Introduction to Statistical Machine Learning
Homework 3

Yota Toyama

November 17, 2016

1. A decision tree can classify linearly separable data. A boundary made by such a tree looks like stairs approximating $\mathbf{w}^T\mathbf{x} + w_0 = 0$. And, in the worst case, its depth is $lceil \log \frac{N}{2} \rceil + 1$ because we can balance the tree along x_1.

2. A decision tree can classify data points which are not linearly separable by separating a space of $\mathbf{x}$ into $N - 1$ thin regions along x_1. And, in the worst case, its depth is $\lceil \log(N-1) \rceil$ balancing

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

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

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

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

$$\sum_{i=N+1}^{N+M} a^2 \|\mathbf{w}\|^2 = \lambda \|\mathbf{w}\|^2$$

$$Ma^2 \|\mathbf{w}\|^2 = \lambda \|\mathbf{w}\|^2$$

$$Ma^2 = \lambda$$

$$\therefore\ \mathbf{y}' = \begin{bmatrix} y_1 \\ \vdots \\ y_N \\ 0 \\ \vdots \\ 0 \end{bmatrix}, X' = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N1} & \cdots & x_{Nd} \\ 0 & a & \vdots & a \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a & \vdots & a \end{bmatrix}$$

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

4.

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

Let $\log p(c_i|x) = \mathbf{w}_i \cdot \mathbf{x}$. This doesn't break generality of the model above because $\mathbf{w}_{ij} = -\mathbf{w}_{ji}$ obviously and for all i and j we can pick any $\mathbf{w}_{ij}$ even if either $\mathbf{w}_i$ or $\mathbf{w}_j$ is fixed.

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

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

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

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

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

5.

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

$\therefore$ The update equasions are the below.

$$
\begin{aligned}
W_{ci} &\leftarrow W_{ci} - \eta\left(\left(\frac{e^{W_c\cdot\mathbf{x}+\mathbf{b}_c}}{\sum_{c=1}^{C} e^{W_c\cdot\mathbf{x}+\mathbf{b}_c}} - p(y=c)\right)\mathbf{x}_i + \lambda W_{ci}\right) \\
\mathbf{b}_c &\leftarrow \mathbf{b}_c - \eta\left(\frac{e^{W_c\cdot\mathbf{x}+\mathbf{b}_c}}{\sum_{c=1}^{C} e^{W_c\cdot\mathbf{x}+\mathbf{b}_c}} - p(y=c)\right)
\end{aligned}
$$

6. Please, see a Jupyter notebook file submitted together.