CSE250A Homework4 Answer

Yue Wang, A53102167

November 1, 2015

4.1 Gradient-based learning

(a)

We know $\log P(y_t|\vec{x}_t) = \log {p_t}^{y_t}(1-p_t)^{1-y_t}$ and $p_t = f(\vec{w}\cdot\vec{x})$.

So: $\mathcal{L} = \sum_t \log P(y_t|\vec{x}_t) = \sum_t \log {p_t}^{y_t}(1-p_t)^{1-y_t} = \sum_t [y_t \log p_t + (1-y_t)\log(1-p_t)]$

$\frac{\partial \mathcal{L}}{\partial w_i} = \sum_{t=1}^{T}[\frac{y_t f'(\vec{w}\cdot\vec{x_t})x_{it}}{p_t} - \frac{(1-y_t)f'(\vec{w}\cdot\vec{x_t})x_{it}}{(1-p_t)}] = \sum_{t=1}^{T}[\frac{y_t f'(\vec{w}\cdot\vec{x_t})x_{it}(1-p_t)}{p_t(1-p_t)} - \frac{(1-y_t)f'(\vec{w}\cdot\vec{x_t})x_{it}p_t}{p_t(1-p_t)}]$

$= \sum_{t=1}^{T}[\frac{f'(\vec{w}\cdot\vec{x_t})}{p_t(1-p_t)}](y_t - p_t)x_{it}$

(b)

We know $\sigma'(z) = \sigma(z)\sigma(-z) = \sigma(z)(1-\sigma(z))$.

So:

$\frac{\partial \mathcal{L}}{\partial w_i} = \sum_{t=1}^{T}[\frac{f'(\vec{w}\cdot\vec{x_t})}{p_t(1-p_t)}](y_t - p_t)x_{it}$

$= \sum_{t=1}^{T}[\frac{\sigma'(\vec{w}\cdot\vec{x_t})}{\sigma(\vec{w}\cdot\vec{x_t})(1-\sigma(\vec{w}\cdot\vec{x_t}))}](y_t - \sigma(\vec{w}\cdot\vec{x_t}))x_{it}$

$= \sum_{t=1}^{T}[\frac{\sigma(\vec{w}\cdot\vec{x_t})(1-\sigma(\vec{w}\cdot\vec{x_t})}{\sigma(\vec{w}\cdot\vec{x_t})(1-\sigma(\vec{w}\cdot\vec{x_t}))}](y_t - \sigma(\vec{w}\cdot\vec{x_t}))x_{it}$

$= \sum_{t=1}^{T}(y_t - \sigma(\vec{w}\cdot\vec{x_t}))x_{it}$

which is the result in lecture.

4.2 Multinomial logistic regression

$\frac{\partial p_{it}}{\partial \vec{w}_i} = \frac{e^{\vec{w}_i \cdot \vec{x}_t} \cdot \vec{x}_t}{\sum_{j=1}^{c} e^{\vec{w}_j \cdot \vec{x}_t}} - \frac{e^{\vec{w}_i \cdot \vec{x}_t} \cdot e^{\vec{w}_i \cdot \vec{x}_t} \cdot \vec{x}_t}{(\sum_{j=1}^{c} e^{\vec{w}_j \cdot \vec{x}_t})^2} = p_{it}\vec{x}_t - {p_{it}}^2\vec{x}_t$

$\frac{\partial p_{kt}}{\partial \vec{w}_i} = -\frac{e^{\vec{w}_k \cdot \vec{x}_t} \cdot e^{\vec{w}_i \cdot \vec{x}_t} \cdot \vec{x}_t}{(\sum_{j=1}^{c} e^{\vec{w}_j \cdot \vec{x}_t})^2} = -p_{kt}p_{it}\vec{x}_t$

$P(y_t|\vec{x}_t) = \prod_k p_{kt}^{y_{kt}}$

$\mathcal{L} = \sum_t \log P(y_t|\vec{x}_t) = \sum_t \log \prod_k p_{kt}^{y_{kt}} = \sum_t \sum_k y_{kt} \log p_{kt}$

$\frac{\partial \mathcal{L}}{\partial \vec{w}_i} = \sum_t (\frac{y_{it}(p_{it} - {p_{it}}^2)}{p_{it}} - \sum_{k \neq i} \frac{y_{kt}p_{it}p_{kt}}{p_{kt}})\vec{x}_t = \sum_t (y_{it} - p_{it}\sum_k y_{kt})\vec{x}_t$

Because $\sum_k y_{kt} = 1$

So:

$\frac{\partial \mathcal{L}}{\partial \vec{w}_i} = \sum_t (y_{it} - p_{it})\vec{x}_t$

4.3 Convergence of gradient descent

(a)

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

$f'(x_n) = \alpha(x_n - x_*) = \alpha\varepsilon_n$

$\varepsilon_{n+1} = x_{n+1} - x_* = x_n - \eta f'(x_n) - x_* = \varepsilon_n - \eta\alpha\varepsilon_n = (1 - \eta\alpha)\varepsilon_n$

$So:$

$\varepsilon_n = (1 - \eta\alpha)^n \varepsilon_0$

(b)

If the update rule converges to the minimum, then it needs: $0 <= \varepsilon_{n+1} < \varepsilon_n$

which means:

$0 <= (1 - \eta\alpha) < 1$

then:

$0 < \eta <= 1/\alpha$

If we want to get the fastest convergence, simply set $(1 - \eta\alpha) = 0$ and $\eta = 1/\alpha$

and because $f'(x_n) = \alpha(x_n - x_*)$, $f''(x_n) = \alpha$

So the step size corresponding to the fastest convergence is the reciprocal of $f''(x_n)$

(c)

$\varepsilon_{n+1} = x_{n+1} - x_* = x_n - \eta f'(x_n) + \beta(x_n - x_{n-1}) - x_* = x_n - x_* - \eta\alpha(x_n - x_*) + \beta(x_n - x_* - x_{n-1} + x_*) = \varepsilon_n - \eta\alpha\varepsilon_n + \beta(\varepsilon_n - \varepsilon_{n-1}) = (1 - \alpha\eta + \beta)\varepsilon_n - \beta\varepsilon_{n-1}$

(d)

$\varepsilon_{n+1} = (1 - \alpha\eta + \beta)\varepsilon_n - \beta\varepsilon_{n-1} = (1 - 4/9 + 1/9)\varepsilon_n - (1/9)\varepsilon_{n-1} = (2/3)\varepsilon_n - (1/9)\varepsilon_{n-1}$

$\varepsilon_{n+1} - (1/3)\varepsilon_n = (1/3)(\varepsilon_n - (1/3)\varepsilon_{n-1})$

So:

$\varepsilon_{n+1} - (1/3)\varepsilon_n = (1/3)^n(\varepsilon_1 - (1/3)\varepsilon_0)$

$(1/3)\varepsilon_n - (1/3)^2\varepsilon_{n-1} = (1/3)^n(\varepsilon_1 - (1/3)\varepsilon_0)$

$\dots$

$(1/3)^{n-1}\varepsilon_2 - (1/3)^n\varepsilon_1 = (1/3)^n(\varepsilon_1 - (1/3)\varepsilon_0)$

$\varepsilon_{n+1} - (1/3)^n\varepsilon_1 = (1/2)(1 - (1/3)^n)n(\varepsilon_1 - (1/3)\varepsilon_0)$

$\varepsilon_n - (1/3)^{n-1}\varepsilon_1 = (1/2)(1 - (1/3)^{n-1})(n-1)(\varepsilon_1 - (1/3)\varepsilon_0)$

If we set the momentum term at t = 0 is $(-1/2)\varepsilon_0$, which means $\varepsilon_1 = (1/3)\varepsilon_0$

then:

$\varepsilon_n = (1/3)^n\varepsilon_0$

So: $\lambda = 1/3$ in this case.

The rate of convergence is bigger than the result in (a), where the $\varepsilon_n = (5/9)^n\varepsilon_0$ because the ε decreases faster.

4.4 Newton's method

(a)

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

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

$\varepsilon_n = |x_n - x_*| = |x_{n-1} - \frac{f'(x_{n-1})}{f''(x_{n-1})} - x_*| = |x_{n-1} - x_* - \frac{x_{n-1}-x_*}{2p-1}| = |\varepsilon_{n-1} - \frac{\varepsilon_{n-1}}{2p-1}| = |\frac{2p-2}{2p-1}\varepsilon_{n-1}|$

So:

$\varepsilon_n = |(\frac{2p-2}{2p-1})^n \varepsilon_0|$

(b)

$\varepsilon_n <= \delta\varepsilon_0$

$|(\frac{2p-2}{2p-1})^n \varepsilon_0| <= \delta\varepsilon_0$

$|\frac{2p-2}{2p-1}|^n <= \delta$

$n >= \frac{\log\delta}{\log|\frac{2p-2}{2p-1}|}$(if p$\neq$ 1)

n is 1 when p = 1

$\log|\frac{2p-2}{2p-1}| = \log\frac{2p-2}{2p-1} <= \frac{2p-2}{2p-1} - 1 = \frac{-1}{2p-1}$ if $p \neq 1$

$\frac{\log\delta}{\log|\frac{2p-2}{2p-1}|} >= \frac{-\log\delta}{\frac{1}{2p-1}} = (2p-1)\log\frac{1}{\delta}$

So:

$n >= (2p-1)\log\frac{1}{\delta}$ (if p $\neq$ 1)

n is 1 when p = 1

(c)

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

$f''(x) = \frac{x_*}{x^2} > 0$ for all x

Let $f'(x) = 0$, we get $x = x_*$ and we know for all x, $f''(x) > 0$

We can confirm the minimum occurs at $x = x_*$