Introduction to Statistical Machine Learning
Homework 1

Yota Toyama

October 20, 2016

1.

$$
\begin{aligned}
\frac{\partial}{\partial \mathbf{w}} R(\mathbf{w}) &= 0 \\
\frac{\partial}{\partial \mathbf{w}} E_{p(\mathbf{x},y)} \left[(y - \mathbf{w}^T \mathbf{x})^2 \right] &= 0 \\
E_{p(\mathbf{x},y)} \left[2(y - \mathbf{w}^T \mathbf{x})(-\mathbf{x}) \right] &= 0 \\
E_{p(\mathbf{x},y)} \left[(y - \mathbf{w}^T \mathbf{x})\mathbf{x} \right] &= 0 \\
\mathbf{a}^T E_{p(\mathbf{x},y)} \left[(y - \mathbf{w}^T \mathbf{x})\mathbf{x} \right] &= 0 \\
E_{p(\mathbf{x},y)} \left[(y - \mathbf{w}^T \mathbf{x})\mathbf{a}^T \mathbf{x} \right] &= 0
\end{aligned}
$$

2.

$$
\begin{aligned}
E_{p(\mathbf{x},y)} \left[y - \mathbf{w}^T \mathbf{x} \right] &= 0 \\
E_{p(\mathbf{x},y)} \left[(y - \mathbf{w}^T \mathbf{x}) E_{p(\mathbf{x})} \left[\mathbf{x} \right] \right] &= 0 \\
E_{p(\mathbf{x},y)} \left[(y - \mathbf{w}^T \mathbf{x})\mathbf{x} \right] + E_{p(\mathbf{x},y)} \left[(y - \mathbf{w}^T \mathbf{x}) E_{p(\mathbf{x})} \left[\mathbf{x} \right] \right] &= 0 \\
E_{p(\mathbf{x},y)} \left[(y - \mathbf{w}^T \mathbf{x})(\mathbf{x} - E_{p(\mathbf{x})} \left[\mathbf{x} \right]) \right] &= 0
\end{aligned}
$$

$\therefore$ the correration between data and prediction errors is 0.

3.

$$
\begin{aligned}
\hat{\mathbf{w}} &= \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 \\
&= (X^T X)^{-1} X^T y
\end{aligned}
$$

$$
\text{Let } C \in \mathbb{R}^{(d+1)\times(d+1)} \text{ be a diagonal matrix \; s.t. } \tilde{X} = XC
$$

$$
\begin{aligned}
\hat{\tilde{\mathbf{w}}} &= \underset{\mathbf{w}}{\operatorname{argmin}} \sum_{i=1}^{N} (y_i - \mathbf{w}^T \tilde{\mathbf{x}}_i)^2 \\
&= (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T y \\
&= ((XC)^T XC)^{-1} (XC)^T y \\
&= (C X^T X C)^{-1} C X^T y \\
&= C^{-1} (X^T X)^{-1} C^{-1} C X^T y \\
&= C^{-1} (X^T X)^{-1} X^T y \\
\tilde{X} \hat{\tilde{\mathbf{w}}} &= X C C^{-1} (X^T X)^{-1} X^T y \\
&= X (X^T X)^{-1} X^T y \\
&= X \hat{\mathbf{w}} \text{ as required}
\end{aligned}
$$

4.

$$\begin{aligned}
\hat{\sigma^2} &= \operatorname*{argmax}_{\sigma^2} \sum_{i=1}^{N} \log p(y_i|\mathbf{x}_i;\mathbf{w},\sigma) \\
&= \operatorname*{argmax}_{\sigma^2} -\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i;\mathbf{w}))^2 - N \log \sigma\sqrt{2\pi} \\
&= \operatorname*{argmin}_{\sigma^2} \frac{1}{\sigma^2} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i;\mathbf{w}))^2 + N \log 2\pi\sigma^2 \\
&= \operatorname*{argmin}_{\sigma^2} \frac{1}{\sigma^2} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i;\mathbf{w}))^2 + N \log \sigma^2
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial\sigma^2} \frac{1}{\sigma^2} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i;\mathbf{w}))^2 + N \log \sigma^2 &= 0 \\
-\frac{1}{\sigma^4} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i;\mathbf{w}))^2 + \frac{N}{\sigma^2} &= 0 \\
\sum_{i=1}^{N} (y_i - f(\mathbf{x}_i;\mathbf{w}))^2 - N\sigma^2 &= 0
\end{aligned}$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i;\mathbf{w}))^2$$

$$\therefore\ \hat{\sigma^2} = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i;\mathbf{w}))^2$$

5.

$$
\begin{aligned}
\text{Let } \beta_{\hat{y},y} &= \begin{cases} 1 & \text{if } \hat{y} \leq y \\ \alpha & \text{otherwise} \end{cases} \\
l_\alpha(\hat{y}, y) &= \begin{cases} (\hat{y} - y)^2 & \text{if } \hat{y} \leq y \\ \alpha(\hat{y} - y)^2 & \text{otherwise} \end{cases} \\
&= \beta_{\hat{y},y}(\hat{y} - y)^2 \\
\frac{\partial}{\partial \mathbf{w}} L_\alpha &= \frac{\partial}{\partial \mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} l_\alpha(\hat{y}, y) \\
&= \frac{\partial}{\partial \mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \beta_{\hat{y},y}(\hat{y}_i - y_i)^2 \\
&= \frac{\partial}{\partial \mathbf{w}} \frac{1}{N} \sum_{i=1}^{N} \beta_{\hat{y},y}(\mathbf{w}^T \mathbf{x}_i - y_i)^2 \\
&= \frac{1}{N} \sum_{i=1}^{N} 2\beta_{\hat{y},y}(\mathbf{w}^T \mathbf{x}_i - y_i)\mathbf{x}_i
\end{aligned}
$$