Introduction to Statistical Machine Learning
Homework 2

Yota Toyama

November 2, 2016

1.

$$
\begin{aligned}
R(h_r; q) &= \int_{\mathbf{x}} \sum_{c=1}^{C} \sum_{c'=1}^{C} L_{0/1}(c', c) q(c_r = c' | \mathbf{x}) p(\mathbf{x}, y = c) d\mathbf{x} \\
&= \int_{\mathbf{x}} R(h_r | \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\
\text{where } R(h_r | \mathbf{x}) &= \sum_{c=1}^{C} \sum_{c'=1}^{C} L_{0/1}(c', c) q(c_r = c' | \mathbf{x}) p(y = c | \mathbf{x}) \\
&= \sum_{c=1}^{C} \sum_{c' \neq c}^{C} q(c_r = c' | \mathbf{x}) p(y = c | \mathbf{x}) \\
&= \sum_{c=1}^{C} (1 - q(c_r = c | \mathbf{x})) p(y = c | \mathbf{x}) \\
R(h^*) &= \int_{\mathbf{x}} R(h^* | \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\
\text{where } R(h^* | \mathbf{x}) &= \sum_{c=1}^{C} L_{0/1}(h^*(\mathbf{x}), c) p(y = c | \mathbf{x}) \\
&= \sum_{c \neq h^*}^{C} p(y = c | \mathbf{x}) \\
&= 1 - p(y = h^*(\mathbf{x}) | \mathbf{x})
\end{aligned}
$$

$$
\begin{aligned}
R(h_r | \mathbf{x}) - R(h^* | \mathbf{x}) &= \sum_{c=1}^{C} (1 - q(c_r = c | \mathbf{x})) p(y = c | \mathbf{x}) - (1 - p(y = h^*(\mathbf{x}) | \mathbf{x})) \\
&= p(y = h^*(\mathbf{x}) | \mathbf{x}) - \sum_{c=1}^{C} q(c_r = c | \mathbf{x}) p(y = c | \mathbf{x}) \\
&= \sum_{c=1}^{C} q(c_r = c | \mathbf{x}) (p(y = h^*(\mathbf{x}) | \mathbf{x}) - p(y = c | \mathbf{x})) \\
&\geq 0
\end{aligned}
$$

$$
\begin{aligned}
&\therefore\ R(h_r | \mathbf{x}) \geq R(h^* | \mathbf{x}) \\
&\therefore\ R(h_r; q) \geq R(h^*)
\end{aligned}
$$

2. Let M be the number of augmented data points.

$$\sum_{i=1}^{N+M}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 = \sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \lambda \left\|\mathbf{w}\right\|^2$$
$$\sum_{i=N+1}^{N+M}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 = \lambda \left\|\mathbf{w}\right\|^2$$

Let $y_i = 0$ and $\mathbf{x}_i = [0, a, \dots, a]^T$.

$$\sum_{i=N+1}^{N+M} a^2 \left\|\mathbf{w}\right\|^2 = \lambda \left\|\mathbf{w}\right\|^2$$
$$Ma^2 \left\|\mathbf{w}\right\|^2 = \lambda \left\|\mathbf{w}\right\|^2$$
$$Ma^2 = \lambda$$
$$\therefore\ \mathbf{y}' = \begin{bmatrix} y_1 \\ \vdots \\ y_N \\ 0 \\ \vdots \\ 0 \end{bmatrix}, X' = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nd} \\ 0 & a & \vdots & a \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a & \vdots & a \end{bmatrix}$$

s.t. $Ma^2 = \lambda$ (where M is the number of augmented data points)

3.

$$\forall i, j, \log \frac{p(c_i|\mathbf{x})}{p(c_j|\mathbf{x})} = \mathbf{w}_{ij} \cdot \mathbf{x}$$
$$\log p(c_i|\mathbf{x}) - \log p(c_j|\mathbf{x}) = \mathbf{w}_{ij} \cdot \mathbf{x}$$

Let $\log p(c_i|x) = \mathbf{w}_i \cdot \mathbf{x}$.

$$\mathbf{w}_i \cdot \mathbf{x} - \mathbf{w}_j \cdot \mathbf{x} = \mathbf{w}_{ij} \cdot \mathbf{x}$$
$$\mathbf{w}_{ij} = \mathbf{w}_i - \mathbf{w}_j$$

$$p(c_i|\mathbf{x}) = e^{\mathbf{w}_{ij}\cdot\mathbf{x}} p(c_j|\mathbf{x})$$
$$p(c_i|\mathbf{x}) = e^{\mathbf{w}_i\cdot\mathbf{x}} e^{-\mathbf{w}_j\cdot\mathbf{x}} p(c_j|\mathbf{x})$$
$$1 = \sum_{i=1}^{C} e^{\mathbf{w}_i\cdot\mathbf{x}} e^{-\mathbf{w}_j\cdot\mathbf{x}} p(c_j|\mathbf{x})$$
$$p(c_j|\mathbf{x}) = \frac{e^{\mathbf{w}_j\cdot\mathbf{x}}}{\sum_{i=1}^{C} e^{\mathbf{w}_i\cdot\mathbf{x}}}$$

$\therefore$ the softmax model corresponds to modeling the log-odds between any two classes.
If the number of classes equals 2,

$$\begin{aligned}
\frac{e^{\mathbf{w}_j \cdot \mathbf{x}}}{\sum_{i=1}^{C} e^{\mathbf{w}_i \cdot \mathbf{x}}} &= \frac{1}{\sum_{i=1}^{C} e^{(\mathbf{w}_i - \mathbf{w}_j) \cdot \mathbf{x}}} \\
&= \sigma((\mathbf{w}_i - \mathbf{w}_j) \cdot \mathbf{x}) \\
&= \sigma(\mathbf{v} \cdot \mathbf{x})
\end{aligned}$$

$\therefore$ In the binary case the softmax model is equivalent to the logistic regression model.

4.

$$\begin{aligned}
L(Y|X; W, \mathbf{b}) &\approx L(y|\mathbf{x}; W, \mathbf{b}) \\
&= -\log \hat{p}(y|\mathbf{x}; W, \mathbf{b}) + \lambda \left\| W \right\|^2 \\
&= -\log \frac{e^{W_y \cdot \mathbf{x} + \mathbf{b}_y}}{\sum_{c=1}^{C} e^{W_c \cdot \mathbf{x} + \mathbf{b}_c}} + \lambda \left\| W \right\|^2 \\
\frac{\partial}{\partial W_{ci}} L(y|\mathbf{x}; W, \mathbf{b}) &= -p(y = c)\mathbf{x}_i + \frac{\mathbf{x}_i e^{W_c \cdot \mathbf{x} + \mathbf{b}_c}}{\sum_{c=1}^{C} e^{W_c \cdot \mathbf{x} + \mathbf{b}_c}} + \lambda \left\| W \right\|^2 \\
&= (\frac{\mathbf{x}_i e^{W_c \cdot \mathbf{x} + \mathbf{b}_c}}{\sum_{c=1}^{C} e^{W_c \cdot \mathbf{x} + \mathbf{b}_c}} - p(y = c))\mathbf{x}_i + \lambda \left\| W \right\|^2 \\
\frac{\partial}{\partial \mathbf{b}_c} L(y|\mathbf{x}; W, \mathbf{b}) &= p(y = c) - \frac{\mathbf{x}_i e^{W_c \cdot \mathbf{x} + \mathbf{b}_c}}{\sum_{c=1}^{C} e^{W_c \cdot \mathbf{x} + \mathbf{b}_c}}
\end{aligned}$$