Machine Learning Assignment-1 Report

102062703 魏偉哲

October 4, 2014

1. 平方和的結果會把誤差大的嚴重性放大，除了要求錯誤差距少以外更希望的是錯誤的差距不能太大，寧可是比較多小錯誤，而不太容許有大錯誤的出現；第二個取絕對值得和反而就沒有這些考量，只是單純的希望平均錯誤差距小而已。

2.
$$
\begin{aligned}
l(\beta) &= \sum_{t=1}^{N} \left\{ y^{(t)} log\pi \left(x^{(t)}; \beta \right) + \left(1 - y^{(t)} \right) log \left(1 - \pi \left(x^{(t)}; \beta \right) \right) \right\} \\
&= \sum_{t=1}^{N} \left\{ y^{(t)} log \left(\frac{e^{\beta^T \tilde{x}^{(t)}}}{e^{\beta^T \tilde{x}^{(t)}} + 1} \right) + \left(1 - y^{(t)} \right) log \left(1 - \frac{e^{\beta^T \tilde{x}^{(t)}}}{e^{\beta^T \tilde{x}^{(t)}} + 1} \right) \right\} \\
&= \sum_{t=1}^{N} \left\{ y^{(t)} log \left(e^{\beta^T \tilde{x}^{(t)}} \right) - y^{(t)} log \left(e^{\beta^T \tilde{x}^{(t)}} + 1 \right) + \left(y^{(t)} - 1 \right) log \left(e^{\beta^T \tilde{x}^{(t)}} + 1 \right) \right\} \\
&= \sum_{t=1}^{N} \left\{ y^{(t)} \beta^T \tilde{x}^{(t)} - log \left(1 + e^{\beta^T \tilde{x}^{(t)}} \right) \right\}
\end{aligned}
$$

3. (a) $\forall x, y \in C_i \cap C_j$
   Because $x, y \in C_i$, we have $(1-\theta)x + \theta y \in C_i$ for any $\theta \in [0,1]$
   Because $x, y \in C_j$, we have $(1-\theta)x + \theta y \in C_j$ for any $\theta \in [0,1]$
   Then, we know $(1-\theta)x + \theta y \in C_i \cap C_j$ for any $\theta \in [0,1]$

   It shows that intersection of two convex sets is convex. We can also show that the intersection of convex sets is convex by continue to intersect two convex sets.

   (b) According Appendix C, we know that affine function, $f(x) = e^{ax}$ and $f(x) = -log(x)$ are both convex functions. Moreover, we know non-negative weighted sum of convex functions is convex, and composition with monotone convex functions is convex too.

   $\left(y^{(t)} \beta^T \tilde{x}^{(t)} \right)$ is an affine function, which is convex. Let $f(x) = -log(x)$, $g(x) = e^x$, $h(x) = \beta^T x$, $(-log(1+e^{\beta^T \tilde{x}^{(t)}})) = f(1+g(h(\tilde{x})))$

is convex.(because of composition with convex functions). So the log-likelihood function for logistic regression is concave(because of non-negative weighted sum of convex functions).

4. (a)

$$X = \begin{bmatrix} 1 & x^{(1)} \\ 1 & x^{(2)} \\ \vdots & \vdots \\ 1 & x^{(N)} \end{bmatrix} r = \begin{bmatrix} r^{(1)} \\ r^{(2)} \\ \vdots \\ r^{(N)} \end{bmatrix} L = \frac{1}{2} \begin{bmatrix} l^{(1)} & 0 & \cdots & 0 \\ 0 & l^{(2)} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & l^{(N)} \end{bmatrix}$$

$$\begin{aligned}
(Xw-r)^T L(Xw-r) &= (Xw-r)^T \frac{1}{2} \begin{bmatrix} l^{(1)} & 0 & \cdots & 0 \\ 0 & l^{(2)} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & l^{(N)} \end{bmatrix} \begin{bmatrix} \begin{bmatrix} 1 & x^{(1)} \end{bmatrix} w - r^{(1)} \\ \begin{bmatrix} 1 & x^{(2)} \end{bmatrix} w - r^{(2)} \\ \vdots \\ \begin{bmatrix} 1 & x^{(N)} \end{bmatrix} w - r^{(N)} \end{bmatrix} \\
&= \frac{1}{2} \begin{bmatrix} \begin{bmatrix} 1 & x^{(1)} \end{bmatrix} w - r^{(1)} \\ \begin{bmatrix} 1 & x^{(2)} \end{bmatrix} w - r^{(2)} \\ \vdots \\ \begin{bmatrix} 1 & x^{(N)} \end{bmatrix} w - r^{(N)} \end{bmatrix} \begin{bmatrix} l^{(1)} \left(\begin{bmatrix} 1 & x^{(1)} \end{bmatrix} w - r^{(1)} \right) \\ l^{(2)} \left(\begin{bmatrix} 1 & x^{(2)} \end{bmatrix} w - r^{(2)} \right) \\ \vdots \\ l^{(N)} \left(\begin{bmatrix} 1 & x^{(N)} \end{bmatrix} w - r^{(N)} \right) \end{bmatrix} \\
&= \frac{1}{2} \sum_{i=1}^{N} l^{(i)} \left(w^T \begin{bmatrix} 1 \\ x^{(i)} \end{bmatrix} - r^{(i)} \right)^2
\end{aligned}$$

(b) let $L = L'^T L'$

$$\begin{aligned}
Xw-r)^T L(Xw-r) &= (Xw-r)^T L'^T L'(Xw-r) \\
&= (L'^T(Xw-r))^T L(L'^T(Xw-r))
\end{aligned}$$

let $X' = L'X$ and $r' = L'r$, the formula will become $(X'w-r')^T(X'w-r') = \|X'w-r'\|^2$, so

$$w = (X'^T X')^{-1} X'^T r' = (X^T LXT)^{-1} XT^L$$

(c)

$$
\begin{aligned}
\prod_{i=1}^{N} p\left(r^{(i)}|x^{(i)};w\right) &\propto \prod_{i=1}^{N} \exp\left(-\frac{\left(r^{(i)} - w^T\left[\begin{array}{c}1\\x^{(i)}\end{array}\right]\right)^2}{2\sigma^{(i)2}}\right) \\
&\propto \sum_{i=1}^{N} -\frac{\left(r^{(i)} - w^T\left[\begin{array}{c}1\\x^{(i)}\end{array}\right]\right)^2}{2\sigma^{(i)2}} \\
&= \frac{1}{2}\sum_{i=1}^{N} l^{(i)}\left(w^T\left[\begin{array}{c}1\\x^{(i)}\end{array}\right] - r^{(i)}\right)^2 \quad where \quad l^{(i)} = \frac{-1}{\sigma^{(i)2}}
\end{aligned}
$$

(d) see Figure 1, 在 train 時，用左除算出 w

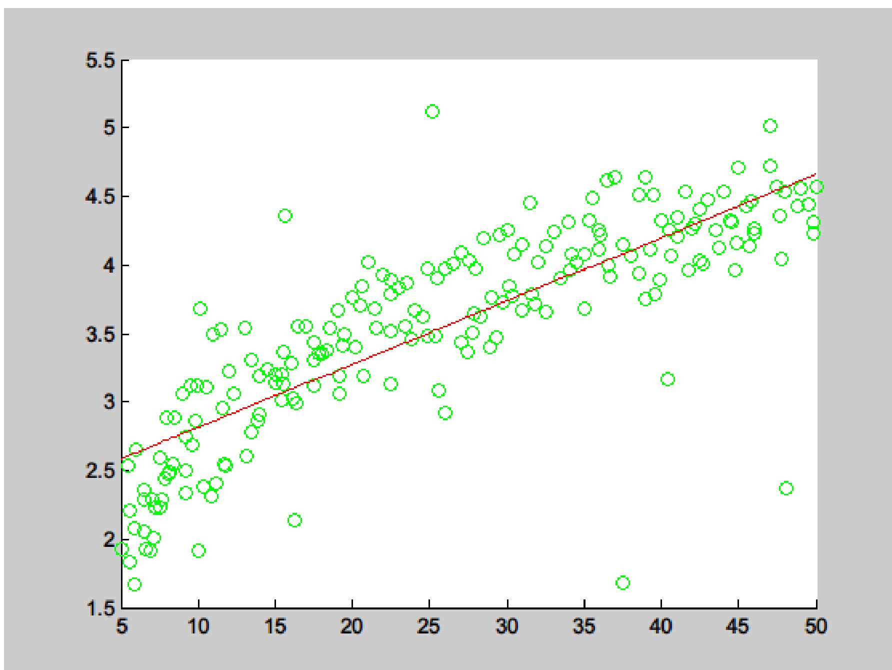
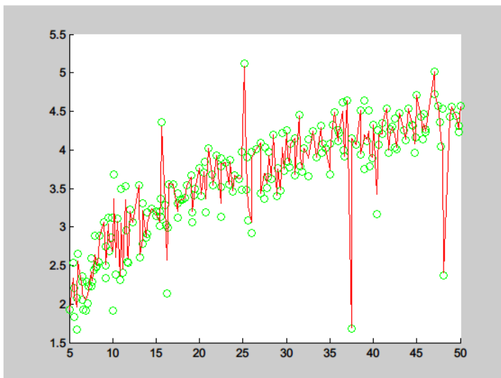


Figure 1: 4.(d)

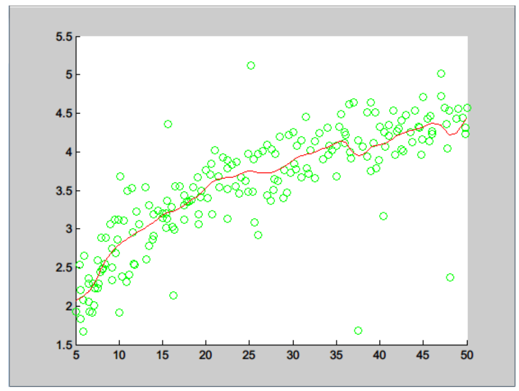
(e) see Figure 2, 在 predict 時，對於每一個 instance $\tilde{x}$，算出 L', X', r' (4-b 提到的)，再用 X', r' 左除算出的 w 後，拿去預測 $\tilde{x}$

(f) 當 τ 越小的話預測的曲線看起來越複雜，比較接近 training dataset 分布的樣子；反之，則是越接近直線，結果會越接近單純的 Linear Regression 得出的結果。有點像是 τ 的目的是調整要 training 的

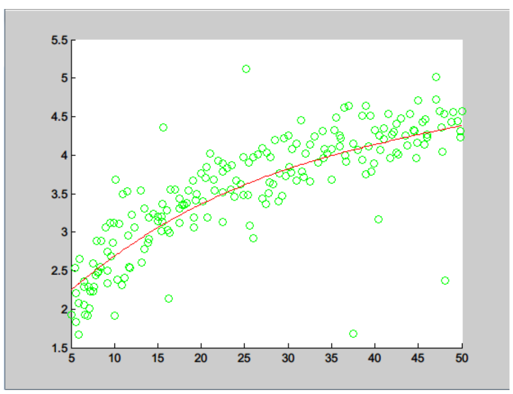
dataset 的範圍，範圍越小的話代表 traing dataset 跟 $\tilde{x}$ 越接近，predict 的結果就會很接近那個區間內的 dataset 結果，反之，當 τ 越大的話就會涵蓋較多的 dataset，當大到一定的程度後就等於是在做原來的 Linear Regression，因為每一個 dataset 裡面的 instance 都是 weight 1。



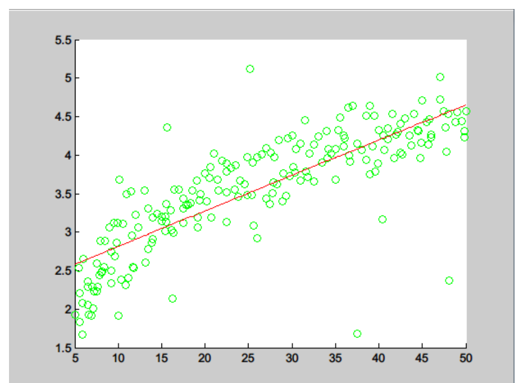
(a) $\tau = 0.1$



(b) $\tau = 1$



(c) $\tau = 10$



(d) $\tau = 100$

Figure 2: 4.(e)