Machine Learning Assignment 01

103061527 李豪韋

1. What is the difference in terms of the performance between hypotheses based on the objective $\arg_\theta \min \sum_{t=1}^{N}[r^{(t)} - h(x^{(t)};\theta)]^2$ and $\arg_\theta \min \sum_{t=1}^{N} |r^{(t)} - h(x^{(t)};\theta)|$ respectively?

$[r^{(t)} - h(x^{(t)};\theta)]^2$ has an analytic form that makes finding solutions easier, but suffers from difference accuracy issue. $|r^{(t)} - h(x^{(t)};\theta)|$ cannot be differentiated directly, but can estimate difference in a more accurate way.

---

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

$$
\begin{aligned}
l(\beta) = log(p(y|\widetilde{x};\beta)) &= log\left(\prod_{t=1}^{N}(\pi(\widetilde{x}^{(t)};\beta))^{y^{(t)}}(1-\pi(\widetilde{x}^{(t)};\beta))^{(1-y^{(t)})}\right) \\
&= \sum_{t=1}^{N} y^{(t)} log(\pi(\widetilde{x}^{(t)};\beta)) + \sum_{t=1}^{N}(1-y^{(t)})log(1-\pi(\widetilde{x}^{(t)};\beta)) \\
&= \sum_{t=1}^{N} y^{(t)} log(\frac{1}{1+\exp(-\beta^T\widetilde{x}^{(t)})}) + \sum_{t=1}^{N}(1-y^{(t)})log(\frac{\exp(-\beta^T\widetilde{x}^{(t)})}{1+\exp(-\beta^T\widetilde{x}^{(t)})}) \\
&= \sum_{t=1}^{N} y^{(t)}\beta^T\widetilde{x}^{(t)} - \sum_{t=1}^{N} log(1+\exp(\beta^T\widetilde{x}^{(t)}))
\end{aligned}
$$

---

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

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

Let $x, y \in \bigcap_{i\in N} C_i$, then $\forall k \in N, \quad x, y \in C_k$

By definition, $(1-\theta)x + \theta y \in C_k, \quad \forall k \in N, \quad \theta \in [0,1]$

And because $(1-\theta)x + \theta y \in C_k, \quad \forall k \in N, \quad \theta \in [0,1]$

Hence $(1-\theta)x + \theta y \in \bigcap_{i\in N} C_i, \quad \theta \in [0,1]$

Therefore, $\bigcap_{i\in N} C_i$ is also a convex.

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

Let $\beta \subseteq \mathbb{F}^{n\times 1}$, which is a vector space whose elements are contained by a field $\mathbb{F}$, where n is the dimension of $x^{(t)} \quad \forall t \in [1, N], t \in \mathbb{N}$. Obviously, vector spaces are convex.
Define $f(\beta; v) = -\langle \beta, v\rangle = -\beta^T v$, where $v \in \mathbb{F}^{n\times 1}$.
Because of linearity of matrix computation, we have

$$\begin{aligned}
f(\theta x + (1-\theta)y; v) = -\langle \theta x + (1-\theta)y, v\rangle &= -(\theta\langle x, v\rangle + (1-\theta)\langle y, v\rangle) \\
&= \theta f(x; v) + (1-\theta) f(y; v) \\
&\leq \theta f(x; v) + (1-\theta) f(y; v) \quad \forall x, y \in \beta \quad \theta \in [0, 1]
\end{aligned}$$

Hence $f : \mathbb{F}^{n\times 1} \to \mathbb{F}$ is convex.
Define $g(v) = log(1 + e^v)$ where $v \in \mathbb{F}$, and v can be expressed as $\theta x + (1-\theta)y$ where $x, y \in \mathbb{F} \quad \theta \in [0, 1]$, then

$$\begin{aligned}
g(\theta x + (1-\theta)y) &= log(1 + e^{(\theta x + (1-\theta)y)}) \\
&\leq log(1 + e^{\theta x} + e^{(1-\theta)y} + e^{\theta x + (1-\theta)y}) \\
&= log\big((1 + e^{\theta x})(1 + e^{(1-\theta)y})\big) \\
&\leq log\big((1 + e^{x})^{\theta}(1 + e^{y})^{(1-\theta)}\big) \\
&= \theta log(1 + e^x) + (1-\theta) log(1 + e^y) \\
&= \theta g(x) + (1-\theta) g(y)
\end{aligned}$$

Hence $g : \mathbb{F} \to \mathbb{F}$ is convex.
Then consider $-l(\beta)$:

$$-l(\beta) = -\sum_{t=1}^{N} y^{(t)} \beta^T x^{(t)} + \sum_{t=1}^{N} log(1 + \exp(\beta^T x^{(t)})) = \sum_{t=1}^{N} \Big[y^{(t)} f(\beta; x^{(t)}) + g(f(\beta; x^{(t)})) \Big]$$

$y^{(t)} f(\beta; x^{(t)})$ is convex $\forall t \in [1, N], t \in \mathbb{N} \quad (\because y^{(t)}$ is a scalar with the value of either 1 or 0). In addition, $g(f(\beta; v))$ is also convex because it is the convex function of a convex set. Therefore, $-l(\beta)$ is convex because it can be expressed as a summation of convex sets. Hence, $l(\beta)$ is concave.

---

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

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

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

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

$$(Xw - r)^T L(Xw - r).$$

Specify clearly what X, r, and L are.

$$\frac{1}{2}\sum_{t=1}^{N} l^{(i)} \; (w^T \begin{bmatrix} 1 \\ x^{(i)} \end{bmatrix} - r^{(i)})^2 = \langle A,\, B \rangle$$

$$A = \begin{pmatrix} 1 & x_1^{(1)} & \cdots & x_d^{(1)} \\ 1 & x_1^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(N)} & \cdots & x_d^{(N)} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix} - \begin{pmatrix} r^{(1)} \\ r^{(2)} \\ \vdots \\ r^{(N)} \end{pmatrix} = Xw - r$$

$$\begin{aligned} B &= \frac{1}{2}\left[\begin{pmatrix} l^{(1)}(1 & x_1^{(1)} & \cdots & x_d^{(1)}) \\ l^{(2)}(1 & x_1^{(2)} & \cdots & x_d^{(2)}) \\ \vdots & \vdots & \vdots & \vdots \\ l^{(N)}(1 & x_1^{(N)} & \cdots & x_d^{(N)}) \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix} - \begin{pmatrix} l^{(1)}r^{(1)} \\ l^{(2)}r^{(2)} \\ \vdots \\ l^{(N)}r^{(N)} \end{pmatrix} \right] \\ &= \begin{pmatrix} \frac{l^{(1)}}{2} & 0 & \cdots & 0 \\ 0 & \frac{l^{(2)}}{2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{l^{(N)}}{2} \end{pmatrix} \left[\begin{pmatrix} 1 & x_1^{(1)} & \cdots & x_d^{(1)} \\ 1 & x_1^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(N)} & \cdots & x_d^{(N)} \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{pmatrix} - \begin{pmatrix} r^{(1)} \\ r^{(2)} \\ \vdots \\ r^{(N)} \end{pmatrix} \right] \\ &= L(Xw - r) \end{aligned}$$

Hence:
$X^{(t)} = \begin{pmatrix} 1 & x^{(t)} \end{pmatrix}$, $\forall t \in [1, N] \quad t \in \mathbb{N}$
$r_t = r^{(t)}$, $\forall t \in [1, N] \quad t \in \mathbb{N}$
$L_{ij} = \frac{l^{(i)}}{2}\delta_{ij}$, $\forall i, j \in [1, N] \quad i, j \in \mathbb{N}$ (i.e. L is a diagonal matrix, $L_{ii} = \frac{l^{(i)}}{2}$)

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

Find w_{exm} such that $\triangledown \left[\frac{1}{2}\sum_{t=1}^{N} l^{(i)} \; (w^T \begin{bmatrix} 1 \\ x^{(i)} \end{bmatrix} - r^{(i)})^2 \right] \Bigg|_{w = w_{exm}} = 0.$

$$\begin{aligned} (Xw - r)^T L(Xw - r) &= (w^T X^T - r^T) L(Xw - r) \\ &= w^T X^T L X w - w^T X^T L r - r^T L X w + r^T L r \\ \triangledown \left[(Xw - r)^T L(Xw - r) \right] &= \triangledown (w^T X^T L X w - w^T X^T L r - r^T L X w + r^T L r) \\ &= 2 X^T L^T X w - X^T L r - X^T L^T r \\ &= 2(X^T L X w - X^T L r) \end{aligned}$$

$$2(X^T L X w_{exm} - X^T L r) = 0 \implies w_{exm} = (X^T L X)^{-1} (X^T L) r$$

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

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

where $\sigma^{(i)}$'s are constants. Show that finding the maximum likelihood of 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.

Suppose the transformation from dataset to label can be expressed as:

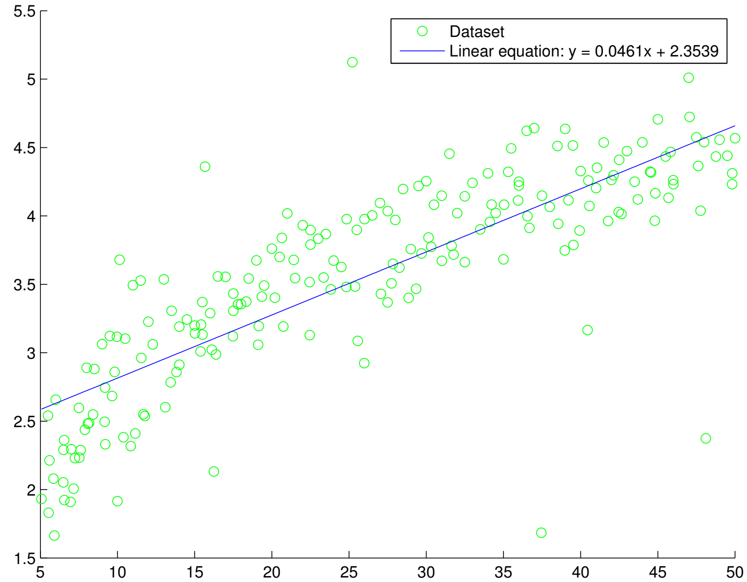
$$r^{(i)} = h(x^{(i)}; w) + \epsilon, \text{ where } h(x^{(i)}; w) = w^T \begin{bmatrix} 1 \\ x^{(i)} \end{bmatrix} \text{ and } \epsilon \sim \mathcal{N}(0, \sigma).$$

$$\begin{aligned}
log\big(p(r\big|x; w)\big) &= log\big(p(h(x; w) + \epsilon\big|x; w)\big) \\
&= log\Big(\prod_{x^{(i)} \in x} p(h(x^{(i)}; w) + \epsilon\big|x^{(i)}; w)\Big) \\
&= log\Bigg(\Big(\frac{1}{\sqrt{2\pi\sigma^{(i)}}}\Big)^{|x|} \prod_{x^{(i)} \in x} \exp\Big(- \frac{\big(r^{(i)} - h(x^{(i)}; w)\big)^2}{2\sigma^{(i)2}}\Big)\Bigg) \\
&= O\Bigg(- \sum_{x^{(i)} \in x} \frac{\big(r^{(i)} - h(x^{(i)}; w)\big)^2}{2\sigma^{(i)2}}\Bigg) \\
&= O\Bigg(- \frac{1}{2} \sum_{x^{(i)} \in x} l^{(i)} \big(r^{(i)} - w^T \begin{bmatrix} 1 \\ x^{(i)} \end{bmatrix}\big)^2\Bigg) \\
\implies l^{(i)} &= - \frac{1}{\sigma^{(i)2}}
\end{aligned}$$

(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)

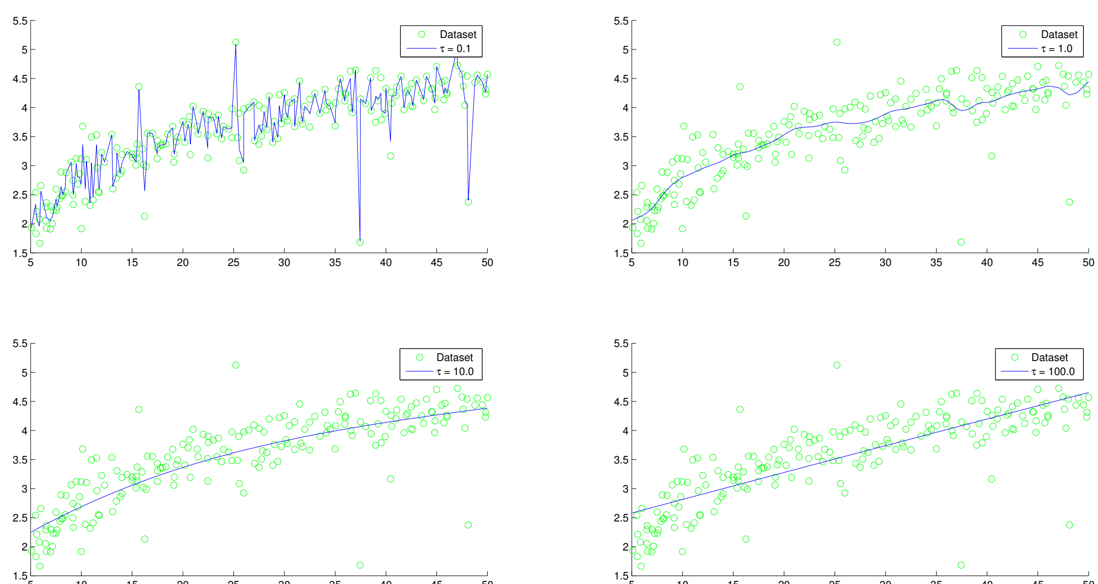
The regression is implemented with LMS algorithm.

$$w = \arg \min_{w \in \mathbb{R}^{2\times 1}} \left\| \begin{pmatrix} 1 & x^{(1)} \\ 1 & x^{(2)} \\ \vdots & \vdots \\ 1 & x^{(N)} \end{pmatrix} \begin{pmatrix} w^{(1)} \\ w^{(2)} \end{pmatrix} - \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{pmatrix} \right\|^2$$



Line equation: $y = 0.0461x + 2.3539$, MSE $= 0.1971$.

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



τ	0.1	1	10	100
MSE	0.0234	0.1478	0.1650	0.1962

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

Obviously, $\lim_{\tau \to \infty} l^{(i)} = 1$. Therefore, the larger τ we give, the more similar locally weighted linear regressors are to linear regressors. On the other hand, the line we obtain tends to fit the training set when τ is getting smaller, for minimizing the objective function because the nearer points have greater weights. Therefore, the constant τ can be regarded as the sensitivity of regressors to the data noise.
