

# CS 229 : Problem Set #1 : Supervised Learning

## 1) Linear Classifiers (logistic & GDA)

- we will look at log reg & GDA; both of which generate a linear decision boundary to separate data into two classes

Log Reg : discriminative linear classifier

GDA : generative linear classifier.

- data set  $\{x_i, y_i\}$  where  $x_i = \{x_{i0}, x_{i1}\}$

$$\vdots \quad y_i = g_i$$

- we will investigate using log reg & GDA to perform bin classification

(a) empirical loss for log reg :

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

where  $y^i \in \{0, 1\}$ ,  $h_\theta(x) = g(\theta^T x)$  and

$$g(z) = \frac{1}{1+e^{-z}}$$

$$\bullet J(\theta) = ?$$

①

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

~~$x = 10$~~   $x = h_\theta$   $v = h_\theta(x)$

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

~~$v = \log(u) = \log'(u) \cdot u^{(x)}$~~

$$= \frac{1}{v} \cdot u^{(x)}$$

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

~~$= \sum_{i=1}^m -\frac{1}{m} y \frac{1}{h_\theta(x)} \cdot h'_\theta(x) + (1-y) \frac{1}{1-h_\theta(x)} \cdot h'_\theta(x)$~~

~~$= \sum_{i=1}^m h'_\theta(x) \left( -\frac{1}{m} y \frac{1}{h_\theta(x)} + (1-y) \frac{1}{1-h_\theta(x)} \right)$~~

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

$$h_\theta(x) = g(\theta^T x) \quad z = \theta^T x$$

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

$$v = g(z)$$

$$= -\frac{1}{m} \sum_{i=1}^m y \log(v) + (1-y) \log(1-v)$$

$$= -\frac{1}{m} \sum_{i=1}^m y \frac{1}{v} \cdot v' + (1-y) \frac{1}{1-v} \cdot (1-g(z)) \cdot v'$$

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

$$= -\frac{1}{m} \sum_{i=1}^m y \frac{1}{g(z)} \cdot g(z)(1-g(z)) \frac{1}{1-g(z)} \cdot g(z)(1-g(z)) \times$$

$$= -\frac{1}{m} \sum_{i=1}^m y (1-g(z)) \cancel{(1-y)(g(z))} x \Rightarrow = \frac{1}{m} \sum_{i=1}^m x (g(\theta^T x) - y)$$

~~$= \cancel{\left( y - y g(z) + g(z) - y g(z) \right)} x + \cancel{\left( y g(z) - y g(z) \right)}$~~

$$= -\frac{1}{m} \sum_{i=1}^m x (y(1-g(z)) - (1-y)g(z))$$

~~$= -\frac{1}{m} \sum_{i=1}^m x (y - yg(z) - yg(z) + yg(z))$~~

$$= -\frac{1}{m} \sum_{i=1}^m x (y - g(z))$$

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

$$\frac{\frac{1}{m} \sum_{i=1}^m (g(z) - y) x}{1 - \frac{1}{m} \sum_{i=1}^m (g(z) - y) x}$$

$$\frac{\partial}{\partial \theta} g(\theta^T x)$$

OK

$$\therefore \nabla_{\theta} J(\theta) = \frac{1}{m} X^T (g(X\theta) - y)$$

$$\frac{\partial \theta}{\partial j} = \frac{1}{m} \sum_{i=1}^m [g(\theta^T x^i) - y_i] x_j^i$$

$$H_{jk} = \frac{\partial^2 J(\theta)}{\partial \theta_j \partial \theta_k}$$

$$\begin{aligned} \frac{\partial^2 \theta}{\partial j \partial k} &= \frac{\partial \theta}{\partial \theta_k} \frac{1}{m} \sum_{i=1}^m [g(\theta^T x^i) - y_i] x_j^i \\ &= \frac{\partial \theta}{\partial k} \frac{1}{m} \sum_{i=1}^m [y(z) - y_i] x_j^i \\ &= \frac{1}{m} \sum_{i=1}^m y'(z) = y(z)(1 - y(z)) x_j^i \end{aligned}$$

$$\frac{\partial^2 \theta}{\partial j} = \frac{1}{m} \sum_{i=1}^m [g(\theta^T x^i) - y_i] x_j^i \Rightarrow g(\theta^T x^i)^j$$

$$= \frac{1}{m} \sum_{i=1}^m [g(z) - y_i] x_j^i$$

$$= \frac{1}{m} \sum_{i=1}^m [y(z) - y_i] x_j^i$$

$$\frac{\partial y(\theta^T x^i)}{\partial \theta_k} = g(\theta^T x^i)(1 - g(\theta^T x^i)) \cdot x_k^i$$

$$= \frac{1}{m} \sum_{i=1}^m [g(z)(1 - g(z)) - y_i] x_j^i x_k^i$$

$$= \frac{1}{m} \sum_{i=1}^m [(g(\theta^T x^i) - y_i)] x_j^i x_k^i$$

$$= \frac{1}{m} \sum_{i=1}^m [(g(\theta^T x^i) - y_i)] x_j^i x_k^i$$

$$H = \frac{1}{m} \sum_{i=1}^m [g(\theta^T x^i)(1 - g(\theta^T x^i)) - y_i] x_j^i x_k^i$$

$$= H = \frac{1}{m} [X^T \cdot g(X\theta) \cdot (1 - g(X\theta))] X$$

1c) GDA is

• show that GDA results in a classifier that has a linear decision boundary; Show that the posterior distribution can be written as:

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

$$p(y) = \begin{cases} \phi & \text{if } y=1 \\ 1-\phi & \text{if } y=0 \end{cases}$$

$$P(X|y=1) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (X - \mu_1)^T \Sigma^{-1} (X - \mu_1)\right)$$

$\Sigma$  is the covariance matrix:

$$\Sigma = \begin{bmatrix} \text{Var}(x_1) & & & \\ \text{Cov}(x_1, x_2) & \text{Var}(x_2) & \dots & \\ & & \ddots & \end{bmatrix}$$

- diagonals  
=  $\text{Var}(x_i) R$
- off diagonals  
=  $\text{Cov}(x_i, x_j)$

$$P(X|y=0) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2} (X - \mu_0)^T \Sigma^{-1} (X - \mu_0)\right)$$

$$P(X|Y=1) = \frac{1}{(\sqrt{2\pi})^{\frac{n}{2}} |\Sigma|^{-\frac{1}{2}}} \exp\left(-\frac{1}{2} (X - \mu)^T \Sigma^{-1} (X - \mu)\right)$$

Bayes

$$P(Y=1|X) = \frac{P(X|Y=1) P(Y=1)}{P(X)}$$

$$P(Y=1|X) =$$

(2) apply rule

$$P(Y=1|X) = \frac{P(X|Y=1) P(Y=1)}{P(X|Y=1) P(Y=1) + P(X|Y=0) P(Y=0)}$$

③ Sub in for prior

$$= \frac{P(X|Y=1) \phi}{P(X|Y=1) \phi + P(X|Y=0) (1-\phi)}$$

$$\phi (Y=1|X) = \frac{P(X|Y=1) \phi}{P(X|Y=1) \phi + P(X|Y=0) (1-\phi)} = \frac{\phi (X|Y=1) \phi}{\phi (X|Y=1) \phi + (1-\phi) (1-\phi)}$$

$$\phi (Y=1|X) = \frac{\phi (X|Y=1) \phi}{\phi (X|Y=1) \phi + (1-\phi) (1-\phi)}$$

$$= \phi(X|y=1) \phi \left( 1 + \frac{\rho(X|y=0)(1-\phi)}{\rho(X|y=1)\phi} \right)$$

• plug back in

$$\overline{\rho(X|y=1)\phi}$$

$$\overline{\rho(X|y=1)\phi} \left( 1 + \frac{\rho(X|y=0)(1-\phi)}{\rho(X|y=1)\phi} \right)$$

$$= \frac{1}{1 + \frac{\rho(X|y=0)(1-\phi)}{\rho(X|y=1)\phi}} \Rightarrow \frac{1}{2\pi^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \\ \text{cancels out...}$$

$$= \frac{1}{1 + \frac{\exp(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1} (x-\mu_0)) (1-\phi)}{\exp(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1} (x-\mu_1)) \phi}} \quad \begin{cases} \exp x = e^x \\ \text{so } \frac{\exp(x)}{\exp(y)} = \frac{e^x}{e^y} = e^{x-y} \\ = \exp(x-y) \end{cases}$$

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

$$\text{we want } \exp(-(\theta^T x + \theta_0))$$

• pull out  $\theta_i$ ; so we got  $\mu_0$  &  $\mu_1$ ... as from  $\theta^T$

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

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

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

- transposes of vectors are scalars & therefore equal

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

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

$$= \frac{1}{2} x^2 - \frac{1}{2} y^2$$

$$= \frac{1}{2} (x+y)(x-y)$$

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

$$\bullet \text{ add } \left( \frac{1-\varphi}{\varphi} \right) \text{ to exp} = \exp \left( \ln \left( \frac{1-\varphi}{\varphi} \right) \right)$$

$$=$$

$$\frac{1 + \exp(\Sigma^{-1}(\mu_0 - \mu_1)^T \chi + \frac{1}{2}(\mu_0 + \mu_1)^T \Sigma^{-1} (\mu_0 - \mu_1) - \ln(\frac{1-\phi}{\phi})\beta)}{1}$$

$$\Theta^T \chi$$

$$\Theta_0$$

$$\Theta = \Sigma^{-1}(\mu_0 - \mu_1)$$

1d) on notebook

1f) For Dataset 1 create a plot of training data with  $x_1$  on horizontal axis, and  $x_2$  on vertical axis.  
on notebook

1g) on notebook

2). of all dataset; we have a subset of all positive examples. All negative examples & the rest of positive examples are unlabelled.

assume a dataset

$$\{(x^i, f^i, y^i)\}_{i=1}^m \text{ where } f^i \in \{0, 1\}$$

is true label and vice versa if  $y^i = \begin{cases} 1 & x^i \text{ is black} \\ 0 & \text{otherwise} \end{cases}$ .

goal is to construct a binary classifier  $h_j$  of the true label +, with only access to partial labels  $y$ .

2A)

• show that  $p(t^i=1|x^i) = p(y^i=1|x^i)/2$

where  $d$  is a factor  $d \in \mathbb{R}$

(Given)

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

$$\because y=1 \Rightarrow t=1 \quad \begin{matrix} \text{* } y=1 \text{ implies} \\ t=1 \end{matrix}$$

$$\underline{p(y=1|x) = p(t=1|x) \cdot p(y=1|t=1, x)}$$

$y=1 \Rightarrow t=1$  tells us that  $p(y=1|x)$  is dependent on the above.

• You know if  $t^i=1$ ; then  $y^i$  is independent of  $x$ .

$$p(y=1|t=1, x) = p(y=1|t=1)$$

$$\therefore p(y=1|x) = p(t=1|x) \cdot p(y=1|t=1)$$

$$\therefore p(t=1|x) = \frac{p(y=1|x)}{p(y=1|t=1)}$$

• Show  $p(t^i=1|x^i) = p(y^i=1|x^i)/2$   
 where  $\frac{1}{2}$  is a constnt

$$\Rightarrow p(t=1|x) = \frac{p(y=1|x)}{p(y=1|t=1)}$$

$\underbrace{\hspace{10em}}$

is a constant; does not depend  
on  $x$ .

$$\therefore p(t=1|x) = \frac{p(y=1|x)}{2}$$

---

b) Suppose we want to estimate  $\frac{1}{2}$  using a trained classifier  $h$ , and a held-out validation set  $V$ . Let  $V_f$  be the set of labeled & pose examples in  $V$ , given by  $V_f = \{x^i \in V | y^i = 1\}$ .  
 Assuming that  $h(x^i) \approx p(y^i=1|x^i)$  for all examples  $x^i$ , show that

$$h(x^i) \approx \frac{1}{2} \text{ for all } x^i \in V_f$$

• You may assume that  $p(t^i=1|x^i) \approx 1$  when  $x^i \in V_f$ .

thoughts:  $h(x) \Rightarrow$  trained classifier

Show:  $h(x^i) \approx 2$ ; for all  $x^i \in V_t$

$h(x^i) \approx p(y^i = 1 | x^i)$  for all examples  $x^i$

$\Rightarrow h(x) \approx p(y=1 | x)$

\* from the fact that  $h(x)$  is a trained classifier to approx:  
 $\Rightarrow p(y=1 | x)$

$\Rightarrow p(t=1 | x) = \underline{p(y=1 | x)}$

$p(y=1 | x) = 2 p(t=1 | x)$

$h(x) \approx p(y=1 | x) = 2 p(t=1 | x)$

$\therefore h(x) \approx 2 p(t=1 | x)$

$= 1$ ; when  $x^i \in V_t$

$\therefore h(x) \approx 2$

c) Coding Problem:

$\Rightarrow$  in notebook.

d) in notebook

e) in notebook

### 3) Poisson Regression

Consider the Poisson distribution parameterized by  $\lambda$ :

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

Show that this distribution is in the exponential family, and clearly state the values for  $b(y)$ ,  $\eta$ ,  $T(y)$ , and  $a(\eta)$

#### GLM

• Poisson distribution  $\Rightarrow$  good for modeling # visitors. Poisson is an exponential family dist; so we can apply a GLM

• To derive a GLM; we need:

3 assumptions:

1)  $y | X; \theta \sim \text{ExponentialFamily}(\eta)$ ; that is; given  $X$  and  $\theta$ , the distribution of  $y$  follows some exponential family distribution, with param  $\eta$ .

$$2) h(x) = E[y|X]$$

that is...  $h(x)$  models the conditional expectation of the target variable  $y$ ; based on the given input  $X$ .

• interpretation of the hypothesis  $h(x)$  is predicting the average value of  $y$  for a given  $x$ .

\* remember; the expectation of a random variable is its mean.

\*  $h(x) = E[y|x]$  therefore say the model is

try to predict the average value of  $y$  for each given  $x$ .

In more b/c in regression, we often assume that  $y$  can be modelled by its mean for a given  $x$ ; assuming that deviations <sup>of</sup> are random & have a mean of 0 (from this mean)

3) The natural parameter  $\eta$  and the inputs  $x$  are related linearly  $\eta = \theta^T x$

\* Not a very justified assumption; more of a "design choice" for this type of model.

### Exponential family

a class of distributions is in the exponential family, if it can be written as

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

$\Rightarrow \eta$  = natural param

$T(y)$  is sufficient statistic =  $y$  for all cases

$a(\eta)$ : log partition function

$e^{-a(\eta)}$  plays role of a normalization constn;  
makes sure distribution  $P(y; \eta)$  integrates to 1

$\Rightarrow$  Write Bernoulli as form

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

$\Rightarrow$  Bernoulli is defined as

$$P(Y; \phi) = \phi^y (1-\phi)^{1-y}$$

$$\ln(\phi^y (1-\phi)^{1-y}) = \text{the number } e \text{ has been raised to, to get } \phi^y (1-\phi)^{1-y}$$

$$\exp(\ln(\phi^y (1-\phi)^{1-y})) = \phi^y (1-\phi)^{1-y}$$

$$\ln(\phi^y (1-\phi)^{1-y})$$

• apply the sum of logs

$$\ln(\phi^y) + \ln(1-\phi)^{1-y}$$

• apply rule of exponents for logs (they come down)

$$y \ln(\phi) + (1-y) \ln(1-\phi)$$

$$\Rightarrow \exp(y \ln \phi + (1-y) \ln(1-\phi))$$

$$\exp(\underbrace{y \ln \phi}_{\text{log m b sub}} + \ln(1-\phi) - \underbrace{y \ln(1-\phi)}_{\text{log m b sub}})$$

$$\exp(y(\ln \phi - \ln(1-\phi)) + \ln(1-\phi))$$

log m b sub

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

swap ln with log ...

$$= \exp\left(y \log\left(\frac{\phi}{1-\phi}\right) + \log(1-\phi)\right)$$

$$\therefore b(y) = 1 \quad T(y) = y \quad \checkmark$$

$$a(n) = -\log(1-\phi) \quad \checkmark$$

$$\eta^T = \log\left(\frac{\phi}{1-\phi}\right) \quad \checkmark !$$

Back to 3...

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

$$= \exp\left(\ln\left(\frac{e^{-\lambda} \lambda^y}{y!}\right)\right)$$

$$= \exp\left(\ln(e^{-\lambda}) + \ln(\lambda^y) - \ln(y!)\right)$$

$$\begin{aligned} &\Rightarrow \ln(e^{-\lambda}) \\ &\Rightarrow \ln(e^{-\lambda}) + \ln(\lambda^y) \\ &\Rightarrow -\lambda \ln(e) + y \ln(\lambda) \end{aligned}$$

$$3c) = \exp(-\lambda \ln(e) + y \ln(\lambda) - \ln(y!))$$

$$-\lambda \ln(e) + y \ln(\lambda) - \ln(y!)$$

$$\Rightarrow e^{-\lambda \ln(e)} \cdot e^{y \ln(\lambda)}$$

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

$$\frac{e^{-\lambda} \cdot e^{y \ln(\lambda)}}{y!}$$

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

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

$$T(y) = y$$

$$\eta = \ln(\lambda)$$

$$a(\eta) = \lambda = e^\eta$$

$$b/c \quad * \quad \begin{aligned} \eta &= \ln(\lambda) \\ e^\eta &= \lambda \end{aligned}$$

3b) Consider performing regression using a GLM model with a Poisson response var. What is the canonical response function for the family. You may use the fact that a Poisson random variable with parameter  $\lambda$  has mean  $\lambda$ .

$\Rightarrow \ln(\lambda)$  is a canonical response function.

$\Rightarrow$  it's a function that directly relates the natural param.  $\eta$  to its mean.

So its mean is  $\lambda$ .  
and natural param from 3a is  $\ln(\lambda)$

$$\therefore f(\eta) = \mu$$

$$\bullet \mu = \lambda \Rightarrow f(\ln(\lambda)) = \lambda$$

$$\bullet \eta = \ln(\lambda) \quad \therefore e^\eta = \lambda \\ e^\eta = \mu$$

$$\Rightarrow e^\eta = \lambda$$

• assumption 3 of GLMs ... the natural param  $\eta$  and inputs  $x$  are related linearly ... aka  $\eta = \theta^T x$

$$\boxed{\therefore e^\eta = \lambda = e^{\theta^T x}}$$

3c) For a training set  $\{(x^i, y^i)\}_{j=1, \dots, m}$ , let the log likelihood of an example be  $\log p(y^i | x^i; \theta)$ .

By taking the derivative of the log likelihood with respect to  $\theta_j$ ; derive the stochastic gradient ascent update rule for  $\theta$  using a GLM model with Poisson responses  $y$ , and the canonical response function

## thoughts

- when training a model using stochastic gradient descent or ascent; we optimize a likelihood function b/c. it's the objective function.
- basically to do gd. or g.a., we need an objective function.
- a good objective function is likelihood; because it measures how well our model  $\Theta$  fits the data. log-likelihood is its equivalent version that's easier to calculate.
- Once we have log likelihood function; we can use that to derive the update rule.

$$\text{log likelihood} : l(\theta) = \log P(y|x; \theta)$$

- log likelihood = how well our model;  $\Theta$  fits the data. To optimize it; we take its gradient w/ respect to  $\theta$ .

$$\nabla_{\theta} l(\theta) = \nabla_{\theta} \log P(y|x, \theta)$$

• what this tells us is the direction of steepest ascent or descent

- With the gradient we derive the update rule.

$$\theta^{(t+1)} = \theta^{(t)} + \alpha \nabla_{\theta} \log P(y|x; \theta)$$

\*stop forgetting log-likelihood  $\rightarrow$  gradient  $\rightarrow$  update rule

\* And if you want log likelihood:

1) Define probability models (linear regression = Gaussian)  
(logistic regression = Bernoulli)

2) Write likelihood probability of observing the given data under model parameters:

$$L(\theta) = P(D|\theta) = \prod_{i=1}^N P(y_i|x_i; \theta)$$

↑  
for independent examples; it's the product over all  
train examples

Back to 3...

$$3c) p(y; \theta) = \frac{e^{\theta^T x} y}{y!} ; \quad \theta = e^{\theta^T x} ; \quad = \frac{e^{-e^{\theta^T x}} e^{\theta^T x} y}{y!} = \frac{\exp(\theta^T x y - e^{\theta^T x})}{y!} = p(y|x; \theta)$$

$$\begin{aligned} \therefore l(\theta) &= \log \left( \frac{\exp(\theta^T x y - e^{\theta^T x})}{y!} \right) = \log \frac{1}{y!} \exp(\theta^T x y - e^{\theta^T x}) \\ &= \log \left( \frac{1}{y!} \right) + \theta^T x y - e^{\theta^T x} \\ &\stackrel{\text{quotient rule}}{=} -\log y! + \theta^T x y - e^{\theta^T x} \end{aligned}$$

$$\begin{aligned} \nabla l(\theta) &= \frac{\partial}{\partial \theta} -\log y! + \frac{\partial}{\partial \theta} \theta^T x y - e^{\theta^T x} \\ &= \frac{\partial}{\partial \theta} (\theta^T x y - e^{\theta^T x}) \\ &= \frac{\partial}{\partial \theta} \theta^T x y - \frac{\partial}{\partial \theta} e^{\theta^T x} \end{aligned}$$

$$\Rightarrow \frac{\partial}{\partial \theta} e^{\theta^T x} = \frac{\partial}{\partial \theta} e^v \cdot \frac{\partial}{\partial v} y = e^{\theta^T x} \cdot x = \underline{x e^{\theta^T x}}$$

$$\frac{\partial}{\partial \theta} \theta^T x = x$$

$$= x y - x e^{\theta^T x} = \underline{xy - \exp(\theta^T x)x} = x(y - \exp(\theta^T x))$$

$$\therefore \underline{\nabla l(\theta) = (y - e^{\theta^T x})x}$$

$$\Rightarrow \text{apply update rule} \therefore \theta^{t+1} = \theta^t + d \nabla l(\theta)$$

$\theta^{t+1} = \theta^t + d \cdot (y - e^{\theta^T x})x$

d) in notebook

• the predict function naturally follows the canonical response functions in GLM's

• recall that the model predicts  $\eta = \mathbf{x}^T \boldsymbol{\theta}$

• canonical response function  $\circ g^{-1}(\eta)$  maps  $\eta$  to the expected value of  $y$ :

$$E[y|\mathbf{x}] = g^{-1}(\mathbf{x}^T \boldsymbol{\theta})$$

• this is what predict function should return.

$\Rightarrow$  why? Because in GLM's the core assumption is that modelling the expected value  $E[y|\mathbf{x}]$  is good enough  $\approx$  representation of actual  $y$ .

$\Rightarrow$  And canonical response function ensures that this expected value is properly linked to  $\mathbf{x}^T \boldsymbol{\theta}$  in a way that fits.

4) Convexity of GLM's,

• Most commonly GLM's are trained using negative-log likelihood (NLL) as the loss funcn. This is mathematically equivalent to Maximum likelihood estimation. In this problem, we want to show that the NLL loss of a GLM is a convex function wrt the model parameters.  
• This is convenient because a convex function is one for which any local minimum is also a global maximum.

Exponential family distribution is one whose prob. density can be modelled as

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

In order to show that NLL loss is convex for GLMs, we break into subparts. First I want to show that the second derivative (Hessian) of the loss w.r.t. model parameters is positive semi-definite at all values of the model parameters.

- restrict to where  $\eta$  is a scalar. Assume  $p(y|X; \theta) \sim \text{Exponential family}(\eta)$  &  $\eta \in \mathbb{R}$  is a scalar &  $T(y) = y$

so...

$$p(y; \eta) = b(y) \exp(\eta y - a(\eta))$$

I) derive an expression for the mean of the distribution.

$$p(y; \eta) \Rightarrow \text{a pdf}$$

$\Rightarrow$  sum rule of PDF's (sum under curve adds up to 1)

$$\int p(y; \eta) dy = 1$$

$\Rightarrow$  take derivative wrt to  $\eta$

$$\frac{d}{d\eta} \int p(y; \eta) dy = 0 \quad (\text{derivative of a constant } 1 = 0)$$

$\Rightarrow$  take the hint

$$\int \frac{\partial}{\partial \eta} p(y; \eta) dy = 0$$

$$\int \frac{\partial}{\partial \eta} b(y) \exp(\eta y - a(\eta)) dy = 0$$



$$\frac{\partial}{\partial n} b(y) \exp(ny - a(n))$$

$$= b(y) \frac{\partial}{\partial n} (\exp(ny - a(n)) + \exp(ny - a(n))) \cdot \cancel{\frac{\partial}{\partial n} b(y)}$$

$$\therefore = b(y) \frac{\partial}{\partial n} \exp[ny - a(n)]$$

$\Rightarrow$  plug back in

$$\int b(y) \frac{\partial}{\partial n} \exp[ny - a(n)] dy = 0$$

$$\Rightarrow \frac{\partial}{\partial n} \exp[ny - a(n)]$$

$$\Rightarrow u = ny - a(n)$$

$$\therefore \frac{\partial}{\partial n} \exp[u] \frac{\partial}{\partial n} u$$

$$= \exp(u) \frac{\partial}{\partial n} u$$

$$= \exp(ny - a(n)) \frac{\partial}{\partial n} ny - a(n)$$

$$= \exp(ny - a(n)) \cdot y - \frac{\partial}{\partial n} a(n)$$

$\Rightarrow$  plug back in

$$\int [b(y) \exp(ny - a(n)) \cdot \underbrace{y - \frac{\partial}{\partial n} a(n)}_{= p(y; n)}] dy = 0$$

$$\int p(y; n) \cdot \{y - \frac{\partial}{\partial n} a(n)\} dy = 0$$

$$\Rightarrow \int p(y; \eta) y - p(y; \eta) \frac{\partial}{\partial \eta} a(\eta) dy = 0$$

$$= \int p(y; \eta) y - \int p(y; \eta) \frac{\partial}{\partial \eta} a(\eta) dy = 0$$

2

$$\int p(y; \eta) dy = E[Y; \eta]$$

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

2

$a(\eta)$  is independent of  $y$ , so we can split the integral; b/c to the integral; an independent var is like a const +

$$\int c f(y) dy = c \cdot \int f(y) dy \dots$$
$$\text{so } \frac{\partial}{\partial \eta} a(\eta) = c \text{ in this case}$$

$$0 = E[Y; \eta] - \int p(y; \eta) dy \cdot \frac{\partial}{\partial \eta} a(\eta)$$

2

$$0 = E[Y; \eta] - \frac{\partial}{\partial \eta} a(\eta)$$

$$\boxed{\frac{\partial}{\partial \eta} a(\eta) = E[Y; \eta] = E[Y|X; \theta]}$$

$\eta$  relates to  $X; \theta$

4b) Derive an expression for the variance of the distribution.  
 In particular, show that  $\text{Var}(Y|X; \theta)$  can be expressed as  
 the derivative of the mean w.r.t  $\eta$  (i.e., the second derivative of  
 the log-partition function  $a(\eta)$  w.r.t the natural parameter  
 $\eta$ ).

$$E[Y|X; \theta] = \frac{\partial}{\partial \eta} a(\eta) - \text{VAR}[Y] = E[Y^2] - E[Y]^2.$$

$$\Rightarrow \frac{\partial}{\partial \eta} \int y p(y; \eta) dy = \frac{\partial^2 a(\eta)}{\partial \eta^2} \quad \Rightarrow \text{Show: } \frac{\partial}{\partial \eta} \int y p(y; \eta) dy \\ = \text{VAR}[Y|X; \theta]$$

$$\Rightarrow \int \underbrace{\frac{\partial}{\partial \eta} y p(y; \eta) dy}$$

$$\frac{\partial}{\partial \eta} y p(y; \eta) = \frac{\partial}{\partial \eta} b(y) \exp(\eta y - a(\eta)) y$$

$$= y b(y) \underbrace{\frac{\partial}{\partial \eta} \exp(\eta y - a(\eta))}_{\frac{\partial}{\partial \eta} \exp(u) \cdot \frac{\partial}{\partial \eta} u}$$

$$= \exp(\eta y - a(\eta)) \cdot \frac{\partial}{\partial \eta} \eta y - a(\eta)$$

$$= \exp(\eta y - a(\eta)) \cdot \frac{\partial}{\partial \eta} \eta y - \frac{\partial}{\partial \eta} a(\eta)$$

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

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

$\Rightarrow$  from Exponential family representation

$$p(y; \eta) = b(y) \exp(\eta y - a(\eta))$$

$$= y p(y; \eta) \left( y - \frac{\partial}{\partial \eta} a(\eta) \right)$$

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

↳ plug back in:

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

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

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

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

$$= \text{Var}[Y; \eta]$$

$$\therefore \text{Var}[Y; \eta] = \frac{\partial^2 a(\eta)}{\partial \eta^2} = \text{Var}[y | x; \theta]$$

4c) Write out the loss function  $l(\theta)$ , the NLL of the distribution, as a function of  $\theta$ . Then, calculate the Hessian of the loss w.r.t.  $\theta$  & show that it is always PSD. This concludes the proof that NLL loss of GLM is convex.

$\Rightarrow$  (GLM's  $\theta$  linked to  $E[X]$  via canonical link function  
With NLL, consider Hessian to establish convexity ...

$$P(Y|X; \theta) \sim \text{Exponential Family}(\eta)$$

$$\Rightarrow P(y; \eta) = b(y) \exp(\eta y - a(\eta))$$

$\Rightarrow$  canonical response function :

$$g(\eta) + \eta = \theta^T X \quad (\text{in GLM's}) \\ \therefore M = g^{-1}(\theta^T X)$$

• goal given  $P(y; \eta)$  find the NLL ...

$$\Rightarrow L(\theta) = P\left(\prod_{i=1}^m y_i | X_i; \theta\right)$$

$$\text{so: } L(\eta) = P(y; \eta)$$

$$L(\eta) = \prod_{i=1}^m P(y_i; \eta)$$

$$\Rightarrow L(\theta) = \prod_{i=1}^m P(y_i | X_i; \theta)$$

$$\text{NLL} = l(\theta) = -\log \left( \prod_{i=1}^m P(y_i | X_i; \theta) \right)$$

$$\log(a \cdot b) = \log(a) + \log(b)$$

! "PSD" : positive semi-definite  $\Rightarrow$  to be PSD  
! over all values of model parameters, it means for any vector,  $v$   
!  $v^T H v \geq 0$  for all values of  $v$  and for all values of  $\theta$ .  
! This guarantees convexity.

$$\begin{aligned}
 \Rightarrow l(\theta) &= -\sum_{i=1}^m \log p(y_i | x_i; \theta) \\
 &= -\sum_{i=1}^m \log (b(y_i) \exp (\theta^\top x_i y_i - a(\theta^\top x_i))) \\
 &= -\sum_{i=1}^m \log b(y_i) + \log (\underbrace{\exp(\theta^\top x_i y_i) - a(\theta^\top x_i)}_{e^{\theta^\top x_i y_i - a(\theta^\top x_i)}}) \\
 &= \log (e^{\theta^\top x_i y_i - a(\theta^\top x_i)}) = e^{\theta^\top x_i y_i - a(\theta^\top x_i)}
 \end{aligned}$$

$$\begin{aligned}
 &= -\sum_{i=1}^m \log b(y_i) + \exp (\theta^\top x_i y_i - a(\theta^\top x_i)) \\
 &= \sum_{i=1}^m -\log b(y_i) - \exp (\theta^\top x_i y_i - a(\theta^\top x_i)) \\
 \frac{\partial}{\partial \theta} l(\theta)
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{\partial}{\partial \theta} \left[ \sum_{i=1}^m -\log (b(y_i)) - \exp (\theta^\top x_i y_i - a(\theta^\top x_i)) \right] \\
 &= \sum_{i=1}^m \frac{\partial}{\partial \theta} \left( -\log (b(y_i)) - \exp (\theta^\top x_i y_i - a(\theta^\top x_i)) \right) \\
 &= \sum_{i=1}^m \frac{\partial}{\partial \theta} -\log (b(y_i)) - \frac{\partial}{\partial \theta} \exp (\theta^\top x_i y_i - a(\theta^\top x_i)) \\
 &= \sum_{i=1}^m -\frac{\partial}{\partial \theta} \exp (\theta^\top x_i y_i - a(\theta^\top x_i))
 \end{aligned}$$

$$\begin{aligned}
 \Rightarrow u &= \theta^\top x_i y_i - a(\theta^\top x_i) \\
 &= -\frac{\partial}{\partial \theta} e^u \cdot \frac{\partial}{\partial \theta} u = e^u \cdot \underbrace{\frac{\partial}{\partial \theta} u}_{u}
 \end{aligned}$$

$$\begin{aligned}
 \frac{\partial}{\partial \theta} u &= \frac{\partial}{\partial \theta} (\theta^\top x_i y_i - a(\theta^\top x_i)) \\
 &= \frac{\partial}{\partial \theta} \theta^\top x_i y_i - \frac{\partial}{\partial \theta} a(\theta^\top x_i)
 \end{aligned}$$

$$= xy - \underbrace{\frac{\partial}{\partial \theta} a(\theta^T x)}_{u}$$

$$u = \theta^T x$$

$$\begin{aligned}\frac{\partial}{\partial \theta} a(u) &= a'(\theta^T x) \cdot \frac{\partial}{\partial \theta} \theta^T x \\ &= \underline{a'(\theta^T x) \cdot x}\end{aligned}$$

$$= xy - a'(\theta^T x) x$$

$$= \sum_{i=1}^m xy - a'(\theta^T x) x$$

$$\frac{\partial}{\partial \theta} l(\theta) = \sum_{i=1}^m x [y - a'(\theta^T x)]$$

$$H = \frac{\partial}{\partial \theta} \left( \frac{\partial}{\partial \theta} l(\theta) \right) = \underbrace{\frac{\partial}{\partial \theta} \sum_{i=1}^m x [y - a'(\theta^T x)]}_{-}$$

$$= \sum_{i=1}^m \frac{\partial}{\partial \theta} (x [y - a'(\theta^T x)])$$

$$= \sum_{i=1}^m \cancel{\frac{\partial}{\partial \theta} xy} - \frac{\partial}{\partial \theta} a'(\theta^T x) x$$

$$= \sum_{i=1}^m - \underbrace{\frac{\partial}{\partial \theta} a'(\theta^T x)}_{x} x$$

$$u = \theta^T x$$

$$= - \frac{\partial}{\partial \theta} x a'(u) \cdot \frac{\partial}{\partial \theta} u$$

$$= x a''(u) \cdot \frac{\partial}{\partial \theta} [\theta^T x]$$

$$= \alpha''(u) \cdot x$$

$$= (\alpha''(\theta^T x) x) x$$

$$H = \text{Var}[\theta^T x] x^2$$

- Show that it is positive semi-definite

$$\Rightarrow v^T H v \geq 0$$

$$\Rightarrow v^T [\text{Var}[\theta^T x] x^2] v \geq 0$$

$$= \text{Var}[\theta^T x] x^2 \cdot v^2$$

$$= \text{Var}[\theta^T x] (x^T v)^2$$

$\underbrace{\quad}_{\text{Variance}}$

$\underbrace{\quad}_{\text{positive}}$

Variance  
is always  
positive

$$\therefore \text{pos} \times \text{pos} = \text{pos}$$

$$\therefore v^T H v \geq 0$$

## 5) Locally weighted Lin Reg.

a) Consider a linear regression problem in which we want to "weight" diff training examples differently. Specifically we want to minimize:

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

• In this problem; we will generalize some of those ideas to the weighted setting.

i. Show that  $J(\theta)$  can also be written ...

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

$\Rightarrow$  first note that  $W$  has to be a diagonal matrix

...  $X\theta - y \Rightarrow$  matrix

$$\begin{bmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \\ a_{20} & a_{21} & a_{22} \end{bmatrix}$$

• For each example  $m$ ; we get  
calculator of forward:

$$a_{i(0-3)}$$

$$\text{so for row } [a_{00} \ a_{01} \ a_{02}]$$

$$\text{second} = [a_{10} \ a_{11} \ a_{12}]$$

• we want to scale exclusively the product of all three weights by  $w$

$$\text{So... } w^T [a_{10} \ a_{11} \ a_{12}] \\ + \\ w^T [a_{20} \ a_{21} \ a_{22}] \\ \vdots$$

So to get this behaviour;  $w$  must be a diagonal matrix; otherwise  $w$  enters the initial weights column

i.e.  $w^T [a_{20} \ a_{21} \ a_{22} \cdot w]$ !

$$\text{So } w = \begin{bmatrix} w & 0 & 0 \\ 0 & w & 0 \\ 0 & 0 & w \end{bmatrix} = D$$


---

$\Rightarrow$  quadratic form  $a = \theta^T x - y$

$V W V^T$

✓

$$\Rightarrow \sqrt{1} \\ [a_1 \ a_2 \ a_3] \begin{bmatrix} w_{01} & 0 & 0 \\ 0 & w_{12} & 0 \\ 0 & 0 & w_{23} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \\ 1 \times 3 \qquad 3 \times 3 \qquad 3 \times 1$$

$$= [w_{01}(a_1 a_2 a_3), w_{12}(a_1 a_2 a_3), w_{23}(a_1 a_2 a_3)] \\ (1 \times 3)$$

$$= [w_{01}(a_1 a_2 a_3), w_{12}(a_1 a_2 a_3), w_{23}(a_1 a_2 a_3)] \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

$\underbrace{a_1}_{\alpha} \Rightarrow 1 \times 1 \underbrace{a_2}_{\alpha} \underbrace{a_3}_{\alpha}$

$$= w_{0,1} a_1 a_j + w_{0,2} a_2 a_j + w_{0,3} a_3 a_j$$

$$= \sum_{i=0}^n \sum_{j=0}^n w_{ij} a_i a_j$$

$$= \sum_{i=0}^n \sum_{j=0}^n w_{ij} (\theta^T x^i - y^i) (\theta^T x^j - y^j)$$

$$= W (\theta^T X - Y)^T (\theta^T X - Y)$$

$$= (\theta^T X - Y)^T W (\theta^T X - Y)$$

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

ii) If all the  $w^{(i)}$ 's equal 1, then we saw in class that the Normal equation is

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

and that the value of  $\theta$  that minimizes  $J(\theta)$  is given by  $(X^T X)^{-1} X^T y$ . By finding the derivative  $\frac{\partial}{\partial \theta} J(\theta)$

and setting that to zero, generate the normal equation to the weighted setting and give the new value of  $\theta$  that minimizes  $J(\theta)$  in closed form as a function of  $X$ ,  $W$ , and  $y$ .

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

$$\nabla J(\theta) = \frac{\partial}{\partial \theta} [(X\theta - y)^T W (X\theta - y)]$$

Review: Matrix derivative rules

1) Scalar with respect to a vector

$\frac{df}{dx} \Rightarrow f$  is a scalar function  $x$  is a vector

$$\frac{df}{dx} = \left[ \frac{df}{dx_1}, \frac{df}{dx_2}, \dots, \frac{df}{dx_n} \right]^T$$

• gradient is a vector, where  $f$  is derivated with respect to values of  $x$

2) Scalar with respect to a matrix.

$f \Rightarrow$  scalar function

$X \Rightarrow$  matrix

$\frac{\partial f}{\partial X^{ij}} = \frac{\partial f}{\partial x_{ij}} \Rightarrow$  the outcome is a matrix where each entry  $j$   $f$  is derivative with respect to  $X_{ij}$ .

• grad is same shape as  $X$ ; each element represents the partial derivative of  $f$  with respect to corresponding element of  $X$ .

3) derivatives of linear functions

$$\frac{d}{dx} Ax = A \quad ; \text{where } A \text{ is a constant matrix}$$

$x$  is input.

4) Derivative of Quadratic form:

$$\frac{d}{dx} (x^T A x) = 2Ax$$

$\hookrightarrow$  No linear shift (only multiplication)

5) Derivative of Transpose

$$f(x) = a^T x$$

$$\frac{d}{dx}(a^T x) = a$$

6) Derivative of Matrix-Vector Product

$$\frac{d}{dx}(Ax) = A$$

7) Derivative of an inverse Matrix

$$\frac{\partial}{\partial X} X^{-1} = -X^{-1} \left( \frac{\partial X}{\partial x} \right) X^{-1}$$

$\Rightarrow$  back to ii

$$\nabla J(\theta) = \frac{\partial}{\partial \theta} [(x\theta - y)^T w (x\theta - y)]$$

$$\nabla J(\theta) = \frac{\partial}{\partial \theta} [(x^T \theta^T - y^T) w (x\theta - y)]$$

$$= \frac{\partial}{\partial \theta} [w(x^T x \theta^T \theta - x^T \theta^T y - x \theta y^T + y y^T)]$$

$$= \frac{\partial}{\partial \theta} [w x^T x \theta^T \theta - w x^T \theta^T y - w x \theta y^T + w y y^T]$$

$$= \frac{\partial}{\partial \theta} [w x^T x \theta^T \theta - 2 y^T w x \theta]$$

$$= \frac{\partial}{\partial \theta} [\underline{\theta^T w x^T x \theta^T} - 2 y^T w x \theta]$$

$$= 2 x^T w x \theta - 2 y^T w x \theta \quad \begin{matrix} \text{quadratic form} \\ \theta^T x \theta = 2 x \theta \end{matrix}$$

$$= 2 x^T w x \theta - 2 x^T w y \quad \begin{matrix} \text{also symmetric} \\ b/c w! \end{matrix}$$

iii. Suppose we have a dataset  $\{(x^i, y^i); i = 1 \dots m\}$  of  $m$  independent examples, but we model each  $y^i$ 's as drawn from conditional dist. with diff levels of variance,  $(\sigma^i)^2$

$$p(y^i | x^i; \theta) = \frac{1}{\sqrt{2\pi\sigma^i}} \exp\left(-\frac{(y^i - \theta^T x^i)^2}{2(\sigma^i)^2}\right)$$

$$\Rightarrow l(\theta) = \sum \log p(y^i | x^i; \theta)$$

$$l(\theta) = \sum \log \left( \frac{1}{\sqrt{2\pi\sigma^i}} \exp\left(-\frac{(y^i - \theta^T x^i)^2}{2(\sigma^i)^2}\right) \right)$$

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

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

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

$$\Rightarrow \frac{\partial}{\partial \theta} l(\theta) = \frac{\partial}{\partial \theta} \left( -\log(\sqrt{2\pi\sigma^i}) - \frac{(y^i - \theta^T x^i)^2}{2(\sigma^i)^2} \right)$$

$$= \bar{-} \frac{\partial}{\partial \theta} \frac{(y^i - \theta^T x^i)^2}{2(\sigma^i)^2}$$

$$= \sum -\frac{1}{2(\sigma^i)^2} \frac{\partial}{\partial \theta} (y^i - \theta^T x^i)^2$$

$$= \bar{-} \frac{1}{2(\sigma^i)^2} \cdot 2(y^i - \theta^T x^i) \cdot x$$

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

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

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

also  $y^i - \theta^T x^i = \text{MSE}$

SB) weighted linear regression:

- in notebook

$$\theta^{t+1} = \theta^t - 2 \nabla l(\theta)$$

$\Rightarrow$  weighted linear regression  $\Rightarrow$  no explicit "training phase"; it is "non-parametric" meaning the model is retrained dynamically for each query upon  $x^i$  at test time. This is because we don't train a global model once and never it for all predictions.

$\Rightarrow$  model is retrained dynamically

$\Rightarrow$  also loss function allows you to get  $\theta$  directly

$$\nabla J(\theta) = 2x^T W X \theta - 2x^T W y$$

$$0 = 2x^T w \times \theta - 2x^T w y$$

$$2x^T w x \theta = 2x^T w y$$

$$x^T w x \theta = x^T w y$$

$\Rightarrow$

$$x^T w x \theta (x^T w x)^{-1}$$

=

$$x^T w y (x^T w x)^{-1}$$

\* to isolate  $\theta$ ;  
we can't just  
divide; because  
these are matrices;  
instead mult. by  
inverse  
 $(x^T w x)^{-1}$

$$\therefore \theta = x^T w y (x^T w x)^{-1}$$

---

5c) ... in notebook