ML CS229.

PS 1.1

1. Linear Classifiers (Logistic Regression & GDA)

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

1°: $\nabla_\theta J(\theta)$

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

$g'(\theta^T x^{(i)}) \cdot (\theta^T x^{(i)})' = g(\theta^T x^{(i)})(1-g(\theta^T x^{(i)})) \cdot x_j^{(i)}$

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

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

$$\Rightarrow \nabla_\theta J(\theta) = +\frac{1}{m}\left[\sum_{i=1}^{m} g(\theta^T x^{(i)}) - y^{(i)}\right] x_j^{(i)}$$

$$\Rightarrow \nabla_\theta J(\theta) = \frac{1}{m}\left(g(\theta X) - Y\right)^T X$$

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

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

$z^T X^T \cdot g(X\theta)\cdot(1-g(X\theta)) \cdot X \cdot z$ ($g(z) \in [0,1]$)

$= k(Xz)^T \cdot (Xz)$

$= k z^T X^T X z = k(z^T X)^2 \geq 0$ ✓

— □

or

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

(row of X)

$$= \frac{1}{m}\sum_{i}\sum_{j}\sum_{k} z_j \cdot x_k^{(i)} \cdot x_j^{(i)} \cdot z_k$$

$$= \frac{1}{m}\sum_{i}\left(z^T x^{(i)}\right)^2 \geq 0$$

□

$\therefore H \geq 0$

(c) GDA. Gaussian Discriminant Analysis

• Bayes' Theorem $p(y=1|x) = \frac{p(x|y=1;\mu_1,\mu_0,\Sigma)\, p(y=1;\phi)}{p(x)}$

$\phi, \mu_0, \mu_1, \Sigma$

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

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

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

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

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

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

$$= \frac{1}{\exp(-(\theta^T x + \theta_0)) + 1}.$$

where $\theta^T = \frac{\mu_1-\mu_0}{\Sigma}$, $\theta_0 = \frac{\mu_1^2-\mu_0^2}{2\Sigma} - \ln\frac{1-\phi}{\phi}$. □

(d) maximize $L(\phi, \mu_0, \mu_1, \Sigma)$:

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

where $\mu_{y^{(i)}} = \begin{cases}\mu_0, & y^{(i)}=0 \\ \mu_1, & y^{(i)}=1\end{cases} \Rightarrow \mu_{y^{(i)}} = \{y^{(i)}=0\}\mu_0 + \{y^{(i)}=1\}\mu_1$

② $p(y^{(i)};\phi) = \begin{cases}\phi & y=1 \\ 1-\phi & y=0\end{cases} \Rightarrow p(y^{(i)}) = \phi^{\{y^{(i)}=1\}}\cdot(1-\phi)^{\{y^{(i)}=0\}}$

$\therefore L = \sum_{i=0}^{m} \log p(x^{(i)}|y^{(i)};\cdot) + \sum_{i=0}^{m}\log p(y^{(i)};\phi)$

plug ① ② into L:

$$= -m(\log(2\pi)^{\frac{n}{2}}|\Sigma^{\frac{1}{2}}|) - \frac{1}{2}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})^T\Sigma^{-1}(x^{(i)}-\mu_{y^{(i)}}) + \sum_{i=1}^{m}\{y^{(i)}=1\}\cdot\log\phi + \left(\sum_{i=1}^{m}\{y^{(i)}=0\}\right)\log(1-\phi)$$

$$= -m\left[\frac{n}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma|\right] + \cdots$$

$$\Delta\frac{\partial L}{\partial\phi} = \frac{1}{\phi}\cdot\sum_{i=1}^{m}\{y^{(i)}=1\} + \frac{1}{\phi-1}\left[m - \sum_{i=1}^{m}\{y^{(i)}=1\}\right]$$

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

$$\Delta\frac{\partial L}{\partial\Sigma} = -\frac{m}{2}\cdot\Sigma^{-1} + \frac{1}{2}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})^T\cdot(x^{(i)}-\mu_{y^{(i)}})\cdot\Sigma^{-2}$$

$$\frac{\partial L}{\partial\Sigma} = 0 \Rightarrow \Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-\mu_{y^{(i)}})^T(x^{(i)}-\mu_{y^{(i)}})$$

$$\frac{\partial l}{\partial \mu_0} = \frac{\partial L}{\partial \mu_{y^{(i)}}} \cdot \frac{\partial \mu_{y^{(i)}}}{\partial \mu_0} = \frac{\partial \left\{ -\frac{1}{2} \sum_{i=1}^{m} [x^{(i)} - 2\mu_{y^{(i)}} \cdot x^{(i)} + \mu_{y^{(i)}}^2] \right\} \{y^{(i)}=0\}}{\partial \mu_{y^{(i)}}}$$

$$= \Sigma^{-1}\left(-\frac{1}{2}\right) \sum_{i=1}^{m} \left[-2x^{(i)} + 2\mu_{y^{(i)}} \right] \{y^{(i)}=0\}$$

$$= \Sigma^{-1} \sum_{i=1}^{m} \left(x^{(i)} - \mu_{y^{(i)}} \right) \cdot \{y^{(i)}=0\}$$

μ_0

maximize, Hence

$$\frac{\partial l}{\partial \mu_0} = 0 \Rightarrow \mu_0 = \frac{\sum_{i=1}^{m} \{y^{(i)}=0\} \cdot x^{(i)}}{\sum_{i=1}^{m} \{y^{(i)}=0\}}$$

By symmetry: $\mu_1 = \frac{\sum_{i=1}^{m} \{y^{(i)}=1\} \cdot x^{(i)}}{\sum_{i=1}^{m} \{y^{(i)}=1\}}$ □

---

3. Incomplete, Positive-Only Labels

(a). conditionally independent. By def.

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

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

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

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

Given that $p(t^{(i)}=1 \mid y^{(i)}=1) = 1$

t & x conditionally independent

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

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

(b) V_+ positive (labeled) examples in V

i.e. $V_+ = \{ x^{(i)} \in V \mid y^{(i)}=1 \}$

$\therefore p(t^{(i)}=1 \mid x^{(i)}) \approx 1$ when $x \in V_+$

Given those $\alpha = p(y^{(i)}=1 \mid t^{(i)}=1)$

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

Take α as weighted average.

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

3. Poisson Regression

(a). Poisson Regression.

$$p(y;\lambda) = \frac{\lambda^y \exp(-\lambda)}{y!} = \frac{\exp(-\lambda + y \ln \lambda)}{y!}$$

Exponential Family: P22

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

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

sufficient statistic $T(y) = y$

natural parameter / canonical parameter $\eta = \ln \lambda$

log partition function $a(\eta) = \lambda = e^{\eta}$ · sigmoid function P22

(b) GLM. Generalized Linear Model

P26 $T(y) = y$ $\quad h(x) := E[y \mid x; \theta]$

canonical response fun $h(x) = E[y \mid x] = \lambda = e^{\eta} = e^{\theta^T x}$

η natural parameter:

(c) $\log p(y^{(i)} \mid x^{(i)}; \theta) = \log \frac{1}{y^{(i)}!} \exp[\theta^T x^{(i)} y^{(i)} - e^{\theta^T x^{(i)}}]$

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

$(x + \theta_j \cdot x_j^{(i)} + \dots)$

$$\frac{\partial \log p(\dots)}{\partial \theta_j} = y^{(i)} \cdot x_j^{(i)} - e^{\theta^T x^{(i)}} x_j^{(i)}$$

$$= [y^{(i)} - e^{\theta^T x^{(i)}}] \; x_j^{(i)}$$

derive. ⇓ < batch

Stochastic gradient update ascent rule:

$$\theta_j := \theta_j + \alpha (y^{(i)} - h_\theta(x^{(i)})) x_j^{(i)}$$

$h_\theta(x^{(i)}) = \lambda = e^{\theta^T x^{(i)}}$

LMS

4. Convexity of GLM

(a) $E[Y|X;\theta] = \int y\, p(y;\eta)\, dy$ ①

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

Hint. * using derivatives

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

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

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

$= E[Y|x;\theta] - \int p(y;\eta)\, dy \cdot \frac{\partial a(\eta)}{\partial \eta}$ (By ①)

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

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

Hint.

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

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

$= E[Y^2;\eta] - E^2[Y;\eta]$ (μ^2)

$= Var[Y;\eta] = Var[Y|X;\eta]$.

(c) p13 negative log likelihood. P13

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

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

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

$H_{jk} = \frac{\partial^2 l(\theta)}{\partial \theta_j \partial \theta_k} = \sum_{i=1}^m a''(\theta^T x^{(i)}) \cdot x_j^{(i)} \cdot x_k^{(i)}$

$\cdot$ P51) $z^T H z = \sum_i^m \sum_j^n \sum_k^n a''(\theta^T x^{(i)})\, x_j^{(i)} x_k^{(i)} z_j z_k$

$= \sum_i^m a''(\theta^T x^{(i)}) \left(x^{(i)T} \cdot z \right)^2 \geq 0.$

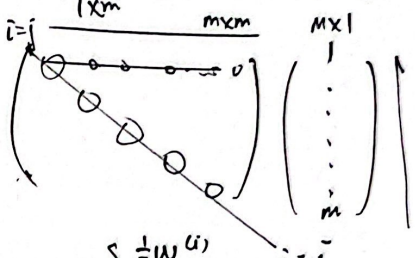
$\Rightarrow z^T H z \geq 0$

---

5. Locally weighted Linear Regression. P110

(a) i. $w \in R^{m \times m}$

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



$\therefore w_{ij} = \begin{cases} \frac{1}{2} w^{(i)} & i = j \\ 0 & i \neq j \end{cases}$

---

ii. normal equation P11

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

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

$= \nabla_\theta (\theta^T X^T W X \theta - y^T W X \theta - \theta^T X^T W y + y^T W y)$ ($\theta^T X^T W y$: 1×m, m×m, m×1)

$= \nabla_\theta (\theta^T X^T W X \theta - 2 y^T W X \theta)$ ($\sim R$)

$= \nabla_\theta \operatorname{tr} (\theta^T X^T W X \theta I - 2 y^T W X \theta)$

$= X^T W X \theta + (X^T W X)^T \theta - 2 (y^T W X)^T$ ($\nabla_\theta \operatorname{tr}(2\theta y^T W X)$)

$= X^T \cdot W^T X$ (As $W^T = W$ by symmetry)

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

$\nabla_\theta J(\theta) = 0 \Rightarrow \theta = (X^T W X)^{-1} X^T W y$

$\operatorname{tr} A = \operatorname{tr} A^T$.

$\nabla_A \operatorname{tr} AB = B^T$.

$\nabla_{A^T} f(A) = (\nabla_A f(A))^T$

$\nabla_A \operatorname{tr} BA = B^T$.

$\nabla_A \operatorname{tr} ABA^T C = BA^T C^T + BA^T C$

---

iii. $l(\theta) = \sum_{i=1}^m \log p(y^{(i)}|x^{(i)};\theta)$

$= \sum_{i=1}^m -\log(\sqrt{2\pi}\sigma^{(i)}) - \frac{(y^{(i)} - \theta^T x^{(i)})^2}{2(\sigma^{(i)})^2}$ ($\uparrow w^{(i)}$)

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

(b) underfitting

(c) bandwidth parameter

$\tau = 0.05$

Python. src.

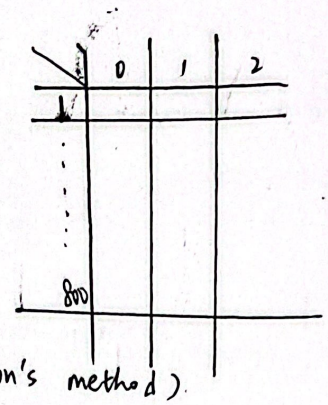
1.(b).

x. shape

m, n = x. shape. (800, 3)

m = x. shape[0]

n = x. shape[1]



✓ Logistic regression (Newton's method).

$$\theta := \theta - H^{-1} \nabla_\theta L(\theta)$$

↑ np.linalg.inv( )

where.

- $g(x) = \frac{1}{1+e^{-\theta^T x}}$ ← np.exp( )
- $H = g(x) \cdot [1 - g(x)] \cdot x^T \cdot x \,/ m$

$$|\theta'_k - \theta_{k-1}| < \varepsilon$$

Prediction

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

---

1. (e) Logistic vs GDA

$p(x|y)$

DB1: Logistic > GDA (Not Gaussian Distribution)

(h). Box-Cox transformation.

2. (e).

decision boundary: $p(t^{(i)} = 1 \mid x^{(i)}) = \frac{1}{2}$.

$$\Rightarrow \frac{1}{2} \; p(y^{(i)} = 1 \mid x^{(i)}) = \frac{1}{2}$$

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

$$\theta^T x + \underbrace{\ln\left(\frac{\beta_2}{\beta_1} - 1\right)}_{\text{add to } \theta_0} = 0.$$

3. (d).

$$+= \alpha \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) \, x_j^{(i)} \,/ m \quad \text{LMS}$$

step size

x: $x_1 \cdots x_4$ (m, n) ~~(4, 3000)~~

x_i (3000, 1)

x^T (4, 3000)

x^T: $x_1^{(i)} \cdots x_4^{(i)}$

y: $y^{(i)}$ (3000, 1)

if np.linalg.norm(self.theta − theta, ord=1) < self.eps:

breaks

5. (b) $w^{(i)} = \exp\left(-\frac{\|x^{(i)} - x\|_2^2}{2\tau^2}\right)$

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

$y' = \theta^T x$.