

CS229: Problem Set #1

Andrew Ng

0130

Problem 1

1.

$$\begin{aligned}\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} - \frac{1}{m} \sum_{i=1}^m y^{(i)} \log(h_\theta(x^{(i)})) + (1 - y^{(i)} \log(1 - h_\theta(x^{(i)}))) \\ &= -\frac{1}{m} \sum_{i=1}^m [y^{(i)} \frac{1}{h_\theta(x^{(i)})} - (1 - y^{(i)}) \frac{1}{1 - h_\theta(x^{(i)})}] \frac{\partial}{\partial \theta_j} h_\theta(x^{(i)})\end{aligned}$$

Then we calculate:

$$\frac{\partial}{\partial \theta_j} h_\theta(x^{(i)}) = \frac{\partial}{\partial \theta_j} g(\theta^T x^{(i)}) = g(\theta^T x^{(i)}) (1 - g(\theta^T x^{(i)})) x_j^{(i)} = h_\theta(x^{(i)}) (1 - h_\theta(x^{(i)})) x_j^{(i)}$$

And we can further simplify the above equation:

$$\begin{aligned}\frac{\partial}{\partial \theta_j} J(\theta) &= -\frac{1}{m} \sum_{i=1}^m [(y^{(i)} - h_\theta(x^{(i)})) - (1 - y^{(i)}) h_\theta(x^{(i)})] x_j^{(i)} \\ &= -\frac{1}{m} \sum_{i=1}^m (y^{(i)} - h_\theta(x^{(i)})) x_j^{(i)}\end{aligned}$$

Then we can get second-order derivative:

$$\begin{aligned}H_{ij} &= \frac{\partial^2}{\partial \theta_i \partial \theta_j} J(\theta) = \frac{1}{m} \sum_{k=1}^m x_j \frac{\partial}{\partial \theta_i} h_\theta(x^{(k)}) \\ &= \frac{1}{m} \sum_{k=1}^m x_i^{(k)} x_j^{(k)} h_\theta(x^{(k)}) (1 - h_\theta(x^{(k)}))\end{aligned}$$

for each vector z , consider the quadratic form of Hessian matrix:

$$\begin{aligned}z^T H z &= \sum_{i=1}^n \sum_{j=1}^n z_i H_{ij} z_j = \frac{1}{m} \sum_{i=1}^n \sum_{j=1}^n \sum_{k=1}^m z_i x_i^{(k)} z_j x_j^{(k)} h_\theta(x^{(k)}) (1 - h_\theta(x^{(k)})) \\ &= \frac{1}{m} \sum_{k=1}^m (\sum_{i=1}^n \sum_{j=1}^n z_i x_i^{(k)} z_j x_j^{(k)}) h_\theta(x^{(k)}) (1 - h_\theta(x^{(k)})) \\ &= \frac{1}{m} \sum_{k=1}^m (x^{(k)T} z)^2 h_\theta(x^{(k)}) (1 - h_\theta(x^{(k)})) \geq 0 \Leftrightarrow H \succeq 0\end{aligned}$$

2. Codes are shown in src director, see `src/p01b_logreg.py`.

3.

$$\begin{aligned}P(y = 1|x; \phi, \mu_0, \mu_1, \Sigma) &= \frac{P(x|y; \phi, \mu_0, \mu_1, \Sigma)}{P(x|y = 1; \phi, \mu_0, \mu_1, \Sigma)P(y = 1) + P(x|y = 0; \phi, \mu_0, \mu_1, \Sigma)P(y = 0)} \\ &= \frac{\exp(-1/2(x - \mu_1)^T \Sigma^{-1}(x - \mu_1))}{\exp(-1/2(x - \mu_1)^T \Sigma^{-1}(x - \mu_1))\phi + \exp(-1/2(x - \mu_0)^T \Sigma^{-1}(x - \mu_0))(1 - \phi)} \\ &= \frac{1}{1 + ((1 - \phi)/\phi) \exp(-1/2(x - \mu_0)^T \Sigma^{-1}(x - \mu_0) + 1/2(x - \mu_1)^T \Sigma^{-1}(x - \mu_1))}\end{aligned}$$

then we calculate the portion of the denominator inside the exponential term using the fact that: (1) Σ is symmetric (2) if A is symmetric, then $(A^{-1})^T = (A^T)^{-1}$.

$$\begin{aligned}1/2((x - \mu_1)^T \Sigma^{-1}(x - \mu_1) - (x - \mu_0)^T \Sigma^{-1}(x - \mu_0)) \\ = (\mu_0 - \mu_1)^T \Sigma^{-1}x + 1/2(\mu_0^T \Sigma^{-1}\mu_0 - \mu_1^T \Sigma^{-1}\mu_1)\end{aligned}$$

and this allow us to simplify the first equation and prove that the decision boundary is linear:

$$\begin{aligned} P(y = 1|x; \phi, \mu_0, \mu_1, \Sigma) &= \frac{1}{1 + \exp(-(\theta^T x + \theta_0))} \\ \theta &= \Sigma^{-1}(\mu_1 - \mu_0) \\ \theta_0 &= 1/2(\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1) + \log(1 - \phi) - \log \phi \end{aligned}$$

4. Firstly, simplify the formula of l :

$$\begin{aligned} l(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^m p(x^{(i)}, y^{(i)}, \phi, \mu_0, \mu_1, \Sigma) \\ &= -\frac{mn}{2} \log(2\pi) - \frac{m}{2} \log(|\Sigma|) - \frac{1}{2} \sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) + \sum_{i=1}^m y^{(i)} \log(\phi) + (1 - y^{(i)}) \log(1 - \phi) \end{aligned}$$

let $\frac{\partial}{\partial \phi} l = 0$, we have

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

let $\frac{\partial}{\partial \mu_{y^{(i)}}} l = 0$, we have

$$\begin{cases} \frac{\partial}{\partial \mu_{y^{(i)}}} l &= \frac{1}{2} \sum_{i=1}^m 2\Sigma^{-1}(x^{(i)} - \mu_{y^{(i)}}) = \Sigma^{-1} \sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}}) = 0 \\ \frac{\partial}{\partial \mu_0} \mu_{y^{(i)}} &= 1\{y^{(i)} = 0\}, \frac{\partial}{\partial \mu_1} \mu_{y^{(i)}} = 1\{y^{(i)} = 1\} \end{cases}$$

using the chain rule, we get:

$$\frac{\partial}{\partial \mu_k} l = \frac{\partial}{\partial \mu_{y^{(i)}}} l \frac{\partial}{\partial \mu_k} \mu_{y^{(i)}} = 0 \Rightarrow \mu_k = \frac{\sum_{i=1}^m 1\{y^{(i)} = k\} x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)} = k\}}, k = 0, 1$$

we don't need to assume $n = 1$, let's consider a more general case, let $\frac{\partial}{\partial \Sigma} l = 0$:

$$\frac{\partial}{\partial \Sigma} l = \frac{\partial}{\partial \Sigma} \left\{ -\frac{m}{2} \log(|\Sigma|) - \frac{1}{2} \sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})^T \Sigma^{-1} (x^{(i)} - \mu_{y^{(i)}}) \right\} = 0$$

let's first calculate the derivative of the determinant, suppose $D = (\delta_1, \delta_2, \dots, \delta_n)$ and $E = (\epsilon_1, \epsilon_2, \dots, \epsilon_n)$, we have:

$$\det(D + tE) - \det(D) = \sum_{i=1}^n \det(\delta_1, \delta_2, \dots, t\epsilon_i, \dots, \delta_n)$$

let $\epsilon_i = \sum_j f_{ij} \delta_j$ which is equivalent to $E = DF$ where $F_{ij} = f_{ij}$, then the above equation can be written as:

$$\sum_{i=1}^n \det(\delta_1, \delta_2, \dots, t\epsilon_i, \dots, \delta_n) = \sum_{i=1}^n t f_{ii} \det(D) = \text{tr}(D^{-1}E) \cdot t \det(D)$$

then we know the derivative of the determinant is:

$$\frac{\partial}{\partial \Sigma} \det(\Sigma) = \det(\Sigma) \text{tr}(\Sigma^{-1}E)$$

then we can calculate the derivative of the log determinant:

$$\frac{\partial}{\partial \Sigma} \log(\det(\Sigma)) = \frac{1}{\det(\Sigma)} \det(\Sigma) \text{tr}(\Sigma^{-1}E) = \text{tr}(\Sigma^{-1}E)$$

next, we calculate the derivative of the inverse of the matrix:

$$\begin{aligned}(D + tE)^{-1} - D^{-1} &= (D(I + tD^{-1}E))^{-1} - D^{-1} \\ &= (I - tD^{-1}E + O(t^2))D^{-1} - D^{-1} \\ &= -tD^{-1}ED^{-1} + O(t^2)\end{aligned}$$

then we know the derivative of the inverse of the matrix is:

$$\frac{\partial}{\partial \Sigma} \Sigma^{-1} = -\Sigma^{-1} E \Sigma^{-1}$$

using the fact that $\frac{\partial}{\partial A} AB = B^T$, let $u^{(i)} = (x^{(i)} - \mu_{y^{(i)}})$, we know the equation of $\frac{\partial}{\partial \Sigma} l = 0$ is equivalent to:

$$\begin{aligned}-\frac{m}{2} \text{tr}(\Sigma^{-1} E) + \frac{1}{2} \sum_{i=1}^m u^{(i)T} \Sigma^{-1} E \Sigma^{-1} u^{(i)} &= 0 \\ \Leftrightarrow \text{tr}(\sum_{i=1}^m u^{(i)T} \Sigma^{-1} E \Sigma^{-1} u^{(i)} - m \Sigma^{-1} E) &= 0 \\ \Leftrightarrow \text{tr}((\sum_{i=1}^m \Sigma^{-1} u^{(i)} u^{(i)T} - m I) \Sigma^{-1} E) &= 0, \forall E \\ \Leftrightarrow \sum_{i=1}^m \Sigma^{-1} u^{(i)} u^{(i)T} &= m I \\ \Leftrightarrow \Sigma &= \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T\end{aligned}$$

5. Codes are shown in src director, see `src/p01e_gda.py`.
6. Figure is shown in `src/output/p01f_plot.png`. Logistic regression is better than GDA in this case, since the $p(x|y)$ may not be Gaussian.
7. Figure is shown in `src/output/p01g_plot.png`.
8. The idea of this problem is that we can choose a transformation of x such that $p(x|y)$ is Gaussian. (See Box-Cox transformation)

Problem 2

- Using the chain of probability, $p(y^{(i)} = 1|x^{(i)})$ can be written as:

$$p(y^{(i)} = 1|x^{(i)}) = \sum_{k=0}^1 p(y^{(i)} = 1|t^{(i)} = k, x^{(i)})p(t^{(i)} = k|x^{(i)})$$

and since we know $p(y^{(i)} = 1|t^{(i)} = 1, x^{(i)}) = p(y^{(i)} = 1|t^{(i)} = 1)$ and $p(y^{(i)} = 1|t^{(i)=0, x^{(i)}}) = 0$, we can further simplify the above equation:

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

then we know $\alpha = p(y^{(i)} = 1|t^{(i)} = 1)$.

-

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

- Codes are shown in src directory, see `src/p02cde_posonly.py`, result is stored in `src/output/p02c_pred.txt`.
- Codes are shown in src directory, see `src/p02d_roc.py`, result is stored in `src/output/p02d_pred.txt`.
- Codes are shown in src directory, see `src/p02e_roc.py`, result is stored in `src/output/p02e_pred.txt`. We need correction here since we train on dataset with label `y`, but we need to predict on dataset with label `t`. Let θ' represents the θ after correction, we need to find a relationship between θ and θ' using the fact that $p(t^{(i)} = 1|x^{(i)}) = \alpha p(y^{(i)} = 1|x^{(i)})$. Here I just use the correction in official solution.

Problem 3

1.

$$\begin{aligned}
 p(y; \lambda) &= \frac{e^{-\lambda} \lambda^y}{y!} = \exp(\log(\frac{e^{-\lambda} \lambda^y}{y!})) \\
 &= \exp(-\lambda + y \log \lambda - \log y!) = \frac{1}{y!} (\log \lambda y - \lambda) \\
 \Rightarrow b(y) &= \frac{1}{y!}, \quad \eta = \log \lambda, \quad T(y) = y, \quad \alpha(\eta) = \lambda = e^\eta
 \end{aligned}$$

2. Let canonical response function represented as g , then we have:

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

then we know $g = \exp$.

3.

$$\begin{aligned}
 \log p(y|\lambda) &= -\lambda + y \log \lambda - \log y! \Rightarrow \log p(y^{(i)}|x^{(i)}; \theta) \\
 &= -e^{\theta^T x} + y^{(i)} \theta^T x^{(i)} - \log y^{(i)}!
 \end{aligned}$$

and we can calculate the derivative of $\log p(y^{(i)}|x^{(i)}; \theta)$ with respect to θ_j :

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

we then get the gradient ascent update rules as follows:

$$\theta_j := \theta + \alpha \frac{\partial}{\partial \theta_j} \log p(y^{(i)}|x^{(i)}; \theta) = \theta_j + \alpha (y^{(i)} - e^{\theta^T x}) x_j^{(i)}$$

In fact, the member in GLM has similar stochastic gradient ascent update rules:

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

since $\mathbb{E}(y|x; \theta) = \exp(\theta^T x)$, we get the same answer.

4. Codes are shown in src directory, see `src/p03d_poisson.py`.

Problem 4

1. From the property of the probability space we get:

$$\int_{\Omega} p(y; \eta) dy = 1 = \frac{1}{\exp(\alpha(\eta))} \int_{\Omega} b(y) \exp(\eta y) dy$$

which is equivalent to:

$$\exp(\alpha(\eta)) = \int_{\Omega} b(y) \exp(\eta y) dy$$

apply the partial derivatives to both sides, we have:

$$\frac{\partial}{\partial \eta} \exp(\alpha(\eta)) = \exp(\alpha(\eta)) \frac{\partial}{\partial \eta} \alpha(\eta) = \frac{\partial}{\partial \eta} \int_{\Omega} b(y) \exp(\eta y) dy = \int_{\Omega} b(y) \exp(\eta y) y dy$$

after some algebraic manipulation, we get:

$$\frac{\partial}{\partial \eta} \alpha(\eta) = \int_{\Omega} b(y) y \exp(\eta y - \alpha(\eta)) dy = \mathbb{E}[Y|X; \theta]$$

- 2.

$$\begin{aligned} \frac{\partial^2}{\partial \eta^2} \alpha(\eta) &= \frac{\partial}{\partial \eta} \left(\frac{\partial}{\partial \eta} \alpha(\eta) \right) = \frac{1}{\exp(\alpha(\eta))} \int_{\Omega} b(y) y^2 \exp(\eta y) dy - \frac{1}{\exp(\alpha(\eta))} \frac{\partial}{\partial \eta} \alpha(\eta) \int_{\Omega} b(y) y \exp(\eta y) dy \\ &= \mathbb{E}[Y^2|X; \theta] - \mathbb{E}[Y|X; \theta]^2 = \text{Var}[Y|X; \theta] \end{aligned}$$

3. We can formulate the loss function as follows:

$$l(\theta) = -\log J(\theta) = -\log P(Y|X; \theta)$$

where $J(\theta)$ is the likelihood function. In order to get the hessian of the loss function, we first calculate the first-order derivative of the loss function:

$$\frac{\partial}{\partial \theta_j} l(\theta) = \frac{-1}{p(y; \eta)} \frac{\partial}{\partial \eta} p(y; \eta) \frac{\partial}{\partial \theta_j} \eta = \frac{-x_j}{p(y; \eta)} \frac{\partial}{\partial \eta} p(y; \eta)$$

then we calculate the second-order derivative of the loss function:

$$\begin{aligned} \frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\theta) &= \frac{x_j}{p(y; \eta)^2} \frac{\partial}{\partial \eta} p(y; \eta) \frac{\partial}{\partial \theta_i} \eta \frac{\partial}{\partial \eta} p(y; \eta) - \frac{x_j}{p(y; \eta)} \frac{\partial^2}{\partial \eta^2} p(y; \eta) \frac{\partial}{\partial \theta_i} \eta \\ &= \frac{x_i x_j}{p(y; \eta)^2} \left(\frac{\partial}{\partial \eta} p(y; \eta) \right)^2 - \frac{x_i x_j}{p(y; \eta)} \frac{\partial^2}{\partial \eta^2} p(y; \eta) \end{aligned}$$

by calculating the first-order and second-order derivatives of $p(y; \eta)$, we have:

$$\begin{aligned} \frac{\partial}{\partial \eta} p(y; \eta) &= b(y) \exp(\eta y - \alpha(\eta)) (y - \frac{\partial}{\partial \eta} \alpha(\eta)) = p(y; \eta) (y - \frac{\partial}{\partial \eta} \alpha(\eta)) \\ \frac{\partial^2}{\partial \eta^2} p(y; \eta) &= p(y; \eta) (y - \frac{\partial}{\partial \eta} \alpha(\eta))^2 - p(y; \eta) \frac{\partial^2}{\partial \eta^2} \alpha(\eta) \end{aligned}$$

then we can further simplify the second-order derivative of the loss function by some algebraic manipulation, and the result is:

$$\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\theta) = x_i x_j \text{Var}[Y|X; \theta]$$

consider the quadratic form of the hessian matrix, we have:

$$z^T H z = \text{Var}[Y|X; \theta] \sum_i \sum_j x_i x_j z_i z_j = \text{Var}[Y|X; \theta] (z^T x)^2 \geq 0, \forall z \in \mathbb{R}^n$$

Problem 5

1. (a) Let

$$W = 1/2 \begin{bmatrix} w^{(1)} & 0 & \dots & 0 \\ 0 & w^{(2)} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w^{(n)} \end{bmatrix}$$

then we have: $J(\theta) = (X\theta - y)^T W (X\theta - y)$.

- (b) Let $\frac{\partial}{\partial \theta} J(\theta) = 0$, which is equivalent to:

$$\frac{\partial}{\partial \theta} (X\theta - y)^T W (X\theta - y) = \frac{\partial}{\partial A} A^T W A \circ \frac{\partial}{\partial \theta} (X\theta - y) \quad (\text{where } A = X\theta - y)$$

since we know:

$$\begin{aligned} X(\theta + E) - y - (X\theta - y) &= XE \\ (A + E)^T W (A + E) - A^T W A &= A^T W E + E^T W A + E^T W E \end{aligned}$$

we can calculate that the differential of $J(\theta)$ is (Given input E):

$$A^T W X E + E^T X^T W A$$

let it be zero, we have:

$$\begin{aligned} A^T W X E + E^T X^T W A &= 0, \forall E \\ \Leftrightarrow X^T W^T A &= X^T W^T (X\theta - y) = 0 \\ \Leftrightarrow X^T W^T X \theta &= X^T W^T y \\ \Leftrightarrow \theta &= (X^T W^T X)^{-1} X^T W^T y \end{aligned}$$

- (c) Let

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

define $l(\theta) = \log J(\theta)$, maximize $l(\theta)$ is equivalent to maximize $J(\theta)$, $l(\theta)$ can be written as:

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

and we find that maximize $l(\theta)$ is equivalent to:

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

which is equivalent to solve the locally weighted linear regression with $w_i = -\frac{1}{(\sigma^{(i)})^2}$.

- Codes are shown in src directory, see `src/p05b_lwr.py`.
- Codes are shown in src directory, see `src/p05c_tau.py`, figure is shown in `src/output/p05c_lwr_tau0.05_test.png`. $\tau = 0.05$ achieves the lowest MSE on the `valid` split, and the MSE on the `test` split is 0.012400076150475756 with $\tau = 0.05$.