}(\mu_0-\mu_1)$$
$$\mu_0^T\Sigma^{-1}(\mu_0-\mu_1) + \mu_1^T\Sigma^{-1}(\mu_0-\mu_1)$$
$$\mu_0\Sigma^{-1}\mu_0 - \mu_0^T\Sigma^{-1}\mu_1 + \underbrace{\mu_1^T\Sigma^{-1}\mu_0}_{\downarrow \text{symmetry}} - \mu_1^T\Sigma^{-1}\mu_1$$

cancel

$$\mu_0^T\Sigma^{-1}\mu_1$$

$$\Rightarrow \mu_0\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1$$

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

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

$$\Rightarrow \theta = (\mu_1-\mu_0)^T\Sigma^{-1}$$

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

(d) Derive the MLE estimates for GDA.

$$\ell(\phi, \mu_0, \mu_1, \Sigma) = \ln \prod_{i=1}^{m} P(x^{(i)} | y^{(i)}; \ldots) \, P(y^{(i)}; \phi)$$

Take partials wrt each parameter assume all just scalar (so no matrix derivatives?)

Recall here, $y^{(i)} \in \{0, 1\}$. or, $y^{(i)} = 1_{\{y^{(i)}=1\}}$. (a little redundant.)

$$= \ln \prod_{i=1}^{m} \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu_{y^{(i)}})^T \Sigma^{-1} (x - \mu_{y^{(i)}})\right) \cdot \phi^{y_i} (1-\phi)^{1-y_i}$$

(simplify the log likelihood.)

$$= \sum_{i=1}^{m} \ln\left[\ldots \right]$$

$$= \sum_{i=1}^{m} \ln\left(\frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \phi^{y_i} (1-\phi)^{1-y_i}\right) + \boxed{-\frac{1}{2}(x - \mu_{y^{(i)}})^T \Sigma^{-1} (x - \mu_{y^{(i)}})}$$

$$\boxed{\ln\left(\frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}}\right)} + \ln\left(\phi^{y^{(i)}} (1-\phi)^{1-y^{(i)}}\right)$$

$$\left[\ldots \right] + \boxed{\ln(\phi) y^{(i)} + \ln(1-\phi)(1-y^{(i)})} + \ldots$$

Estimator of ϕ:

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

Set 0, solve for ϕ

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

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

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

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

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

D

Estimators of μ_0/μ_1:

$$\frac{\partial \ell}{\partial \mu_0} = \sum_{i=1}^{m} \frac{\partial}{\partial \mu_0}\left[-\frac{1}{2}(x^{(i)}-\mu_{y^{(i)}})^T \Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}})\right]$$

Note if $y^{(i)}=1$, the derivative of the inner term is 0.

But, if $y^{(i)}=0$, we can get a nonzero derivative. Also rewrite for univariate case: $\Sigma=\sigma^2$, $\Sigma^{-1}=\frac{1}{\sigma^2}$

$$= \sum_{i=1}^{m} \frac{\partial}{\partial \mu_0}\left[\frac{1}{2\sigma^2}(x^{(i)}-\mu_{y^{(i)}})^2\right]$$

$$0 = \sum_{i \in \{1 \le i \le m \mid y^{(i)}=0\}} \frac{-1}{2\sigma^2} \cdot 2(x^{(i)}-\mu_0)(-1) = \sum \frac{1}{\sigma^2}(x^{(i)}-\mu_0)$$

$$= \sum (x^{(i)}-\mu_0) = \sum_{i=1}^{m} x^{(i)} 1_{\{y^{(i)}=0\}} - \mu_0 \sum \ldots$$

$$0 = \sum_{i=1}^{m} x^{(i)} 1_{\{y^{(i)}=0\}} - \mu_0 \sum_{i=1}^{m} 1_{\{y^{(i)}=0\}}$$

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

And the same derivations should hold for μ_1: just change the label.

Estimator of $\Sigma=\sigma^2$: $\sigma^2 = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})^2$

$$\frac{\partial \ell}{\partial \sigma^2} = \sum_{i=0}^{m} \frac{\partial}{\partial \sigma^2}\left[\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \frac{-1}{2\sigma^2}(x^{(i)}-\mu_{y^{(i)}})^2\right]$$

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{-1} \cdot \left(-\frac{1}{2}\right)(2\pi\sigma^2)^{-3/2} \cdot 2\pi + \left(\frac{1}{2}\right)(-1)(\sigma^2)^{-2}(x^{(i)}-\mu_{y^{(i)}})^2$$

$$\ldots (2\pi\sigma^2)^{-1} \cdot 2\pi \cdot \left(-\frac{1}{2}\right) + \left(\frac{1}{2}\right)\cdot(\sigma^2)^{-2}(x^{(i)}-\mu_{y^{(i)}})^2$$

$$\ldots \frac{1}{2}(\sigma^2)^{-2}(x^{(i)}-\mu_{y^{(i)}})^2 - \frac{1}{2}(\sigma^2)^{-2}\sigma^2$$

$$\sum_{i=1}^{m} \frac{1}{2}(\sigma^2)^{-2}\left[(x^{(i)} - \mu_{y^{(i)}})^2 - \sigma^2\right]$$

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

$$\frac{1}{2(\sigma^2)^2}\left[\sum_{i=1}^{m}(x^{(i)} - \mu_{y^{(i)}})^2 - m\sigma^2\right]$$

Set equal to 0:

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

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

which is equivalent to the expression given in the univariate case, i.e

$$(x^{(i)} - \mu_{y^{(i)}})^2 = (x^{(i)} - \mu_{y^{(i)}})^T (x^{(i)} - \mu_{y^{(i)}})$$

when $x \in \mathbb{R}$.

(2) Incomplete, Positive Only Labels.

Scenario: Binary labeled data, but only a subset of the positive examples are labeled. The rest are unlabeled. They could be pos. or neg.

Goal: Construct a binary classifier that models the true label as closely as possible.

Notation

$$y^{(i)} = \begin{cases} 1 & \text{if example } i \text{ is labeled} \\ 0 & \text{otherwise} \end{cases}$$

$$t^{(i)} = \begin{cases} 1 & \text{if the true label is 1} \\ 0 & \text{if the true label is 0} \end{cases}$$

(a) Suppose the labeled examples were selected uniformly randomly from the set of positive examples. Then,

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

In other words, it didn't matter what the features of the examples were when we selected the ones to be the labeled examples.

↳ e.g. violated if the only labeled ones were $x > 100$, due to measurement sensitivity?

Show: $P(t^{(i)}=1 \mid x^{(i)}) = \frac{1}{\alpha} \cdot P(y^{(i)}=1 \mid x^{(i)})$

- The probability of being labeled differs by a constant factor from the prob. of having a true label of 1.

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

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

$P(y^{(i)}=1 \mid t^{(i)}=1) = \alpha$ — constant but might vary w.r.t. $x^{(i)}$?

Constant: roughly equal to the proportion

$$= \frac{\text{labeled examples}}{\text{all positive examples}}$$

(b) Estimate α using some trained model h validation set V

Let $V_+ := \{x^{(i)} \in V \mid y^{(i)} = 1\}$
↳ the labeled examples in V.

Assume:
- $h(x^{(i)}) \approx P(y^{(i)}=1 \mid x^{(i)})$ ← the classifier predicts prob of being labeled
- $x^{(i)} \in V_+ \Rightarrow P(t^{(i)}=1 \mid x^{(i)}) \approx 1$ — If in labeled set, that x's true label is 1.

Show: $h(x^{(i)}) \approx \alpha \quad \forall x^{(i)} \in V_+$

Assume $x^{(i)} \in V_+$. Then, by the last problem,

$$\underbrace{P(y^{(i)}=1 \mid x^{(i)})}_{\hookrightarrow \approx h(x^{(i)})} = P(y^{(i)}=1 \mid t^{(i)}=1, x^{(i)})\, P(t^{(i)}=1 \mid x^{(i)})$$
$$= P(y^{(i)}=1 \mid t^{(i)}=1)\, \underbrace{P(t^{(i)}=1 \mid x^{(i)})}_{\hookrightarrow \approx 1 \text{ by assumption}}$$

$= \alpha$ by last problem

$$\Rightarrow h(x^{(i)}) \approx \alpha$$

Plotting decision boundary for α-corrected prediction

$$\frac{1}{\alpha}\,\frac{1}{1+\exp(-\theta^T x)} = 0.5$$

$$\frac{1}{1+\exp(-\theta^T x)} = \alpha \cdot 0.5$$

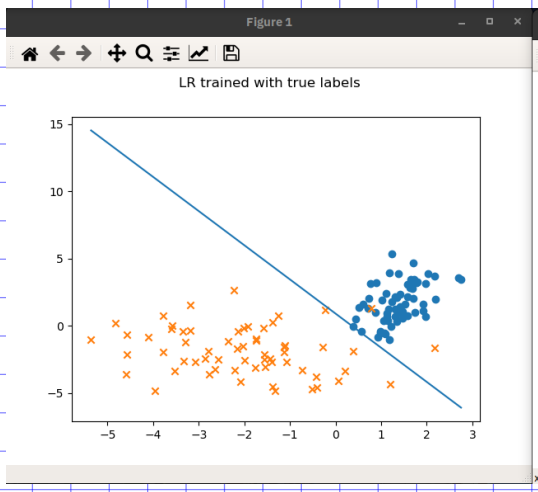
$$1+\exp(-\theta^T x) = \frac{1}{\alpha \cdot 0.5}$$

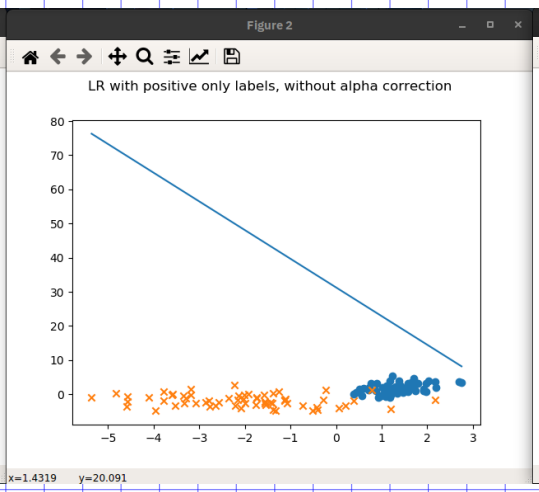
$$\exp(-\theta^T x) = \frac{1}{\alpha \cdot 0.5} - 1$$

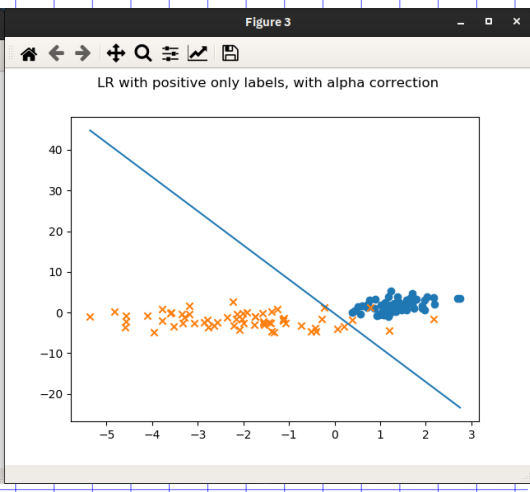
$$-\theta^T x = \ln\left(\frac{1}{\alpha \cdot 0.5} - 1\right)$$

$$-\theta_0 + -\theta_1 x_1 - \theta_2 x_2 = \ln\left(\frac{1}{\alpha \cdot 0.5} - 1\right)$$

$$= \frac{\ln\left(\frac{1}{\alpha \cdot 0.5} - 1\right) + \theta_1 x_1 + \theta_0}{-\theta_2}$$








```
LR trained with access to true labels
[[14  0]
 [50 60]]
LR trained with access to partial positive labels, no rescale
[[  0  14]
 [  0 110]]
LR trained with access to partial positive labels
[[14  0]
 [52 58]]
```

(3) Poisson Regression

(a) Show that the Poisson dist. is in the exponential family.

$$P_Y(y;\lambda) = \frac{e^{-\lambda}\lambda^y}{y!} = \frac{1}{y!}\cdot \exp(\ln(\lambda^y))e^{-\lambda}$$

$$= \frac{1}{y!}\exp(y\ln(\lambda) - \lambda)$$

$$= \frac{1}{y!}\exp(y\ln(\lambda) - \exp(\ln(\lambda)))$$

The Poisson distribution is in the Exponential Family w/...

$T(y) = y$ $\quad a(\eta) = e^{\eta}$

$b(y) = \frac{1}{y!}$ $\quad \eta = \ln(\lambda) \Rightarrow \boxed{\lambda = e^{\eta}}$

(b) What is the canonical response f'n when we do GLM regression w/ a Poisson R.V?

& in fact that's the prediction f'n, since Poisson regr. would assume

$$y|x \sim \text{Poisson}(\lambda), \text{ but } \lambda = e^{\eta} = e^{\theta^T x}$$

$$\Rightarrow h_\theta(x) = \mathbb{E}[y|x;\theta] = \lambda = e^{\eta} = e^{\theta^T x} \Rightarrow h_\theta(x) = e^{\theta^T x}$$

(c) Derive the stochastic gradient ascent rule by maximizing the log likelihood.

Gradient ascent:

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

$$\ell(\theta) = \log \prod_{i=1}^{m} P(y^{(i)}|x^{(i)};\theta) = \log \prod_{i=1}^{m} \frac{e^{-\lambda(\theta)}(\lambda(\theta))^{y^{(i)}}}{y^{(i)}!}$$

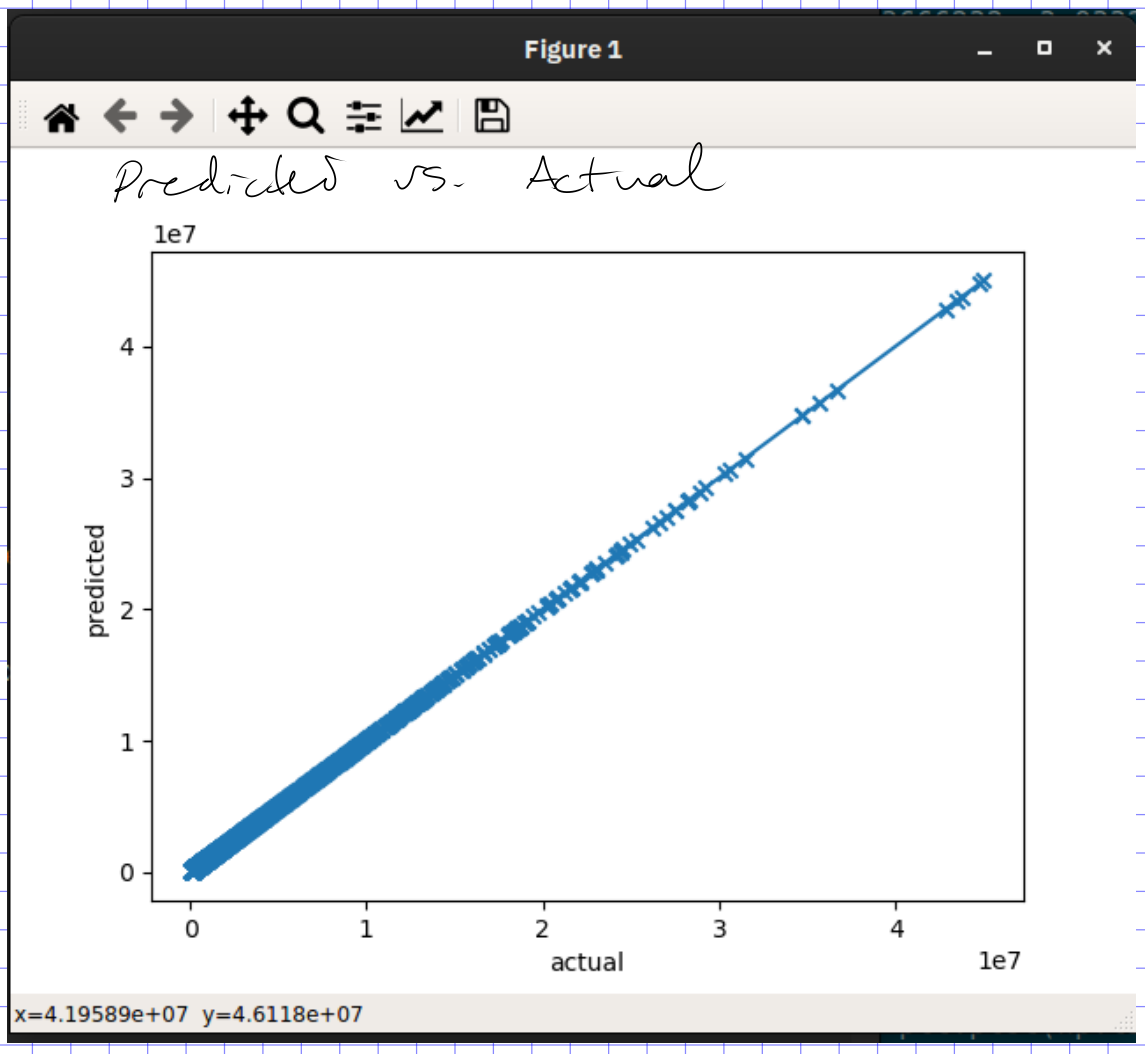
$$= \sum_{i=1}^{m} \log(\dots) = \sum_{i=1}^{m} -\lambda(\theta) + y^{(i)}\log(\lambda(\theta)) + \log\left(\frac{1}{y^{(i)}!}\right)$$

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

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

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

$$= \sum_{i=1}^{m} x_j^{(i)}\left(y^{(i)} - \underbrace{e^{\theta^T x^{(i)}}}_{h_\theta(x^{(i)})}\right)$$




```
[7.36659576 3.93332251 3.43327325 1.99993815 4.39995804]
```

Final θ vector

Key: need to choose a pretty extreme learning rule $\alpha \approx 10^{-8}$ since otherwise gradient vectors are too large: since magnitude of y too large.

Using a fixed number of SGD iterations worked okay in this case. ($\approx$ 15 iters)

(4) Convexity of GLMs

Goal: Show NLL (Negative Log-Likelihood) of a GLM is convex in the natural parameter η.
$\Rightarrow$ local maxima are global maxima.
Show this by showing the Hessian is PSD.

Setup: $Y|X \sim \text{ExpFamily}(\eta)$
where $\eta \in \mathbb{R}$ & the sufficient statistic is $T(y) = y$.

$$\Rightarrow f_{Y|X}(y;\eta) = b(y)\exp(\eta y - a(\eta))$$

Recall that $\eta(x) = \theta^T x$. But we'll get there in a second.

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

Let A be the support set of $Y|X$.

$$= \int_A y \cdot f_{Y|X}(y;\eta)\,dy = \int_A y \cdot b(y)\exp(\eta y - a(\eta))\,dy$$

$$= \int_A b(y)\cdot y\exp(\eta y - a(\eta))\,dy$$

$$\frac{\partial}{\partial \eta}\left[\exp(\eta y - a(\eta))\right] = \exp(\eta y - a(\eta))\cdot\left[y - \frac{\partial a}{\partial \eta}\right]$$
$$= \exp(\eta y - a(\eta))\,y - \frac{\partial a}{\partial \eta}\exp(\eta y - a(\eta))$$

$$= \int b(y)\left[\frac{\partial}{\partial \eta}\left[\exp(\eta y - a(\eta))\right] + \frac{\partial a}{\partial \eta}\exp(\eta y - a(\eta))\right]dy$$

$$= \frac{\partial}{\partial \eta}\int_A b(y)\exp(\eta y - a(\eta))\,dy + \frac{\partial a}{\partial \eta}\int_A b(y)\exp(\eta y - a(\eta))\,dy$$

$$= \frac{\partial}{\partial \eta}\left[\int_A f_{Y|X}(y;\eta)\,dy\right] + \frac{\partial a}{\partial \eta}\int_A f_{Y|X}(y;\eta)\,dy$$

$$= \frac{\partial}{\partial \eta}[1] + \frac{\partial a}{\partial \eta}\cdot 1 = \frac{\partial a}{\partial \eta}$$

(b) Show $\mathrm{Var}(Y|X;\theta) = \frac{\partial^2 a}{\partial \eta^2}$.

$$\mathbb{E}[(Y|X)^2;\theta] = \int_A y^2 \cdot b(y)\exp(\eta y - a(\eta))\,dy$$

$$\frac{\partial^2}{\partial \eta^2}\left[\exp(\eta y - a(\eta))\right]$$

$$= \frac{\partial}{\partial \eta}\left[\exp(\eta y - a(\eta))\, y - \frac{\partial a}{\partial \eta}\exp(\eta y - a(\eta))\right]$$

$$= y\frac{\partial}{\partial \eta}\left[\exp(\eta y - a(\eta))\right] - \left[\frac{\partial^2 a}{\partial \eta^2}\exp(\eta y - a(\eta)) + \frac{\partial a}{\partial \eta}\frac{\partial}{\partial \eta}\left[\exp(\eta y - a(\eta))\right]\right]$$

Shorthand: let $M = \exp(\eta y - a(\eta))$

$$\Rightarrow \frac{\partial M}{\partial \eta} = My - \frac{\partial a}{\partial \eta}M$$

$$= y\left(My - \frac{\partial a}{\partial \eta}M\right) - \frac{\partial^2 a}{\partial \eta^2}M + \frac{\partial a}{\partial \eta}\left(My - \frac{\partial a}{\partial \eta}M\right)$$

$$= My^2 - \frac{\partial a}{\partial \eta}My - \frac{\partial^2 a}{\partial \eta^2}M + \frac{\partial a}{\partial \eta}My - \left(\frac{\partial a}{\partial \eta}\right)^2 M$$

$$\Rightarrow My^2 = \frac{\partial^2 M}{\partial \eta^2} + \frac{\partial^2 a}{\partial \eta^2}M + \left(\frac{\partial a}{\partial \eta}\right)^2 M$$

$$\Rightarrow \cdots = \int_A b(y)\left[\frac{\partial^2 M}{\partial \eta^2} + \frac{\partial^2 a}{\partial \eta^2}M + \left(\frac{\partial a}{\partial \eta}\right)^2 M\right]dy$$

$$= \frac{\partial^2}{\partial \eta^2}\left[\int_A b(y)M\,dy\right] + \frac{\partial^2 a}{\partial \eta^2}\int_A b(y)M\,dy + \left(\frac{\partial a}{\partial \eta}\right)^2 \int_A b(y)M\,dy$$

($\int_A b(y)M\,dy = 0$; $\int_A b(y)M\,dy = 1$; $\int_A b(y)M\,dy = 1$)

$$\Rightarrow \mathbb{E}[(Y|X)^2;\eta] = \frac{\partial^2 a}{\partial \eta^2} + \left(\frac{\partial a}{\partial \eta}\right)^2$$

$$\mathrm{Var}(Y|X;\eta) = \mathbb{E}[(Y|X)^2;\eta] - (\mathbb{E}[Y|X;\eta])^2$$

$$= \frac{\partial^2 a}{\partial \eta^2} + \left(\frac{\partial a}{\partial \eta}\right)^2 - \left(\frac{\partial a}{\partial \eta}\right)^2 = \frac{\partial^2 a}{\partial \eta^2}.$$

(c) $\ell(\theta)$: NLL. Then show $\ell(\theta)$'s Hessian is always PSD, which shows the convexity of GLMs.

$$\ell(\theta) = -\log \prod_{i=1}^{m} f_{Y|X}(y^{(i)} | x^{(i)}; \theta)$$

$$= -\log \prod_{i=1}^{m} b(y^{(i)}) \exp(\eta y^{(i)} - a(\eta))$$

$$= -\sum_{i=1}^{m} \log(b(y^{(i)})) + [\eta y^{(i)} - a(\eta)]$$

& recall: $\eta = \theta^T x^{(i)}$

$$= -\sum_{i=1}^{m} \log(b(y^{(i)})) + \theta^T x^{(i)} \cdot y^{(i)} - a(\theta^T x^{(i)})$$

Compute gradient + hessian:

$$\frac{\partial \ell}{\partial \theta_j} = -\sum_{i=1}^{m} \frac{\partial}{\partial \theta}[\theta^T x^{(i)}] y^{(i)} - \left[\frac{\partial a}{\partial \eta}\Big|_{\eta=\theta^T x^{(i)}}\right] \frac{\partial}{\partial \theta}[\theta^T x^{(i)}]$$

$x_j^{(i)}$; $h_\theta(x^{(i)})$; $x_j^{(i)}$

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

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

where $h_\theta(x^{(i)}) = \frac{\partial a}{\partial \eta}\Big|_{\eta=\theta^T x^{(i)}} = \mathbb{E}[Y|X=x^{(i)}]$.

$$\frac{\partial^2 \ell}{\partial \theta_j \partial \theta_k} = \frac{\partial}{\partial \theta_k}\left[-\sum_{i=1}^{m} x_j^{(i)} y^{(i)} - x_j^{(i)} \cdot \frac{\partial a}{\partial \eta}\Big|_{\eta=\theta^T x^{(i)}}\right]$$

$$= \sum_{i=1}^{m} x_j^{(i)} x_k^{(i)} \frac{\partial^2 a}{\partial \eta^2}\Big|_{\eta=\theta^T x^{(i)}}$$

$$= \sum_{i=1}^{m} x_j^{(i)} x_k^{(i)} \mathrm{Var}(Y|X=x^{(i)})$$

$H = X^T D X$ where

$X = \begin{bmatrix} - x^{(1)} - \\ \vdots \\ - x^{(m)} - \end{bmatrix}$ m by n

$D = \text{diag}([\text{Var}(Y|X=x^{(1)}) \dots \text{Var}(Y|X=x^{(m)})])$

(so D is a diagonal matrix w/ all positive entries on diag.)

To show PSD, WTS $\forall z \in \mathbb{R}^n \quad z^T H z \geq 0$.

- Since D is a diagonal matrix w/ only positive entries on its diagonal, D is PSD.

Easy to prove. Suppose all $d_i \geq 0$.

$$[x_1 \dots x_m] \begin{bmatrix} d_1 & & \\ & \ddots & \\ & & d_m \end{bmatrix} = [x_1 d_1 \dots x_m d_m]$$

$$[x_1 d_1 \dots x_m d_m] \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} = \sum x_i^2 d_i \geq 0.$$

$\Rightarrow z^T (X^T D X) z = (Xz)^T D (Xz)$

Let $y = Xz$. $y \in \mathbb{R}^m$.

$y^T D y \geq 0$ since D PSD.

$\Rightarrow (Xz)^T D (Xz) \geq 0 \Rightarrow z^T H z \geq 0$.

H is PSD.

(5) Locally Weighted Linear Regression

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{m} w^{(i)}(\theta^T x^{(i)} - y^{(i)})^2$$

(i) Show that $J(\theta)$ can be rewritten as

$$J(\theta) = (X\theta - y)^T W (X\theta - y)$$

for an appropriate matrix W.

$$X = \begin{bmatrix} — & x^{(1)} & — \\ & \vdots & \\ — & x^{(m)} & — \end{bmatrix} \qquad y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

Guess:

$W = \text{diag}[w^{(1)} \dots w^{(m)}]$

i.e. $W_{ij} = \begin{cases} w^{(i)} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$

$$X\theta - y = \begin{bmatrix} \theta^T x^{(1)} - y^{(1)} \\ \vdots \\ \theta^T x^{(m)} - y^{(m)} \end{bmatrix}_{m \times 1} \qquad \begin{bmatrix} w^{(1)} & & & \\ & w^{(2)} & & \\ & & \ddots & \\ & & & w^{(m)} \end{bmatrix}$$

$$(X\theta - y)^T W = \begin{bmatrix} w^{(1)}\theta^T x^{(1)} & \dots & w^{(m)}\theta^T x^{(m)} \end{bmatrix}$$

$$\Rightarrow (X\theta - y)^T W (X\theta - y) = \begin{bmatrix} w^{(1)}\theta^T x^{(1)} & \dots & w^{(m)}\theta^T x^{(m)} \end{bmatrix} \begin{bmatrix} \theta^T x^{(1)} \\ \vdots \\ \theta^T x^{(m)} \end{bmatrix}$$

$$= w^{(1)}(\theta^T x^{(1)})^2 + \dots + w^{(m)}(\theta^T x^{(m)})^2$$

$$= \sum_{i=1}^{m} w^{(i)}(\theta^T x^{(i)})^2$$

(ii) Generalize the normal equations to the weighted setting.

(Take the derivative $\nabla_\theta J(\theta)$ & set it equal to 0.)

$$\frac{\partial J}{\partial \theta_j} = \frac{1}{2}\sum_{i=1}^{m} w^{(i)} \frac{\partial}{\partial \theta_j}\left[(\theta^T x^{(i)} - y^{(i)})^2\right] = \frac{2}{2}\sum_{i=1}^{m} w^{(i)} \cdot 1(\theta^T x^{(i)} - y^{(i)})\, x_j^{(i)}$$

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

Somehow express in matrix form.

$(X\theta - y)_i = \theta^T x^{(i)} - y^{(i)}$

$\boxed{WX(X\theta - y)}$

$$WX = \begin{bmatrix} w^{(1)} & & \\ & \ddots & \\ & & w^{(m)} \end{bmatrix} \begin{bmatrix} - & x^{(1)} & - \\ & \vdots & \\ - & x^{(m)} & - \end{bmatrix}$$

$$= \begin{bmatrix} -w^{(1)}x^{(1)} - \\ \vdots \\ -w^{(m)}x^{(m)} - \end{bmatrix}$$

$(m \times m)(m \times n)$

$m \times n$

$$\begin{bmatrix} -w^{(1)}x^{(1)} - \\ \vdots \\ -w^{(m)}x^{(m)} - \end{bmatrix} \begin{bmatrix} \theta^T x^{(1)} - y^{(1)} \\ \vdots \\ \theta^T x^{(m)} - y^{(1)} \end{bmatrix}$$

$m \times 1$

No, dimension mismatch $\Rightarrow x^T W(X\theta - y)$ (this is right)

$$x^T W = \begin{bmatrix} | & & | \\ w^{(1)}x^{(1)} & \cdots & w^{(m)}x^{(m)} \\ | & & | \end{bmatrix} \begin{bmatrix} \theta^T x^{(1)} - y^{(1)} \\ \vdots \\ \theta^T x^{(m)} - y^{(m)} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^m w^{(i)} x_1^{(i)} (\theta^T x^{(i)} - y^{(i)}) \\ \vdots \\ \end{bmatrix}$$

$n \times m$ $\quad$ $m \times 1$

$\Rightarrow X^T W X\theta - X^T W y = 0$

$X^T W X\theta = X^T W y$

Or, you could use matrix calc rules I guess.

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

(iii) Dataset $\{(x^{(i)}, y^{(i)})\}_{i \in \{1, \ldots, m\}}$ but the probability model of each y has a different variance $(\sigma^{(i)})^2$.

i.e, $y^{(i)} | x^{(i)} \sim N(\theta^T x^{(i)}, (\sigma^{(i)})^2)$

Show: The ML estimate of θ in this scenario is a weighted linear regression problem.

What is the weight for $w^{(i)}$?

$$\ell(\theta) = \ln \prod_{i=1}^m P(y^{(i)} | x^{(i)}; \theta)$$

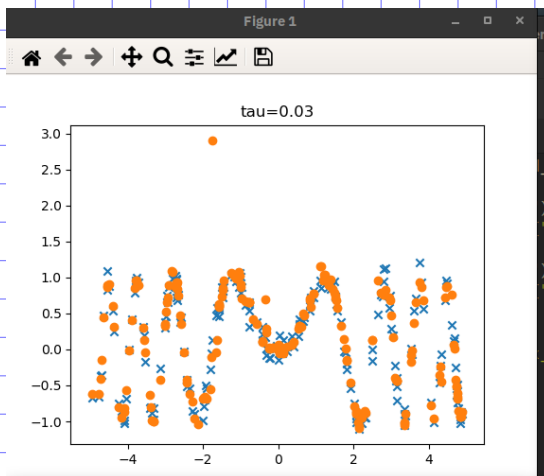
$$= \ln \prod_{i=1}^m \frac{1}{\sqrt{2\pi}\sigma^{(i)}} \exp\left(- \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2} \right)$$

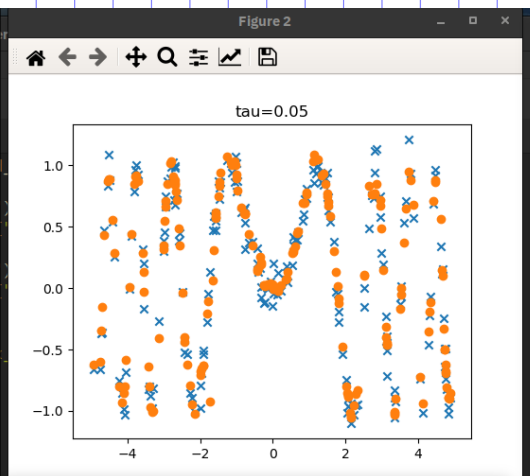
$$= \sum_{i=1}^{m} \ln\left(\frac{1}{\sqrt{2\pi}\,\sigma^{(i)}}\right) - \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2}$$

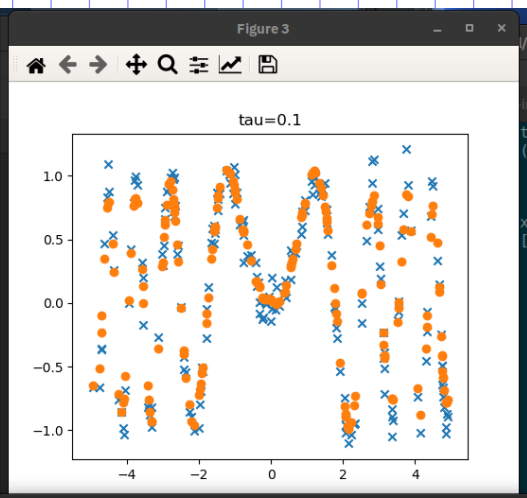
$$\frac{\partial l}{\partial \theta_j} = \sum_{i=1}^{m} -\frac{1}{2(\sigma^{(i)})^2} \cdot 2(y^{(i)} - \theta^T x^{(i)}) \cdot (-x_j^{(i)})$$

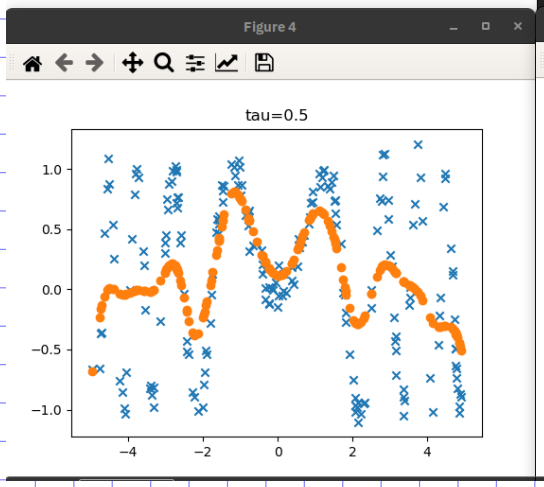
$$= \sum_{i=1}^{m} \frac{1}{(\sigma^{(i)})^2} (y^{(i)} - \theta^T x^{(i)}) x_j^{(i)}$$

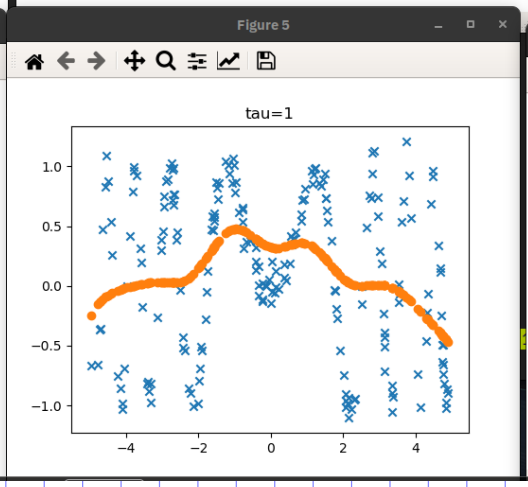
$$\Rightarrow w^{(i)} = (\sigma^{(i)})^{-2}$$

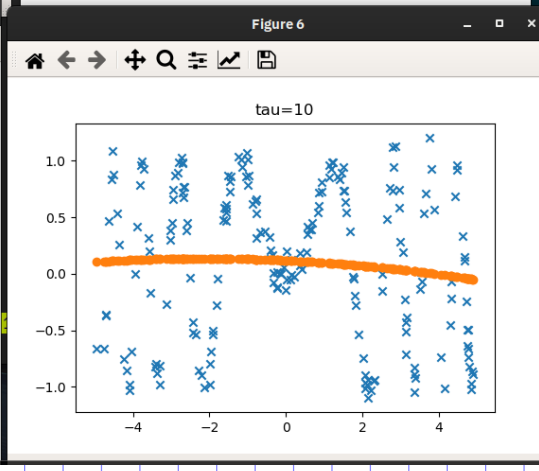













```
--
validation set mse for tau=0.03: 0.059038566452005385
validation set mse for tau=0.05: 0.018944406039864 88
validation set mse for tau=0.1: 0.014461728527284943
validation set mse for tau=0.5: 0.2578133818426481
validation set mse for tau=1: 0.3794515567158331
validation set mse for tau=10: 0.42989004449194135
```