Junda An

Problem Set 4

PRML 5.[1,2,5-8,18]

1. EX 5.1 answer here.

   Let a_k^s to be activation for the neural network whose activation function in the hidden layer is sigmoid function. Denote parameters in this neural network as $w_{kj}^{(2s)}, w_{k0}^{(2s)}, w_{ji}^{(1s)}, w_{j0}^{(1s)}$. Let a_k^t to be activation for the neural network whose activation function in the hidden layer is tanh function. Denote parameters in this neural network as $w_{kj}^{(2t)}, w_{k0}^{(2t)}, w_{ji}^{(1t)}, w_{j0}^{(1t)}$.

   According to Ex 3.1, $tanh(a) = 2\sigma(2a) - 1$. Then,

$$
\begin{aligned}
a_k^t &= \sum_{j=1}^{M} w_{kj}^{(2t)} tanh(\sum_{i=1}^{D} w_{ji}^{(1t)} x_i + w_{j0}^{(1t)}) + w_{k0}^{(2t)} \\
&= \sum_{j=1}^{M} w_{kj}^{(2t)} (2\sigma(2(\sum_{i=1}^{D} w_{ji}^{(1t)} x_i + w_{j0}^{(1t)})) - 1) + w_{k0}^{(2t)} \ (substitute\ tanh(a) = 2\sigma(2a) - 1) \\
&= \sum_{j=1}^{M} 2w_{kj}^{(2t)} \sigma(2(\sum_{i=1}^{D} w_{ji}^{(1t)} x_i + w_{j0}^{(1t)})) - \sum_{j=1}^{M} w_{kj}^{(2t)} + w_{k0}^{(2t)} \ (simplify\ the\ function\ outside\ \sigma) \\
&= \sum_{j=1}^{M} 2w_{kj}^{(2t)} \sigma(\sum_{i=1}^{D} 2w_{ji}^{(1t)} x_i + 2w_{j0}^{(1t)})) - \sum_{j=1}^{M} w_{kj}^{(2t)} + w_{k0}^{(2t)} \ (simplify\ the\ \sigma\ function)
\end{aligned}
$$

   $a_k^s = \sum_{j=1}^{M} w_{kj}^{(2s)} \sigma(\sum_{i=1}^{D} w_{ji}^{(1s)} x_i + w_{j0}^{(1s)}) + w_{k0}^{(2s)}$

   To make two networks equivalent, $a_k^t = a_k^s$. Therefore, $\sum_{j=1}^{M} w_{kj}^{(2s)} \sigma(\sum_{i=1}^{D} w_{ji}^{(1s)} x_i + w_{j0}^{(1s)}) + w_{k0}^{(2s)} = \sum_{j=1}^{M} 2w_{kj}^{(2t)} \sigma(\sum_{i=1}^{D} 2w_{ji}^{(1t)} x_i + 2w_{j0}^{(1t)})) - \sum_{j=1}^{M} w_{kj}^{(2t)} + w_{k0}^{(2t)}$

   Then, through linear transformation, we make $w_{kj}^{(2s)} = 2w_{kj}^{(2t)}$, $w_{k0}^{(2s)} = -\sum_{j=1}^{M} w_{kj}^{(2t)} + w_{k0}^{(2t)}$, $w_{ji}^{(1s)} = 2w_{ji}^{(1t)}$, and $w_{j0}^{(1s)} = 2w_{j0}^{(1t)}$.

2. Ex 5.2 answer here

   The likelihood function under the conditional distribution for a multioutput neural network is $\prod_{n=1}^{N} \mathcal{N}(t_n|y(x_n, w), \beta^{-1} I)$. To maximize this function is equivalent to minimize the negative log of it, which is

$$
\begin{aligned}
-ln(\prod_{n=1}^{N}\mathcal{N}(t_n|y(x_n,w),\beta^{-1}I)) &= -\sum_{n=1}^{N} ln(\mathcal{N}(t_n|y(x_n,w))\ (basic\ log\ operation)\\
&= \frac{1}{2}\sum_{n=1}^{N}(t_n - y(x_n,w))^T(\beta I)(t_n - y(x_n,w)) - \frac{NK}{2}ln(\beta) + const\\
&(According\ to\ (5.13))\\
&(const\ is\ a\ value\ that\ is\ independent\ of\ w\ and\ \beta)\\
&= \frac{\beta}{2}\sum_{n=1}^{N}(t_n - y(x_n,w))^T I(t_n - y(x_n,w)) - \frac{NK}{2}ln(\beta) + const\\
&(take\ \beta\ out,\ since\ it\ is\ independent\ of\ n)\\
&= \frac{\beta}{2}\sum_{n=1}^{N}(t_n - y(x_n,w))^T(t_n - y(x_n,w)) - \frac{NK}{2}ln(\beta) + const\\
&(since\ I\ is\ identity\ matrix)\\
&= \frac{\beta}{2}\sum_{n=1}^{N}||t_n - y(x_n,w)^T||^2 - \frac{NK}{2}ln(\beta) + const\\
&(since\ w^Tw = ||w||^2)
\end{aligned}
$$

The first term of this function is (5.11). Therefore, to minimize the negative log of (5,16) is to minimize (5.11). Hence, maximize (5.16) is equivalent to minimize (5.11).

3. Ex 5.5 answer here.

   Taking the negative log of the likelihood is

$$
\begin{aligned}
E(w) &= -ln(\prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}(x_n,w)^{t_nk}[1-y_{nk}(x_n,w)^{1-t_nk}]) \ (plug\ in\ (5.22)) \\
&= -\sum_{n=1}^{N} ln(\prod_{k=1}^{K} y_{nk}(x_n,w)^{t_nk}[1-y_{nk}(x_n,w)]^{1-t_nk}) \ (basic\ log\ operation) \\
&= -\sum_{n=1}^{N}\sum_{k=1}^{K} y_{nk}ln((x_n,w)^{t_nk}[1-y_{nk}(x_n,w)]^{1-t_nk}) \ (basic\ log\ operation) \\
&= -\sum_{n=1}^{N}\sum_{k=1}^{K} ln(y_{nk}(x_n,w)^{t_nk}) + ln([1-y_{nk}(x_n,w)]^{1-t_nk}) \ (basic\ log\ operation) \\
&= -\sum_{n=1}^{N}\sum_{k=1}^{K} (t_nk)ln(y_{nk}(x_n,w)) + (1-t_nk)ln([1-y_{nk}(x_n,w)]) \ (basic\ log\ operation) \\
&= -\sum_{n=1}^{N}\sum_{k=1}^{K} (t_nk)ln(y_{nk}) + (1-t_nk)ln((1-y_{nk})) \ (Let\ y_{nk} = y_{nk}(x_n,w)) \\
&= -\sum_{n=1}^{N}\sum_{k=1}^{K} (t_nk)ln(y_{nk}) \\
&\ (since\ t_{nk} \in \{0,1\}\ and\ y_{nk} = p(t_{nk}=1|x),\ I\ get\ rid\ of\ the\ last\ term)
\end{aligned}
$$

Therefore, maximizing likelihood for a multiclass neural network model in which the network outputs have the interpretation $y_k(x,w) = p(t_k = 1|x)$ is equivalent to the minimization of the corss-entropy error function (5.24).

4. Ex 5.6 answer here

   Since $y_n = \sigma(a_n) = \frac{1}{1+e^{-a_n}}$, $\frac{\partial y_n}{\partial a_n} = \frac{0-1(-e^{a_n})}{(1+e^{-a_n})^2} = \frac{e^{a_n}}{(1+e^{-a_n})^2} = \frac{1}{1+e^{a_n}}\frac{e^{a_n}}{1+e^{a_n}} = y_n(1-y_n)$

$$\begin{aligned}
\frac{\partial E(w)}{\partial a_n} &= \frac{\partial - (t_k ln(y_k) + (1-t_k) ln(1-y_k))}{\partial a_k} \\
&(plug\ in E(w) = -\sum_{n=1}^{N}(t_n ln(y_n) + (1-t_n) ln(1-y_n))) \\
&((t_k ln(y_k) + (1-t_k) ln(1-y_k)\ is\ the\ only\ term\ that\ depends\ on\ a_k) \\
&= -\frac{t_k ln(y_k) + (1-t_k) ln(1-y_k)}{\partial a_k}\ (take\ summation\ out) \\
&= -(\frac{t_k}{y_k}\frac{\partial y_k}{\partial a_k} - \frac{1-t_k}{1-y_k}\frac{\partial y_k}{\partial a_k})\ (chain\ rule) \\
&= -(\frac{\partial y_k}{\partial a_k}\frac{t_k(1-y_k) - (1-t_k)y_k}{y_k(1-y_k)}) \\
&= -(\frac{\partial y_k}{\partial a_k}\frac{t_k - y_k}{y_k(1-y_k)}) \\
&= -(y_k(1-y_k)\frac{t_k - y_k}{y_k(1-y_k)})\ (plug\ in\ \frac{\partial y_k}{\partial a_k} = y_k(1-y_k)) \\
&= -(t_k - y_k) \\
&= y_k - t_k
\end{aligned}$$

5. Ex 5.7 answer here

   According to previous HW (4.17), $\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j)$.

$$\frac{\partial E(w)}{\partial a_j} = \frac{\partial - \sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} ln(y_k(x_n,w))}{\partial a_j} \ (plug\ in\ E(w) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} ln(y_k)(x_n,w))$$
$$= -\sum_{n=1}^{N}\sum_{k=1}^{K} \frac{\partial t_{kn} ln(y_k(x_n,w))}{\partial a_j} \ (take\ summations\ out)$$
$$= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn}\frac{1}{y_k(x_n,w)}\frac{\partial y_k(x_n,w)}{\partial a_j} \ (chain\ rule)$$
$$= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn}\frac{1}{y_{kn}}\frac{\partial y_{kn}}{\partial a_j} \ (denote\ y_k(x_n,w)\ as\ y_{kn})$$
$$= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn}\frac{1}{y_{kn}} y_{kn}(I_{kj} - y_{jn}) \ (plug\ in \frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j))$$
$$= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} I_{kj} - t_{kn} y_{jn}$$
$$= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} I_{kj} + \sum_{n=1}^{N}\sum_{k=1}^{K} t_{kn} y_{jn}$$
$$= -\sum_{n=1}^{N} t_{jn} + \sum_{n=1}^{N} y_{jn} \ (I_{kj} = 1,\ only\ when\ k = j.\ Otherwise,\ \ I_{kj} = 0. \sum_{k=1}^{K} t_{kn} = 1)$$
$$= \sum_{n=1}^{N} (y_{jn} - t_{jn})$$

6. Ex 5.8 answer here

   According to (5.59), $tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}$.

   Then,

$$d\frac{tanh(a)}{da} = \frac{d\frac{e^a - e^{-a}}{e^a + e^{-a}}}{da} \ (plug\ intanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}})$$
$$= \frac{(e^a + e^{-a})(e^a + e^{-a}) - (e^a - e^{-a})(e^a - e^{-a})}{(e^a + e^{-a})^2} \ (division\ rule)$$
$$= \frac{(e^a + e^{-a})^2}{(e^a + e^{-a})^2} - \frac{(e^a - e^{-a})^2}{(e^a + e^{-a})^2} \qquad = 1 - (\frac{e^a - e^{-a}}{e^a + e^{-a}})^2$$
$$= 1 - tanh(a)^2 \ (plug\ in\ tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}})$$

7. Ex 5.18

Problem Set 4

Suppose the extra parameters corresponding to skip-layer connections that go directly from the inputs to the outputs is given by the matrix s_{ij} . Introducing skip layer weights s_{ij} into the two-layer network of the form shown in Fig. 5.1 would only affect the forward propagation equation Eq. (5.64):

$$y_k = \sum_{j=1}^{M} w_{kj}^{(2)} z_j + \sum_{i=1}^{D} s_{ki} x_i$$

Then, $\frac{\partial y_k}{\partial s_{ki}} = \frac{\partial \sum_{j=1}^{M} w_{kj}^{(2)} z_j + \sum_{i=1}^{D} s_{ki} x_i}{\partial s_{ki}} = x_i$

According to Eq. (5.61), $E_n = \frac{1}{2} \sum_{k=1}^{K} (y_k - t_k)^2$. Therefore, $\frac{\partial E_n}{\partial y_k} = \frac{1}{2} 2(y_k - t_k)$ $(chain\ rule\ and\ y_k = a_k\ due\ to\ linear\ activation\ function)$

$= y_k - t_k = \delta_k$ $(definition\ of\ \delta_k\ in\ Eq.(5.65))$.

$$\begin{aligned}\frac{\partial E_n}{\partial s_{ki}} &= \frac{\partial E_n}{\partial y_k} \frac{\partial y_k}{\partial s_{ki}}\ (chain\ rule) \\ &= \delta_k x_i\ (substitute\ \frac{\partial y_k}{\partial s_{ki}} = x_i,\ and\ \frac{\partial E_n}{\partial y_k} = \delta_k)\end{aligned}$$