HW5

Name: Guanghao Chen

PID: A53276390

Email: guc001@eng.ucsd.edu

5.1 Gradient based learning

(a)

According to the shorthand,

$$p_t = P(Y=1|\overrightarrow{x_t})$$

Therefore the probability for Y=0 could be

$$P(Y=1|\overrightarrow{x_t}) = 1 - p_t$$

Therefore, given the data for x, the conditional probability for variable Y can also be denoted by

$$P(Y=1|\overrightarrow{x_t}) = p_t^{y_t} + (1-p_t)^{1-y_t}$$

Therefore, the log-likelihood can be denoted by

$$\begin{aligned} L &= \sum_t log(P(y_t|\overrightarrow{x_t})) \\ &= \sum_t log(p_t^{y_t} + (1-p_t)^{1-y_t}) \\ &= \sum_t [y_t log(p_t) + (1-y_t) log(1-p_t)] \end{aligned}$$

Further, the gradient of the log-likelihood with respect to a w_i can be denoted by

$$\begin{aligned} \frac{\partial L}{\partial w_i} &= \sum_t [\frac{y_t}{p_t} f^{'}(\vec{w} \cdot \overrightarrow{x_t}) \overrightarrow{x_{it}} - (1-y_t) \frac{f^{'}(\vec{w}, \overrightarrow{x_t})}{1-p_t} x_{it}] \\ &= \sum_t \frac{f^{'}(\vec{w} \cdot \overrightarrow{x_t}) x_{it}}{p_t(1-p_t)} (y_t - p_t) \end{aligned}$$

(b)

If function f is the sigmoid function, accroding to the previous assignment, the derivative of function f can be denoted by

$$f^{'}(z) = f(z) \times (1 - f(z))$$

In our situation, the first rerivative will be

$f^{'}(\vec{w} \cdot \overrightarrow{x_t}) = f(\vec{w} \cdot \overrightarrow{x_t}) \times (1 - f(\vec{w} \cdot \overrightarrow{x_t})) = p_t(1 - p_t)$

Therefore, the gradient of log-lilelihood is

$$\frac{\partial L}{\partial w_i} = \sum_{t=1}^{T}(y_t)x_{it}$$

5.2 Multinomial logistic regression

Similar to the last problem, we can re-denote p_{it} first.

In this situation, the labels for y is not binary but $1, 2, .., c$

Therefore, the marginalization should be

$$p_{1t} + p_{2t} + ... + p_{ct} = 1$$

Therefore, for a given k, the conditional probability should be

$$\begin{cases} p_{kt} & y_t = 1 \\ 1 - \sum_{m} p_{mt} \ (m \neq k) & y_t = 0 \end{cases}$$

Therefore, the probability p_{it} can be denoted by

$$p_{it} = \sum_{k=1}^{c} [p_{kt}^{y_{kt}} + (1 - p_{kt})^{1 - y_{kt}}]$$

Further, the gradient of log-likelihood is

$$\begin{aligned} L &= \sum_{t} log(P(y_t = i|\overrightarrow{x_t})) \\ &= \sum_{t}\sum_{k} [y_{kt} log(p_{kt}) + (1 - y_{kt}) log(1 - p_{kt})] \end{aligned}$$

Therefore,

$$\begin{aligned} \frac{\partial L}{\partial w_i} &= \sum_{t}\sum_{i} \frac{y_{it}}{p_{it}}\overrightarrow{x_t} - \frac{(1 - y_{it})}{1 - p_{it}}\overrightarrow{x_t} \\ &= \sum_{t}\sum_{i} \frac{y_{it} - p_{it}}{p_{it}(1 - p_{it})} \frac{\partial p_{it}}{\partial \overrightarrow{w_t}} \end{aligned}$$

Therefore,

$$p_{it}(1 - p_{it}) = \frac{e^{\overrightarrow{w_j} \cdot \overrightarrow{x_t}}}{\sum_{j=1}^{c} e^{\overrightarrow{w_j} \cdot \overrightarrow{x_t}}} \times \frac{\sum_{j=1}^{c} e^{\overrightarrow{w_j} \cdot \overrightarrow{x_t}} - e^{\overrightarrow{w_j} \cdot \overrightarrow{x_t}}}{\sum_{j=1}^{c} e^{\overrightarrow{w_j} \cdot \overrightarrow{x_t}}} = \frac{e^{\overrightarrow{w_j}}(\sum_{j=1}^{c} e^{\overrightarrow{w_j} \cdot \overrightarrow{x_t}} - e^{\overrightarrow{w_j} \cdot \overrightarrow{x_t}})}{(\sum_{j=1}^{c} e^{\overrightarrow{w_j} \cdot \overrightarrow{x_t}})^2}$$

And

$$\frac{\partial p_{it}}{\partial \overrightarrow{w_t}} = \frac{e^{\overrightarrow{w_j}\cdot\overrightarrow{x_t}}\overrightarrow{x_t}(\sum_{j=1}^{c} e^{\overrightarrow{w_j}\cdot\overrightarrow{x_t}}) - (e^{\overrightarrow{w_j}\cdot\overrightarrow{x_t}})^2\overrightarrow{x_t}}{(\sum_{j=1}^{c} e^{\overrightarrow{w_j}\cdot\overrightarrow{x_t}})^2} = \frac{e^{\overrightarrow{w_j}\cdot\overrightarrow{x_t}}(\sum_{j=1}^{c} e^{\overrightarrow{w_j}\cdot\overrightarrow{x_t}} - e^{\overrightarrow{w_j}\cdot\overrightarrow{x_t}})}{\sum_{j=1}^{c} e^{\overrightarrow{w_j}\cdot\overrightarrow{x_t}}}\overrightarrow{x_t}$$

Therefore,

$$\frac{\partial L}{\partial w_i} = \sum_t (y_{it} - p_{it})\overrightarrow{x_t}$$

5.3

(a)

For the function $f(x)$ in this problem, the first derivative is

$$f^{'}(x) = \alpha(x - x_*)$$

Therefore, the n^{th} iteration can be denoted by

$$x_n = x_{n-1} - \eta\alpha(x_{n-1} - x_*)$$

So, the error at n^{th} could be

$$\epsilon_n = \epsilon_{n-1} - \eta\alpha\epsilon_{n-1} = \epsilon_{n-1}(1 - \eta\alpha)$$

Finally, the error at n^{th} iteration can be denoted in terms of the initial error ϵ_0

$$\epsilon_n = \epsilon_0(1 - \eta\alpha)^{n-1}$$

(b)

First of all, we need to determine the expression of $f^{''}(x_n)$.

$$f^{''}(x_n) = \alpha$$

In order to make the update rule converging to x_*, we should enable the value of $|1 - \eta\alpha| < 1$ so that the error can be less and less.

Therefore, η should locate in the following range

$$0 < \eta < \frac{2}{\alpha}$$

When $|1 - \eta\alpha|$ becomes more closer to zero, the convergence will be faster. The smallest value of $|1 - \eta\alpha|$ is 0. Therefore, at this moment,

$$\eta = \frac{1}{\alpha}$$

Using the expression of $f^{''}(x_n)$, the equation can also be denoted by following

$$\eta = \frac{1}{f''(x_n)}$$

(c)

Still focusing on the second the quadratic function $f(x) = \frac{\alpha}{2}(x - x_*)^2$ and $f'(x) = \alpha(x - x_*)$, the updating rule can be denoted by the following

$$\begin{aligned} x_{n+1} &= x_n - \eta\alpha(x_n - x_*) + \beta(x_n - x_{n-1}) \\ x_{n+1} - x_* &= x_n - x_* - \eta\alpha(x_n - x_*) + \beta(x_n - x_{n-1}) \\ \epsilon_{n+1} &= \epsilon_n - \eta\alpha\epsilon_n + \beta((x_n - x_*) - (x_{n-1} - x_*)) \\ \epsilon_{n+1} &= \epsilon_n - \eta\alpha\epsilon_n + \beta(\epsilon_n - \epsilon_{n-1}) \\ \epsilon_{n+1} &= (1 - \eta\alpha + \beta)\epsilon_n - \beta\epsilon_{n-1} \end{aligned}$$

(d)

Given the parameters

$$\begin{aligned} \epsilon_{n+1} &= (1 - \frac{4}{9} + \frac{1}{9})\epsilon_n - \frac{1}{9}\epsilon_{n-1} \\ \epsilon_{n+1} &= (\frac{2}{3})\epsilon_n - \frac{1}{9}\epsilon_{n-1} \end{aligned}$$

Therefore, we can recursively write down the equation for each iteration

Assume that one solution to the recursion is $\epsilon_n = \lambda^n\epsilon_0$, then substitute it in order to determine parameter λ.

$$\begin{aligned} \lambda^{n+1}\epsilon_0 &= \frac{2}{3}\lambda^n\epsilon_0 - \frac{1}{9}\lambda^{n-1}\epsilon_0 \\ 9\lambda^2 - 6\lambda + 1 &= 0 \\ (3\lambda - 1)^2 &= 0 \\ \lambda &= \frac{1}{3} \end{aligned}$$

Therefore,

$$\epsilon_n = (\frac{1}{3})^n\epsilon_0$$

5.4 Newton's method

(a)

First of all, we need to derive the expression for first derivative and second derivative for the given polynomial function.

$$f^{'}(x_n) = 2p(x_n - x_*)^{2p-1}$$
$$f^{''}(x_n) = 2p(2p-1)(x_n - x_*)^{2p-2}$$

Further, computing the error for iteration n^{th}

$$x_{n+1} = x_n - \frac{x_n - x_*}{2p-1}$$
$$x_{n+1} - x_* = x_n - x_* - \frac{x_n - x_*}{2p-1}$$
$$\epsilon_{n+1} = \epsilon_n \frac{2p-2}{2p-1}$$

Recursively, it has

$$\epsilon_n = (\frac{2p-2}{2p-1})^n \epsilon_0$$

(b)

According to the requirement $\epsilon_n \leq \delta\epsilon_0$, it can be substitute with ϵ_n

$$(\frac{2p-2}{2p-1})^n \epsilon_0 \leq \delta\epsilon_0$$
$$nlog(\frac{2p-2}{2p-1}) \leq log(\delta)$$

According to the hint $logz \leq z - 1$, the denominator can be denoted by

$$n(\frac{2p-2}{2p-1} - 1) \leq log(\delta)$$
$$n(\frac{-1}{2p-1}) \leq log(\delta)$$
$$n \geq -(2p-1)log(\delta)$$
$$n \geq (2p-1)log(\frac{1}{\delta})$$

(c)

In order to compute the minimum of the function, we should make the first derivative equal zero.

$$f^{'}(x) = -x_* \frac{x}{x*} \frac{x_*}{x^2} + 1 = -\frac{x_*}{x} + 1 = 0$$

Therefore, when $x = x_*$, the function occurs minimum.

(d)

$$f^{''}(x) = \frac{x_*}{x^2}$$

According to Newton's method, the updating rule is

$$\begin{aligned} x_{n+1} &= x_n - \frac{f^{'}(x_n)}{f^{''}(x_n)} \\ &= x_n - \frac{x_n(x_n - x_*)}{x_*} \\ &= \frac{2x_n x_* - x_n^2}{x_*} \end{aligned}$$

Therefore, the relative error at n^{th} is

$$\rho_n = \frac{2n - 1x_* - x_n^2 - x_*^2}{x_*^2} = \frac{-(x_* - x_{n-1})^2}{x_*^2} = -\rho_{n-1}^2$$

If the method want's to be converged, $|\rho_0|$ should less than 1, which means

$$\begin{gathered} |\rho_0| < 1 \\ -1 < \rho_0 < 1 \\ -1 < \frac{x_0 - x_*}{x_*} < 1 \\ 0 < x_0 < 2x_* \end{gathered}$$

5.5

(a)

```
[[-0.14978948 -0.23379796 -0.2620632  -0.26735281 -0.15650103  0.05819675
   0.26559247  0.44076956]
 [-0.05206076  0.01651187  0.06728831  0.11427564  0.18547774  0.15784308
  -0.16366097 -0.21322931]
 [ 0.23635434  0.41390298  0.52715968  0.34259459 -0.02491731 -0.50741746
  -0.80810187 -0.53627667]
 [ 0.2355871   0.34742695  0.32223668  0.00385604 -0.2056208  -0.24906787
  -0.22016208 -0.16484149]
 [ 0.10586976  0.08902139  0.03682623 -0.07382843 -0.09211044 -0.1196024
  -0.11450961 -0.24424491]
 [ 0.21715615  0.00884969  0.03549628  0.12870283  0.08384332  0.0158127
  -0.07514835 -0.30496598]
 [ 0.15881496  0.11698806  0.1166019   0.07327297 -0.00699652 -0.06135614
  -0.00213004 -0.23139402]
 [-0.08032515  0.0984083   0.11727974  0.07875866  0.01869178  0.03971351
```

```
 -0.08999777 -0.06336839]]
```

The plots are shown in the attachment.

(b)

```
The error rate of the model is 7.124999999999999%
```

(c)

See in the attachment.