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 in to two classes

log Reg : discriminative linear classifier

GDA : generative linear classifier.

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

- 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

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

- $H(J(\theta)) = ?$

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

$x = h_\theta$ $\quad$ $u = h_\theta(x)$

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

$$\log(u)' = \log'(u)\cdot u'(x) = \frac{1}{u}\cdot u'$$

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

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

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

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

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

$v = g(z)$

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

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

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

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

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

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

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

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

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

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

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

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

$$= \frac{1}{m} X^T(g(X\theta) - Y)$$

$\frac{\partial}{\partial \theta} g(z) x$

ตอบ

$\therefore \nabla_\theta J(\theta) = \frac{1}{m} X^T (g(X\theta) - Y) \overset{\partial\theta/\partial j}{=} \frac{1}{m} \sum_{i=1}^{m} [g(\theta^T x^i) - y^i] x_j^i$

ok

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

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

$= \frac{\partial \theta}{\partial k} \frac{1}{m} \sum_{i=1}^{m} [g(z) - y] x_j$

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

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

$\Rightarrow g(\theta^T x^i)'$

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

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

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

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

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

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

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

(a)

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

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

1c) GDA

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

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

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

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

Σ is the covariance matrix:

$n \times n$

$$\Sigma = \begin{bmatrix} var(x_1) & & \\ cov(x_2,x_1) & var(x_2) & \cdots \\ & & \end{bmatrix}$$

- diagonals = variance
- off diagonals = covariance ...

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

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

① apply Bayes

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

② apply $P(x)$ 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 priors

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

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

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

- Plug back in

$$\frac{\cancel{p(x|y=1)\phi}}{\cancel{p(x|y=1)\phi}\left(1 + \frac{p(x|y=0)(1-\phi)}{p(x|y=1)\phi}\right)}$$

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

$\Rightarrow \frac{1}{2\pi^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}$ cancels out...

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

$\because \exp^x = e^x$

so $\frac{\exp(x)}{\exp(y)} = \frac{e^x}{e^y} = e^{x-y} = \exp(x-y)$

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

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

- pull out x; so we got μ_0 & μ_1... as form θ^T

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

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

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

- transposes of vectors are scalars & transpose equal

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

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

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

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

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

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

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

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

1d) on notebook

1f) For Dataset 1 create a plot of true 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

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

is true label and we $y^i = \begin{cases} 1 & x^i \text{ is labeled} \\ 0 & \text{otherwise.} \end{cases}$

goal is to construct a binary classifier h, of the true label t, with only access to partial labels y.

2A)

- show that $p(t^i=1|x^i) = p(y^i=1|x^i)/\alpha$

where α is a factor $\alpha \in \mathbb{R}$

Given:

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

* $y=1 \Rightarrow t=1$

* $y=1$ implies $t=1$

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

$y=1 \Rightarrow t=1$ 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 $p(t^i = 1 \mid x^i) = p(y^i = 1 \mid x^i)/\alpha$

where α is a constant

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

is a constant; does not depend on x.

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

---

b) Suppose we want to estimate α using a trained classifier h, and a held-out validation set V. Let V_+ be the set of labeled positive examples in V, given by $V_+ = \{x^i \in V \mid y^i = 1\}$

Assuming that $h(x^i) \approx p(y^i = 1 \mid x^i)$ for all examples x^i, show that

$$h(x^i) \approx \alpha \text{ for all } x^i \in V_+$$

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

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

show: $h(x^i) \approx \lambda$; for all $x^i \in V_t$

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

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

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

$\Rightarrow p(t = 1 \mid x) = p(y = 1 \mid x)$

$p(y = 1 \mid x) = \lambda p(t = 1 \mid x)$

$h(x) \approx p(y = 1 \mid x) = \lambda p(t = 1 \mid x)$

$\therefore h(x) \approx \lambda p(t = 1 \mid x)$

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

$\therefore h(x) \approx \lambda$

---

c) Coding Problem:

$\Rightarrow$ in notebook.

---

d) in notebook

---

e) in notebook

3) Poisson Regression

Consider the Poisson distribution parametrized by λ:

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

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

GLM

• Poisson distribution ⇒ 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 θ, the distribution of y follows some exponential family distribution, with param η.

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: 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.

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

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

• works b/c in regression, we often assume that y can be modeled by its mean for a given x; assuming that deviations are random & have a mean of 0 (from this meme)

3) The natural parameter η 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

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

• η => natural param

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

$a(\eta)$: log partition function

$e^{-a(\eta)}$: plays role of a normalization constant, makes sure distribution $p(y;\eta)$ integrates to 1

=> write Bernoulli as form

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

⇒ Bernoulli is defined as

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

$\ln(\phi^{y}(1-\phi)^{1-y})$ = the number e has to be 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)

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

⇒ $\exp(y\ln\phi + (1-y)\ln(1-\phi))$

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

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

log rule subtraction

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

swap ln with log...

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

$\therefore$ $b(y) = 1$ ✓ $T(y) = y$ ✓

$a(\eta) = -\log(1-\phi)$ ✓

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

Back to 3...

3a)

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

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

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

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

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

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

3a)

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

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

$$\Rightarrow e^{-\lambda\ln(e)} \cdot e^{y\ln(\lambda)} \cdot \frac{1}{e^{\ln(y!)}} \quad e^{\ln(y!)} = y!$$

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

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

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

$$T(y) = y$$

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

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

b/c * $\eta = \ln(\lambda)$

$e^{\eta} = \lambda$

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 λ has mean λ.

⇒ wts? is a canonical response function

⇒ its a function that directly relates the natural param. η to its mean.

so: mean is λ.
and natural param from 3a is $\ln(\lambda)$

∴ $f(\eta) = \mu$

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

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

∴ $e^{\eta} = \lambda$
$e^{\eta} = \mu$

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

- assumption 3 of GLM's ... the natural param η and inputs x are related linearly .. aka $\eta = \theta^T x$

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

3c) For a train set $\{(x^i, y^i); i = 1, \ldots m\}$ 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 θ_j; derive the stochastic gradient ascent update rule for using a GLM model with poisson responses y, and the canonical response funct

thoughts:

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

log likelihood: $\ell(\theta) = \log P(y|x;\theta)$

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

$$\nabla_\theta \ell(\theta) = \nabla_\theta \log P(y|x,\theta)$$

- 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 → gradient → update rule

* And if you want log likelihood:

1) Define Probability model (linear regression ⇒ gaussian)
(logistic regression ⇒ Bernoulli)

2) Write likelihoods: 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 training examples

Back to 3...

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

$\therefore \ell(\theta) = \log\left(\frac{\exp(\theta^T x y - e^{\theta^T x})}{y!}\right) = \log \frac{1}{y!} \exp(\theta^T x y - e^{\theta^T x})$

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

$= -\log y! + \theta^T x y - e^{\theta^T x}$

$\nabla \ell(\theta) = \frac{\partial}{\partial \theta} -\log y! + \frac{\partial}{\partial \theta} \theta^T x y - e^{\theta^T x}$

$= \frac{\partial}{\partial \theta}(\theta^T x y - e^{\theta^T x})$

$= \frac{\partial}{\partial \theta} \theta^T x y - \frac{\partial}{\partial \theta} e^{\theta^T x}$

$\Rightarrow \frac{\partial}{\partial \theta} e^{\theta^T x} = \frac{\partial}{\partial \theta} e^v \cdot \frac{\partial}{\partial v} = e^{\theta^T x} \cdot x = x e^{\theta^T x}$ ($v = \theta^T x$)

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

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

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

$\Rightarrow$ apply update rule: $\theta^{t+1} = \theta^t + \alpha \nabla \ell(\theta)$

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

d) in notebook

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

- recall that the model predicts $\eta = x^T\theta$
- canonical response function is $g^{-1}(\eta)$ maps η to the expected value of y:

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

- this is what predict func should return.

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

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

---

4) Convexity of GLM's

- Most commonly GLM's are trained using negative-log likelihood (NLL) as the loss func. 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 3 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 η is a scalar. Assume $p(y|x;\theta) \sim$ Exponential family (η) & $\eta \in \mathbb{R}$ is a scalar & $T(y) = y$)

so...

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

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

$p(y;\eta) \Rightarrow$ a pdf

$\Rightarrow$ from rule of pdf's (sum under curve adds up to 1

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

$\Rightarrow$ take derivate wrt to η

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

$\Rightarrow$ take the hint

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

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

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

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

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

⇒ plug back in

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

$$\Rightarrow \frac{\partial}{\partial \eta} \exp[\eta y - a(\eta)]$$

$$\Rightarrow u = \eta y - a(\eta)$$

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

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

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

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

⇒ plug back in

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

$$= p(y; \eta)$$

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

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

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

$\int p \cdot f_y \, dy = E[Y;\eta]$

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

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

$\int c f(y)\, dy = c \cdot \int f(y)\, dy$ ...

so $\frac{\partial}{\partial \eta} a(\eta) = C$ in this case

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

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

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

η relates to $X;\theta$

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

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

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

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

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

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

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

$\frac{\partial}{\partial \eta} \exp(u) \cdot \frac{\partial}{\partial \eta} u$

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

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

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

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

$\Rightarrow$ 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:

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

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

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

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

$$= Var[y;\eta]$$

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

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

⇒ GLM's θ linked to $E[Y]$ via canonical link function
- With NLL, consider Hessian to establish convexity...

$P(Y|X;\theta) \sim$ ExponentialFamily(η)

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

⇒ canonical response function:

$g(\eta)$ & $\eta = \theta^T x$ (in GLM's)

$\therefore \mu = g^{-1}(\theta^T x)$

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

$\Rightarrow L(\theta) = P(y|X;\theta)$

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

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

$L(\eta) = \prod_{i=1}^{n} P(y;\eta)$

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

$NLL = \ell(\theta) = -\log\left(\prod_{i=1}^{m} P(y|x;\theta)\right)$

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

"PSD": positive semi definite ⇒ 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 θ. This guarantees convexity.

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

$$= -\sum_{i=1}^{m} \log\left(b(y)\exp\left(\theta^T x\, y - a(\theta^T x)\right)\right)$$

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

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

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

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

$$\frac{\partial}{\partial\theta}\ell(\theta)$$

$$\cdots$$

$$= \frac{\partial}{\partial\theta}\left[\sum_{i=1}^{m} -\log(by) - \exp\left(\theta^T x y - a(\theta^T x)\right)\right]$$

$$= \sum_{i=1}^{m} \frac{\partial}{\partial\theta}\left(-\log(b(y) - \exp\left(\theta^T x y - a(\theta^T x)\right)\right)$$

$$= \sum_{i=1}^{m} \frac{\partial}{\partial\theta} \cancel{-\log(b(y))} - \frac{\partial}{\partial\theta}\exp\left(\theta^T x y - a(\theta^T x)\right)$$

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

$$\Rightarrow u = \theta^T x y - a(\theta^T x)$$

$$= -\frac{\partial}{\partial\theta} e^u \cdot \frac{\partial}{\partial\theta} u = e^u \cdot \frac{\partial}{\partial\theta} u$$

$$\frac{\partial}{\partial\theta} u = \frac{\partial}{\partial\theta}\left(\theta^T x y - a(\theta^T x)\right)$$

$$= \frac{\partial}{\partial\theta}\theta^T x y - \frac{\partial}{\partial\theta} a(\theta^T x)$$

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

$u = \theta^T x$

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

$$= a'(\theta^T x) \cdot x$$

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

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

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

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

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

$$= \sum_{i=1}^{m} \frac{\partial}{\partial \theta} \cancel{xy}^{0} - \frac{\partial}{\partial \theta} a'(\theta^T x) x$$

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

$u = \theta^T x$

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

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

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

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

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

- show that it is positive semi definite

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

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

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

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

Variance is always positive

positive

$\therefore$ pos x pos = pos

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

5) Locally weighted Lin Reg.

a) Consider a liner regression problm in which we want to "weight" diff trang examples diffently. Specifically we went to minimize:

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

- In this prob, we will generalize some of those idea to the weighted setting.

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

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

⇒ first note that W has to be a diagonal vectr...

... $X\theta - y$ ⇒ 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 calculus of forward:

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

so first row $[a_{00} \; a_{01} \; a_{02}]$

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

- we want to scale exclusively the product of all three weights by W

So... $W[a_{10}\ a_{11}\ a_{12}) + W[a_{20}\ a_{21}\ a_{22}) + \vdots$

So to get this behavior, W must be a diagonal matrix; otherwise W enters the initial weights column

i.e. $W[a_{20}\ a_{21}\ a_{22} \cdot W)$!

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

⇒ quadratic form $a = \theta^T x - y$

$V W V^T$

$\Rightarrow V^T$ $\begin{bmatrix} w_{01} & 0 & 0 \\ 0 & w_{12} & 0 \\ 0 & 0 & w_{33} \end{bmatrix}$ V $\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$

$[a_1 a_2 a_3]$ 1×3 3×3 3×1

$= [w_{01}(a_1 a_2 a_3),\ w_{12}(a_1 a_2 a_3),\ w_{03}(a_1 a_2 a_3))$ 1×3

$= [w_{01}(a_1 a_2 a_3), w_{12}(a_1 a_2 a_3), w_{03}(a_1 a_2 a_3))$ $\begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$

a ⇒ 1×1 a a

$$= w_{01} a_i a_j + w_{02} a_i a_j + w_{03} a_i a_j$$

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

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

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

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

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

---

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 θ that minimizes $J(\theta)$ is given by $(X^T X)^{-1} X^T y$. By finding the derivative $\nabla_\theta J(\theta)$ and setting that to zero, generate the normal equations to this weighted setting and give the new value of θ that min. $J(\theta)$ in closed form as a function of X, W, and y.

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

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

Review: Matrix derivate rules

1) Scalar with respect to a vector

$\frac{df}{dx}$ ⇒ 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 ⇒ scalar function

X ⇒ matrix

$\frac{\partial f}{\partial X}_{ij} = \frac{\partial f}{\partial X_{ij}}$ ⇒ the outcome is a matrix where each entry; f is derivate w.r.t respect to X_{ij}.

- grad is same shape as X; each element represents the partial derivate of f with respect to corresponding element of X.

3) Derivates of Liner functions

$\frac{d}{dx} Ax = A$; where A is a constant matrix x is input.

4) Derivative of Quadratic form:

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

↳ No liner shift I only multiplization)

5) Derivate of Transpose

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

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

6) Derivate of Matrix-Vector Product

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

7) Derivate of an inverse Matrix

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

=> back to ii

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

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

$$= \frac{\partial}{\partial \theta}\left[W(X^T X \theta^T\theta - X^T\theta^T y - X\theta y^T + y y^T)\right]$$

$$= \frac{\partial}{\partial \theta}\left[W x^T X \theta^T\theta - W X^T \theta^T y - W X \theta y^T + W y y^T\right]$$

$$= \frac{\partial}{\partial \theta}\left[W x^T X (\theta^T \theta - 2 y^T W X \theta\right]$$

$$= \frac{\partial}{\partial \theta}\left[\theta W x^T X \theta^T - 2 y^T W X \theta\right]$$

$$= 2 X^T W X \theta - 2 y^T W X$$

$$= 2 X^T W X \theta - 2 X^T W y$$

→ quadratic form

$\theta^T X \theta = 2X\theta$

- also symmetric b/c W!

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

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

$$\Rightarrow \ell(\theta) = \sum \log p(y|x;\theta)$$

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

$$= \log\left(\frac{1}{\sqrt{2\pi}\sigma^i}\right) + \log\left[\exp\left(-\frac{y^i \dots}{\dots}\right)\right.$$

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

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

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

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

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

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

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

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

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

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

5b) Weighted linear regression:

- in notebook

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

⇒ weighted linear regression ⇒ no explicit "training phase"; it is "non-parametric" meaning the model is retrained dynamically for each query point at test time. This is because we don't train a global model once and reuse it for all predictions.

⇒ model is retrained dynamically.

⇒ also loss such allows you to get θ directly

$$\nabla J(\theta) = 2X^T W X\theta = 2X^T Wy$$

$$0 = 2X^T W X\theta - 2X^T W y$$

$$2X^T W X\theta = 2X^T W y$$

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

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

$\Rightarrow$

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

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

5c) ... in notebook