1. What are the pros and cons between the regression hypotheses based on the objective

$$\arg_\theta \min \sum_{t=1}^{N} \left[r^{(t)} - h(\boldsymbol{x}^{(t)};\theta) \right]^2$$

and $\arg_\theta \min \sum_{t=1}^{N} \left| r^{(t)} - h(\boldsymbol{x}^{(t)};\theta) \right|$ respectively? (Hint: consider the training complexity and prediction accuracy)

A1:令

$$\arg_\theta \min \sum_{t=1}^{N} \left[r^{(t)} - h(\boldsymbol{x}^{(t)};\theta) \right]^2$$

為 1 式 Standard deviation*N

令

$\arg_\theta \min \sum_{t=1}^{N} \left| r^{(t)} - h(\boldsymbol{x}^{(t)};\theta) \right|$ 為 2 式 Mean absolute deviation*N

使用 Mean absolute deviation 的好處是：

單一計算較容易，不需要進行平方的運算

且較直覺,簡單.易於了解.

使用 Mean absolute deviation 的壞處是：

無法微分, 因此,在求此線性方程組近似解時,會較困難.

使用 Standard deviation 的好處是：

可微分,因此容易找到最佳近似解 Eg: $\boldsymbol{w}^* = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{r}$

如果有解,則為最佳解.

使用 Standard deviation 的壞處是：

衡量誤差是根據其平方值，可能有較大誤差。

2. In logistic regression, show that $l(\beta) = \sum_{t=1}^{N} \left\{ y^{(t)} \beta^{\top} \tilde{x}^{(t)} - \log\left(1 + e^{\beta^{\top} \tilde{x}^{(t)}}\right) \right\}$.

A2:因為

$$
\begin{aligned}
l(\beta) &= \log \prod_{t=1}^{N} p\left(x^{(t)}, r^{(t)} \mid \beta\right) \\
&= \log \prod_{t=1}^{N} P\left(r^{(t)} \mid x^{(t)}, \beta\right) p\left(x^{(t)} \mid \beta\right) \\
&\propto \log \prod_{t=1}^{N} \pi\left(x^{(t)}; \beta\right)^{q^{(t)}} \left(1 - \pi\left(x^{(t)}; \beta\right)\right)^{(1-q^{(t)})}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{l}(\beta) &= \sum_{t=1}^{N} q^t * \log \pi(x^t; \beta) + (1 - q^t) * \log(1 - \pi(x^t; \beta)) \\
&= \sum_{t=1}^{N} \log(1 - \pi(x^t; \beta) + q^t * [\log \pi(x^t; \beta) - \log(1 - \pi(x^t; \beta)]
\end{aligned}
$$

$$
\because \log(1 - \pi(x^t; \beta) = \log\left(1 - \frac{1}{1 + e^{-B^T x}}\right) = \log\left(\frac{e^{-B^T x}}{1 + e^{-B^T x}}\right) = \log\left(\frac{1}{\frac{1}{e^{-B^T x}} + 1}\right)
$$

$$
= -\log\left(1 + e^{B^T x}\right)
$$

$$
\because q^t * [\log \pi(x^t; \beta) - \log(1 - \pi(x^t; \beta)] = q^t \left[\log \frac{\pi(x^t; \beta)}{1 - \pi(x^t; \beta)}\right] = q^t \left[\log \frac{\frac{1}{1 + e^{-B^T x}}}{1 - \frac{1}{1 + e^{-B^T x}}}\right]
$$

$$
= q^t \left[\log\left(\frac{1}{1 + e^{-B^T x} - 1}\right)\right] = q^t \left[\log\left(\frac{1}{e^{-B^T x}}\right)\right] = -q^t [\log e^{-B^T x}] = q^t B^T x
$$

$$
\therefore \mathrm{l}(\beta) = \sum_{t=1}^{N} \{-\log\left(1 + e^{B^T x}\right) + q^t B^T x\}
$$

得證

3. Read Appendix C on the definitions of convex set and functions.

(a) Show that the intersection of convex sets, $\bigcap_{i\in\mathbb{N}} C_i$ where $C_i \subseteq \mathbb{R}^n$, is convex.

(b) Show that the log-likelihood function for logistic regression, $l(\boldsymbol{\beta})$, is concave.

A3:

(a) 依照定理

Given a convex set $C_1, C_2 \subseteq \mathbb{R}^n$,

- Scaling: $\beta C = \{\beta \boldsymbol{x} : \boldsymbol{x} \in C\}$ is convex for any $\beta \in \mathbb{R}$
- Sum: $C_1 + C_2 = \{\boldsymbol{x}_1 + \boldsymbol{x}_2 : \boldsymbol{x}_1 \in C_1, \boldsymbol{x}_2 \in C_2\}$ is convex

設 x,y $\in C1 \cap C2$ 且 $0 \le \alpha \le 1$

$\therefore$ x, y 屬於 C1 與 C2

因(依題目設定,C1, C2 皆為 convex set)

$\therefore\ \alpha x + (1-\alpha)y \in C12 = C1 \cap C2$ <-using 定理 Scaling and Sum

證明 C12 為凸集合

重複上述動作 證明 C12 ∩ C3 為凸集合

令此交集為 C123

重複上述動作直到 證明 C12…n-1 ∩ Cn 為凸集合

得證.

(b)(符號使用與 a 小題不同) 依照 concave 定義

A real-valued function *f* on an interval (or, more generally, a convex set in vector space) is said to be *concave* if, for any *x* and *y* in the interval and for any *t* in [0,1],

$$f((1-t)x + (t)y) \ge (1-t)f(x) + (t)f(y).$$

令$\gamma\alpha + (1-\gamma)\beta \in C$(convex set)

$$l(\boldsymbol{\beta}) = \sum_{t=1}^{N} \left\{ y^{(t)} \boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^{(t)} - \log\left(1 + e^{\boldsymbol{\beta}^\top \tilde{\boldsymbol{x}}^{(t)}}\right) \right\}$$

對於任意$\alpha, \beta \in C$ 且 $0 \le \gamma \le 1$

$l(\gamma\alpha + (1-\gamma)\beta) = \sum_{t=1}^{N}\{y^t(\gamma\alpha + (1-\gamma)\beta]^T)X^t - \log(1 + e^{(\gamma\alpha+(1-\gamma)\beta)^T X^t})\}$

$\gamma * l(\alpha) + (1-\gamma) * l(\beta) =$

$$\sum_{t=1}^{N} \gamma\left\{y^t \alpha^T X^t - \log\left(1 + e^{(\alpha)^T X^t}\right)\right\} + \sum_{t=1}^{N} (1-\gamma)\{y^t \beta^T X^t - \log\left(1 + e^{(\beta)^T X^t}\right)\}$$

$$= \sum_{t=1}^{N} y^t\{\gamma\alpha^T + (1-\gamma)(\beta^T)\}X^t - \gamma\left\{\log\left(1 + e^{(\alpha)^T X^t}\right)\right\} - (1-\gamma)\{\log\left(1 + e^{(\beta)^T X^t}\right)\}$$

$= \sum_{t=1}^{N} y^t\{\gamma\alpha^T + (1-\gamma)(\beta^T)\}X^t - \{\log\left(1 + e^{(\alpha)^T X^t}\right)^{\gamma} + \log\left(1 + e^{(\beta)^T X^t}\right)^{(1-\gamma)}\}$

顯然

$l(\gamma\alpha + (1-\gamma)\beta) \ge \gamma * l(\alpha) + (1-\gamma) * l(\alpha)$

所以 $l(\boldsymbol{\beta})$ is concave

4. Consider the locally weighted linear regression problem with the following objective:

$$\arg\min_{\boldsymbol{w}\in\mathbb{R}^{d+1}} \frac{1}{2}\sum_{i=1}^{N} l^{(i)}(\boldsymbol{w}^\top \begin{bmatrix} 1 \\ \boldsymbol{x}^{(i)} \end{bmatrix} - r^{(i)})^2$$

local to a given instance $\boldsymbol{x}'$ whose label will be predicted, where $l^{(i)} = \exp(-\frac{(\boldsymbol{x}'-\boldsymbol{x}^{(i)})^2}{2\tau^2})$ for some constant τ.

(a) Show that the above objective can be written as the form

$$(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{r})^\top \boldsymbol{L}(\boldsymbol{X}\boldsymbol{w}-\boldsymbol{r}).$$

Specify clearly what $\boldsymbol{X}$, $\boldsymbol{r}$, and $\boldsymbol{L}$ are.

(b) Give a close form solution to $\boldsymbol{w}$. (Hint: recall that we have $\boldsymbol{w} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{r}$ in linear regression when $l^{(i)} = 1$ for all i)

(c) Suppose that the training examples $(\boldsymbol{x}^{(i)}, r^{(i)})$ are i.i.d. samples drawn from some joint distribution with the marginal:

$$p(r^{(i)}|\boldsymbol{x}^{(i)};\boldsymbol{w}) = \frac{1}{\sqrt{2\pi\sigma^{(i)}}}\exp\left(-\frac{(r^{(i)} - \boldsymbol{w}^\top \begin{bmatrix} 1 \\ \boldsymbol{x}^{(i)} \end{bmatrix})^2}{2\sigma^{(i)2}}\right)$$

where $\sigma^{(i)}$'s are constants. Show that finding the maximum likelihood of $\boldsymbol{w}$ reduces to solving the locally weighted linear regression problem above. Specify clearly what the $l^{(i)}$ is in terms of the $\sigma^{(i)}$'s.

$$\sum_i l(\boldsymbol{x}^{(i)};\boldsymbol{x}')(r^{(i)} - \boldsymbol{w}^\top \begin{bmatrix} 1 \\ \boldsymbol{x}^{(i)} \end{bmatrix})^2$$

A4: Local Weight Linear Regression

(a) 自行定義符號 l 為 n ∗ 1 矩陣 li 為 $\mathrm{l}(x^i;\mathrm{x}')$, $\mathrm{L} = (\sqrt{l})(\sqrt{l})^T$ *為 n ∗ n矩陣*

原式可寫成 $= \left((\sqrt{l})^T(Xw-r)\right)^T\left((\sqrt{l})^T(Xw-r)\right) = (Xw-r)^T(\sqrt{l})(\sqrt{l})^T(Xw-r)$

$= (Xw-r)^T\,\mathrm{L}(Xw-r)$ 得證

(b)$(Xw-r)^T\,\mathrm{L}(Xw-r)$ ，對 W 作微分 ，並令其一階微分為零

$\because \nabla(\,(Xw-r)^T\,\mathrm{L}(Xw-r)\,) = (X^T)(\mathrm{L}+L^T)(\mathrm{Xw}-\mathrm{r}) = 2(X^T)(\mathrm{L})(\mathrm{Xw}-\mathrm{r}) = 0$

$\therefore\ \mathrm{w} = (X^TLX)^{-1}X^TLr$

(c)$\because \mathrm{P(W|X)} = \frac{P(X|W)*P(W)}{P(X)} \propto \mathrm{P(X|W)}$, and we assume P (W) all same, '(X ,r) are i.i.d'

$\because$ by maxizizing likehood W

$\therefore \log P(X|W) = \log(\prod_{t=1}^{N} P(x^t,r^t|W)) = \log\left(\prod_{t=1}^{N} P(r^t|x^t,W) * P(x^t|W)\right)$

$$\propto \log\left(\prod_{t=1}^{N}\frac{1}{\sqrt{2\pi\sigma^t}}e^{\left(\frac{-\left(r^t-w^t\begin{bmatrix}1\\x^t\end{bmatrix}\right)^2}{2(\sigma^t)^2}\right)}\right) = \sum_{t=1}^{N} -\log\left(\sqrt{2\pi\sigma^t}\right) - \left(\frac{\left(r^t-w^t\begin{bmatrix}1\\x^t\end{bmatrix}\right)^2}{2(\sigma^t)^2}\right)$$

$\propto \sum_{t=1}^{N} -\left(\frac{\left(r^t-w^t\begin{bmatrix}1\\x^t\end{bmatrix}\right)^2}{2(\sigma^t)^2}\right) = -\frac{1}{2}\sum_{t=1}^{N}\frac{1}{(\sigma^t)^2}\left(r^t-w^t\begin{bmatrix}1\\x^t\end{bmatrix}\right)^2$, $l^t = \frac{1}{(\sigma^t)^2}$ 得證.

Coding problem

(d) Implement a linear regressor (see the spec for more details) on the provided 1D dataset. Plot the data and your fitted line. (Hint: don't forget the intercept term)

(e) Implement 4 locally weighted linear regressors (see the spec for more details) on the same dataset with τ = 0.1, 1, 10, and 100 respectively. Plot the data and your 4 fitted curves (for different $\boldsymbol{x}$'s within the dataset range).

(f) Discuss what happens when τ is too small or large.