Homework4 Report

Qi Liu

May 13, 2016

1 PAC Learning

True, PAC says that the error can be less that ϵ for any $\epsilon > 0$, but ϵ must be positive and can not reduce to zero.

2 L2 Norm

2.1 Linear Regression

Calculate the partial derivative of $J(\theta)$ respect to θ, we get

$$\frac{\partial}{\partial\theta}J(\theta) = \frac{\partial}{\partial\theta}(y - X\theta)^T(y - X\theta) + \lambda\frac{\partial}{\partial\theta}||\theta||_2^2 = 2(X^TX\theta - X^Ty) + 2\lambda\theta.$$

Set the partial derivative to 0, we get

$$\begin{aligned}
\frac{\partial}{\partial\theta}J(\theta) &= 0 \\
2(X^TX\theta - X^Ty) + 2\lambda\theta &= 0 \\
X^TX\theta + \lambda\theta &= X^Ty \\
(X^TX + \lambda I)\theta &= X^Ty \\
\theta &= (X^TX + \lambda I)^{-1}X^Ty
\end{aligned}$$

2.2 Logistic Regression

Calculate the partial derivative of $l(\theta)$ respect to θ, we get

$$\frac{\partial}{\partial\theta}l(\theta) = \frac{\partial}{\partial\theta}\log\prod_{i=1}^{m}h_\theta(x^{(i)})^{y^{(i)}}(1-h_\theta(x^{(i)}))^{1-y^{(i)}} - \lambda\frac{\partial}{\partial\theta}||\theta||_2^2 = \sum_{i=1}^{m}[y^{(i)} - h_\theta(x^{(i)})]x^{(i)} - 2\lambda\theta.$$

Thus the updating rule for θ is

$$\theta_{t+1} = \theta_t + \alpha(\sum_{i=1}^{m}[y^{(i)} - h_{\theta_t}(x^{(i)})]x^{(i)} - 2\lambda\theta_t)$$

where α is the learning rate.

3 EM Algorithm

3.1 Correctness of GEM algorithm

First, by Jensen's inequality, we can get

$$\ell(\theta^{(t+1)}) \geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i(z^{(i)})}.$$

And in the M-step of GEM algorithm, the α is chosen small enough that we do not decrease the objective function, which means $\forall i$,

$$p(x^{(i)}, z^{(i)}; \theta^{(t+1)}) \geq p(x^{(i)}, z^{(i)}; \theta^{(t)}).$$

This implies

$$\sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i(z^{(i)})} \geq \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i(z^{(i)})}.$$

And in the E-step, the Q is chosen to satisfy

$$\ell(\theta^{(t)}) = \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta^{(t)})}{Q_i(z^{(i)})}.$$

Thus we have $\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$.

3.2 Gradient Descent

We think the problem statement has a little mistakes. There are no such $Q_i(z^{(i)})$ in the gradient descent. We think the correct updating rule for the gradient descent should be

$$\theta := \theta + \alpha \nabla_\theta \ell(\theta) = \theta + \alpha \nabla_\theta \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta).$$

Assume we use the same learning rate α for gradient descent and GEM algorithm. The updating part in gradient descent is

$$\begin{aligned}
\nabla_\theta \sum_i \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta) &= \sum_i \frac{1}{\sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta)} \sum_{z^{(i)}} \nabla_\theta p(x^{(i)}, z^{(i)}; \theta) \\
&= \sum_i \frac{1}{p(x^{(i)}; \theta)} \sum_{z^{(i)}} \nabla_\theta p(x^{(i)}, z^{(i)}; \theta).
\end{aligned}$$

The updating part for GEM is

$$\nabla_\theta \sum_i \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i(z^{(i)})} = \sum_i \sum_{z^{(i)}} \frac{Q_i(z^{(i)})}{p(x^{(i)}, z^{(i)}; \theta)} \nabla_\theta p(x^{(i)}, z^{(i)}; \theta).$$

We know in the E-step of GEM algorithm, the Q will be chosen to satisfy

$$Q_i(z^{(i)}) = p(z_{(i)}|x_{(i)};\theta) = \frac{p(z_{(i)}, x_{(i)};\theta)}{p(x_{(i)};\theta)},$$

then

$$\begin{aligned}
\sum_i \sum_{z^{(i)}} \frac{Q_i(z^{(i)})}{p(x^{(i)}, z^{(i)};\theta)} \nabla_\theta p(x^{(i)}, z^{(i)};\theta) &= \sum_i \sum_{z^{(i)}} \frac{1}{p(x^{(i)};\theta)} \nabla_\theta p(x^{(i)}, z^{(i)};\theta) \\
&= \sum_i \frac{1}{p(x^{(i)};\theta)} \sum_{z^{(i)}} \nabla_\theta p(x^{(i)}, z^{(i)};\theta).
\end{aligned}$$

Which means gradient descent and GEM algorithm will update the same value.