Solution of Assignment 1

October 30, 2014

1. What is the difference in terms of the performance 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?

   Answer :

   Since $f(x) = x^2$ grows faster than $f(x) = |x|$ with growing x, $\arg_\theta \min \sum_{t=1}^{N} \left[r^{(t)} - h(\boldsymbol{x}^{(t)};\theta)\right]^2$ will be more sensitive to outliner than $\arg_\theta \min \sum_{t=1}^{N} \left|r^{(t)} - h(\boldsymbol{x}^{(t)};\theta)\right|$. Furthermore, $\arg_\theta \min \sum_{t=1}^{N} \left[r^{(t)} - h(\boldsymbol{x}^{(t)};\theta)\right]^2$ is easy to be differential.

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

   Answer :

   As we know, $\phi = \pi(x;\beta) = \frac{e^{\beta^T \tilde{x}}}{e^{\beta^T \tilde{x}}+1} = \frac{1}{e^{-\beta^T \tilde{x}}+1}$.

   $l(\beta) = \sum_{t=1}^{N} \left\{ y^{(t)} \log \pi(x;\beta) + (1-y^{(t)}) \log(1-\pi(x;\beta))\right\}$

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

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

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

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

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.

   Answer :

   Let $x, y \in \bigcap_{i \in N} C_i$, and let $m = (1-\theta)x + \theta y,\ \theta \in [0,1]$. Then $m \in C_1$ because C_1 is convex. Similarly, $m \in C_i,\ \forall i \in N$ because C_i are convex. Therefore, $m \in \bigcap_{i \in N} C_i$, which implies that $\bigcap_{i \in N} C_i$ is convex.

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

   Answer :

   The log-likelyhood function for logistic regression is $l(\beta) = \sum_{t=1}^{N} \left\{ y^{(t)} \beta^T \tilde{x}^{(t)} - log(1 + e^{\beta^T \tilde{x}^{(t)}})\right\}$. Based on the characteristic that the composition with monotone convex function is also convex (p.26 of appendix C), $log(1+e^{\beta^T \tilde{x}^{(t)}})$ is a convex function, so $-log(1+e^{\beta^T \tilde{x}^{(t)}})$ is concave.

$(y^{(t)}\beta^T\tilde{x}^{(t)} - log(1+e^{\beta^T\tilde{x}^{(t)}}))$ is also concave because $y^{(t)}\beta^T\tilde{x}^{(t)}$ is linear. $l(\beta)$ is the sum of concave functions. Therefore, it 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.

(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 $\tau = 0.1$, 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.

Answer :

(a) $X = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_d^{(1)} \\ 1 & x_1^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(N)} & \cdots & x_d^{(N)} \end{bmatrix}$, $w = [w_0, w_1, \cdots, w_d]^T$, $r = [r^{(1)}, r^{(2)}, \cdots, r^{(N)}]$, L is ans identity matrix with diagonal elements $[\frac{l^{(1)}}{2}, \frac{l^{(2)}}{2}, \cdots, \frac{l^{(N)}}{2}]$.

(b) $w = (X^TLX)^{-1}X^TLr$.

(c) $\arg_w \max\ p(w|X) = \arg_w \max\ p(X|w)$ (by Baye's theorem)
$= \arg_w \max\ \prod_{i=1}^{N} p(x^{(i)}, r^{(i)}|w) = \arg_w \max\ \ln \prod_{i=1}^{N} p(r^{(i)}|x^{(i)}, w)p(x^{(i)}|w)$
$= \arg_w \max\ \ln \prod_{i=1}^{N} p(r^{(i)}|x^{(i)}, w) = \arg_w \max\ \sum_{i=1}^{N} \ln\ p(r^{(i)}|x^{(i)}, w)$

$$= \arg_w \max\ \sum_{i=1}^{N} \ln \left(\frac{1}{\sqrt{2\pi\sigma^{(i)}}} \exp\left(-\frac{(r^{(i)} - w^T \begin{bmatrix} 1 \\ x^{(i)} \end{bmatrix})^2}{2\sigma^{(i)2}} \right) \right)$$

$$= \arg_w \max\ \sum_{i=1}^{N} \left(\ln \frac{1}{\sqrt{2\pi\sigma^{(i)}}} + \ln \exp\left(-\frac{(r^{(i)} - w^T \begin{bmatrix} 1 \\ x^{(i)} \end{bmatrix})^2}{2\sigma^{(i)2}} \right) \right)$$

$$= \arg_w \max\ \sum_{i=1}^{N} \left(\ln \frac{1}{\sqrt{2\pi\sigma^{(i)}}} + -\frac{(r^{(i)} - w^T \begin{bmatrix} 1 \\ x^{(i)} \end{bmatrix})^2}{2\sigma^{(i)2}} \right)$$

$= \arg_w \max\ \sum_{i=1}^{N} \left(-\frac{(r^{(i)} - w^T \begin{bmatrix} 1 \\ x^{(i)} \end{bmatrix})^2}{2\sigma^{(i)2}} \right)$ (Since $\ln \frac{1}{\sqrt{2\pi\sigma^{(i)}}}$ is irrelevant to w, it can be ignored)

$$= \arg_w \min\ \sum_{i=1}^{N} \left(\frac{(r^{(i)} - w^T \begin{bmatrix} 1 \\ x^{(i)} \end{bmatrix})^2}{2\sigma^{(i)2}} \right) == \arg_w \min\ \sum_{i=1}^{N} \left(\frac{1}{2\sigma^{(i)2}} (r^{(i)} - w^T \begin{bmatrix} 1 \\ x^{(i)} \end{bmatrix})^2 \right)$$

So, $l^{(i)} = \frac{1}{\sigma^{(i)2}}$.

(d) see the coding solution

(e) see the coding solution

(f) When τ is too large, the predictions become almost the same as linear regression. When τ is too small, the predictions are sensitive to local data points and tend to be influenced by outliers easily.