Junda An

CSCI 420
March 19, 2019

Homework 1

AML EX:3.[1, 7-9]

Answer to the problem goes here.

1. EX 3.1 answer here. If the data is not linearly separable, PLA cannot find a weight, w such that $E_{in}(w) = 0$, which is the only condition that stops PLA. Therefore, if we do not put some restriction on PLA, like setting the maximum number of iterations, PLA will never stop.

2. Ex 3.7 answer here. From (3.9), we get $E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N} ln^{1+e^{-y_n w^T x_n}}$

   Therefore,

$$
\begin{aligned}
\nabla E_{in}(w) &= \frac{1}{N}\sum_{n=1}^{N} \nabla(ln(1+e^{-y_n w^T x_n}))\ (take\ gradient\ of\ each\ term\ and\ combine\ them\ together) \\
&= \frac{1}{N}\sum_{n=1}^{N} \frac{1}{1+e^{y_n w^T x_n}}(-y_n x_n)\ (w\ is\ the\ only\ variable\ and\ use\ chain\ rule) \\
&= -\frac{1}{N}\sum_{n=1}^{N} \frac{y_n x_n}{1+e^{y_n w^T x_n}} \\
&= \frac{1}{N}\sum_{n=1}^{N} -y_n x_n \frac{1\cdot e^{-y_n w^T x_n}}{(1+e^{y_n w^T x_n})\cdot e^{-y_n w^T x_n}} \\
&(simultaneously\ multiply\ dividend\ and\ divisor\ by\ e^{-y_n w^T x_n}) \\
&= \frac{1}{N}\sum_{n=1}^{N} -y_n x_n \frac{e^{-y_n w^T x_n}}{(1+e^{-y_n w^T x_n})} \\
&= \frac{1}{N}\sum_{n=1}^{N} -y_n x_n \theta(-y_n w^T x_n)\ (definition\ of\ the\ logistic\ function)
\end{aligned}
$$

   Suppose (x_i, y_i) is a misclassified point and (x_j, y_j) is a correctly classified point. Then, $sign(w^T x_i) \neq y_i$. $sign(w^T x_j) = y_j$. Therefore, $-y_i w^T x_i > 0$. $-y_j w^T x_j < 0$. Then $\theta(-y_i w^T x_i) > \theta(0) > 0.5 > \theta(-y_j w^T x_j)$ (since the logistic function is monotonically increasing). In the equation of the gradient I derived above, the coefficient of $-y_n x_n$ is $\theta(-y_n w^T x_n)$, which is greater if the point is incorrectly classified. Therefore, a misclassified point contributes more to the gradient than a correctly classified one.

3. Ex 3.8 answer here.

According to the equation in the book,

$$\begin{aligned}\triangle E_{in} &= \eta\nabla E_{in}(w(0)+\eta\hat{v}) - E_{in}(w(0))\\ &= \eta\nabla E_{in}(w(0))^T\hat{v} + O(\eta^2)\ (Tylor\ expansion\ to\ the\ first\ order)\\ &\geq -\eta||\nabla E_{in}(w(0))||\end{aligned}$$

If η is not small, $O(\eta^2)$ cannot be neglected. It has some effect on $\triangle E_{in}$. Therefore, in this case, $\hat{v} = -\frac{\nabla E_{in}(w(0))}{||\nabla E_{in}(w(0))||}$ cannot give largest decrease in E_{in}

4. Ex 3.9 answer here (a) Please see the jupyter notebook, Ex3.9.ipynb.

   (b) If $y = sign(s)$, $e_{class} = 0$ and $e_{sq} = (y-s)^2 \geq 0 = e_{class}$

   IF $y \neq sign(s)$, $e_{class}(s,y) = 1$. There are two cases for $e_{sq}(s,y)$. $y \in \{-1,1\}$. If $y=-1$, since $y \neq sign(s)$, $s > 0$. Then, $e_{sq}(s,y) = (-1-s)^2 \geq 1 = e_{class}(s,y)$. If $y = 1$, $s < 0$. Then, $e_{sq}(s,y) = (1-s)^2 \geq 1 = e_{class}(s,y)$. Therefore, $e_{class}(s,y) \leq e_{sq}(s,y)$

   (c) If $y = sign(s)$, $e_{class}(s,y) = 0$ and $ys > 0$. Then $\frac{1}{ln^2}e_{log}(s,y) = \frac{1}{ln^2}ln^{1+e^{-ys}} = log_2^{1+e^{-ys}}$ $(operation\ of\ log)$ $\geq log_2(1)$ $(since\ e^{-ys} \geq 0) = 0 = e_{class}(s,y)$

   If $y \neq sign(s)$, $e_{class}(s,y) = 1$ and $ys < 0 \Rightarrow -ys \geq 0 \Rightarrow e^{-ys}\ geq1$. Then, $\frac{1}{ln^2}e_{log}(s,y) = \frac{1}{ln^2}ln^{1+e^{-ys}} = log_2^{1+e^{-ys}} \geq log_2^{1+1} = log_2^2 = 1 = e_{class}(s,y)$.

PRML 2.[43]; 3.[1,2]; 4.[1,7,12,13,14,17,18]

1. EX 2.43 answer here.

$$\begin{aligned}\int_{-\infty}^{\infty} p(x|\sigma^2, q)dx &= \int_{-\infty}^{\infty}\frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)}e^{-\frac{|x|^q}{2\sigma^2}}dx\\ &= \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)}\int_{-\infty}^{\infty}e^{-\frac{|x|^q}{2\sigma^2}}dx\ (since\ the\ first\ part\ is\ independent\ of\ x)\end{aligned}$$

$$\begin{aligned}
\int_{-\infty}^{\infty} e^{-\frac{|x|^q}{2\sigma^2}} dx &= \int_{0}^{\infty} e^{-\frac{|x|^q}{2\sigma^2}} dx + \int_{-\infty}^{0} e^{-\frac{|x|^q}{2\sigma^2}} dx \\
&= \int_{0}^{\infty} e^{-\frac{x^q}{2\sigma^2}} dx + \int_{-\infty}^{0} e^{-\frac{(-x)^q}{2\sigma^2}} dx \\
&= 2\int_{0}^{\infty} e^{-\frac{x^q}{2\sigma^2}} dx \ (since\ e^{-\frac{x^q}{2\sigma^2}} = e^{-\frac{(-y)^q}{2\sigma^2}}\ when\ y = -x) \\
&= 2\int_{0}^{\infty} e^{-u} d((2u\sigma^2)^{1/q}) \ (let\ u = \frac{x^q}{2\sigma^2}.\ Then\ x = (2u\sigma^2)^{1/q}) \\
&= 2\int_{0}^{\infty} \frac{1}{q}(2u\sigma^2)^{1/q-1} 2\sigma^2 e^{-u} du \ (chain\ rule) \\
&= 2\frac{1}{q}(2\sigma^2)^{1/q-1}2\sigma^2 \int_{0}^{\infty} u^{1/q-1} e^{-u} du \ (factor\ out\ the\ part\ that\ is\ independent\ of\ u) \\
&= \frac{2}{q}(2\sigma^2)^{1/q}\Gamma(1/q) \ (since\ \Gamma(z) = \int_{0}^{\infty} x^{z-1}e^{-x}dx)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\int_{-\infty}^{\infty} p(x|\sigma^2, q)dx &= \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)} \int_{-\infty}^{\infty} e^{-\frac{|x|^q}{2\sigma^2}} dx \\
&= \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)} \frac{2}{q}(2\sigma^2)^{1/q}\Gamma(1/q) \\
&= 1
\end{aligned}$$

When $q = 2$, $p(x|\sigma^2, q) = \frac{2}{2(2\sigma^2)^{1/2}\Gamma(1/2)} e^{-\frac{x^2}{2\sigma^2}} = \frac{1}{(2\sigma^2)^{1/2}\pi^{1/2}} e^{-\frac{x^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{x^2}{2\sigma^2}}$

Therefore, the distribution is Gaussian with mean equal to 0 and standard deviation equal to σ.

$t = y(x, w) + \epsilon \Rightarrow \epsilon = t - y(x, w)$. Since ϵ is a random noise variable drawn from the distribution $p(x|\sigma^2, q)$, $p(t|X, w, \sigma^2) = \prod_{n=1}^{N} p(y(x_n, w) - t_n|\sigma^2, q) = \prod_{n=1}^{N} \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)} e^{-\frac{|t-y(x_n,w)|^q}{2\sigma^2}}$.

Therefore,

$$
\begin{aligned}
ln^{p(t|X,w,\sigma^2)} &= ln^{\prod_{n=1}^{N} \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)} e^{-\frac{|t_n - y(x_n,w)|^q}{2\sigma^2}}} \quad (substitute\ p(t|X,w,\sigma^2)) \\
&= \sum_{n=1}^{N} (ln^{e^{-\frac{|t_n - y(x_n,w)|^q}{2\sigma^2}}} + ln^{\frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)}}) \\
&= \sum_{n=1}^{N} (ln^{e^{-\frac{|t_n - y(x_n,w)|^q}{2\sigma^2}}}) + N \cdot ln^{\frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)}} \ (since\ ln^{\frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)}}\ is\ independent\ of\ n) \\
&= -\frac{1}{2\sigma^2} \sum_{n=1}^{N} |t_n - y(x_n,w)|^q + N(ln^{(2\sigma^2)^{-1/q}\frac{q}{2\Gamma(1/q)}}) \\
&= -\frac{1}{2\sigma^2} \sum_{n=1}^{N} |y(x_n,w) - t_n|^q + N(ln^{(2\sigma^2)^{-1/q}} + ln^{\frac{q}{2\Gamma(1/q)}}) \\
&= -\frac{1}{2\sigma^2} \sum_{n=1}^{N} |y(x_n,w) - t_n|^q + N \cdot ln^{(2\sigma^2)^{-1/q}} + N \cdot ln^{\frac{q}{2\Gamma(1/q)}} \\
&= -\frac{1}{2\sigma^2} \sum_{n=1}^{N} |y(x_n,w) - t_n|^q - \frac{N}{q} ln^{(2\sigma^2)} + const
\end{aligned}
$$

2. Ex 3.1 answer here

   According to (3.6),

$$
\begin{aligned}
2\sigma(2a) - 1 &= 2\frac{1}{1+e^{-2a}} - 1 \\
&= \frac{2 - 1 - e^{-2a}}{1+e^{-2a}} \\
&= \frac{1 - e^{-2a}}{1+e^{-2a}} \\
&= \frac{e^a}{e^a} \cdot \frac{1 - e^{-2a}}{1+e^{-2a}} \\
&(since\ \frac{e^a}{e^a} = 1,\ multiplying\ the\ function\ by\ it\ does\ not\ change\ the\ value) \\
&= \frac{e^a(1 - e^{-2a})}{e^a(1+e^{-2a})} \\
&= \frac{e^a - e^{-a}}{e^a + e^{-a}} \\
&= tanh(a)
\end{aligned}
$$

Then, let $a_j = \frac{x-\mu_j}{2s}$.

$$
\begin{aligned}
y(x,w) &= w_0 + \sum_{j=1}^{M} w_j \sigma(\frac{x-\mu_j}{s}) \\
&= w_0 + \sum_{j=1}^{M} w_j \sigma(2a_j) \ (substitute\ a_j\ into\ the\ function) \\
&= w_0 + \sum_{j=1}^{M} \frac{w_j}{2}(2\sigma(2a_j) - 1 + 1) \\
&(since\ -1+1=0\ and\ \frac{1}{2}\cdot 2 = 1,\ this\ operation\ does\ not\ change\ the\ function\ value) \\
&= w_0 + \sum_{j=1}^{M} \frac{w_j}{2}(tanh(a_j)+1) \ (substitute\ tanh(a_j)\ into\ the\ function) \\
&= w_0 + \sum_{j=1}^{M} \frac{w_j}{2} tanh(a_j) + \frac{w_j}{2} \\
&= w_0 + \sum_{j=1}^{M} \frac{w_j}{2} + \frac{w_j}{2} tanh(a_j) \\
&= w_0 + \sum_{j=1}^{M} \frac{w_j}{2} + \frac{w_j}{2} tanh(\frac{x-\mu_j}{2s}) \ (substitute\ a_j = \frac{x-\mu_j}{2s}\ back) \\
&= u_0 + \sum_{j=1}^{M} u_j tanh(\frac{x-\mu_j}{2s}) \ (let\ u_0 = w_0 + \sum_{j=1}^{M} \frac{w_j}{2}\ and\ u_j = \frac{w_j}{2})
\end{aligned}
$$

3. Ex 3.2 answer here.

   Let S be the subspace that is spanned by the columns of Φ. Then, any vector v can be decomposed into two vectors relative to S. One is orthogonal to S and the other is in S. Therefore, let $v = x + y$, where $x \in S$ and y is in the space that is orthogonal to S. Then, $x = \Phi z$, since x can be expressed as some linear combination of basis in S, which are columns of Φ. $\Phi y = 0$ and $\Phi^T y = 0$, since y are orthogonal to S.

Then,

$$
\begin{aligned}
\Phi(\Phi^T\Phi)^{-1}\Phi^T v &= \Phi(\Phi^T\Phi)^{-1}\Phi^T(x+y) \ (substitute\ v = x+y) \\
&= \Phi(\Phi^T\Phi)^{-1}\Phi^T x + \Phi(\Phi^T\Phi)^{-1}\Phi^T y \\
&= \Phi(\Phi^T\Phi)^{-1}\Phi^T\Phi z + 0 \ (since\ \Phi^T y = 0) \\
&= \Phi((\Phi^T\Phi)^{-1}(\Phi^T\Phi))z \\
&= \Phi I z \ (since\ the\ product\ of\ a\ matrix\ and\ its\ inverse\ is\ identity\ matrix) \\
&= \Phi z
\end{aligned}
$$

Φz is the linear combination of vectors in S. Therefore, the least-squares solution $y = \Phi(\Phi^T\Phi)^{-1}\Phi^T t$ projects t onto the manifold S which is spanned by columns of Φ. This is shown in Figure 3.2.

4. Ex 4.1 answer here

Proof. Consider two data point sets $\{x_n\}$ and $\{y_n\}$.

$\Rightarrow$: Suppose their convex hulls intersect. This means there exists a point z such that $z = \sum_n \alpha_n x_n = \sum_n \beta_n y_n$, where $\alpha_n, \beta_n \geq 0$ and $\sum_n \alpha_n = \sum_n \beta_n = 1$. Then, assume, by contradiction, that $\{x_n\}$ and $\{y_n\}$ are linearly separable. Then, there exists a vector $\hat{w}$ and a scalar w_0 such that $\hat{w}^T x_n + w_0 > 0$ for all x_n, and $\hat{w}^T y_n + w_0 > 0$ for all y_n.

Then,

$$
\begin{aligned}
\hat{w}^T z + w_0 &= \hat{w}^T(\sum_n \alpha_n x_n) + w_0 \\
&= \hat{w}^T(\sum_n \alpha_n x_n) + (\sum_n \alpha_n) w_0 \ (since\ \sum_n \alpha_n = 1\ by\ definition) \\
&= \sum_n \alpha_n(\hat{w}^T x_n) + \sum_n \alpha_n w_0 \ (since\ w_0\ and\ \hat{w}\ are\ independent\ of\ n) \\
&= \sum_n \alpha_n(\hat{w}^T x_n + w_0) > 0 \ (since\ \alpha_n > 0\ and\ \hat{w}^T x_n + w_0 > 0)
\end{aligned}
$$

However, on the other hand,

$$
\begin{aligned}
\hat{w}^T z + w_0 &= \hat{w}^T(\sum_n \beta_n x_n) + w_0 \\
&= \hat{w}^T(\sum_n \beta_n x_n) + (\sum_n \beta_n) w_0 \ (since\ \sum_n \beta_n = 1\ by\ definition) \\
&= \sum_n \beta_n(\hat{w}^T x_n) + \sum_n \beta_n w_0 \ (since\ w_0\ and\ \hat{w}\ are\ independent\ of\ n) \\
&= \sum_n \beta_n(\hat{w}^T x_n + w_0) < 0 \ (since\ \beta_n > 0\ and\ \hat{w}^T x_n + w_0 < 0)
\end{aligned}
$$

This generates a contradiction. Therefore, $\{x_n\}$ and $\{y_n\}$ are not linearly separable.

$\Leftarrow$: Let $\{x_n\}$ and $\{y_n\}$ be two linearly separable sets. By definition, there exists a vector $\hat{w}$ and a scalar w_0 such that $\hat{w}^T x_n + w_0 > 0$ for all x_n, and $\hat{w}^T y_n + w_0 > 0$ for all y_n. Assume, by contradiction, that their convex hulls intersect. Then, there exists a point z such that $z = \sum_n \alpha_n x_n = \sum_n \beta_n y_n$, where $\alpha_n, \beta_n \geq 0$ and $\sum_n \alpha_n = \sum_n \beta_n = 1$. We can follow the steps in the forward direction prove and get the contradiction that $\hat{w}^T x_n + w_0 < 0$ and $\hat{w}^T x_n + w_0 > 0$. Therefore, their convex hulls do not intersect. $\square$

5. Ex 4.7 answer here

$$\begin{aligned}
1 - \sigma(a) &= 1 - \frac{1}{1+e^{-a}} \; (according\ to\ (4.59)) \\
&= \frac{1+e^{-a}-1}{1+e^{-a}} \\
&= \frac{e^{-a}}{1+e^{-a}} \\
&= \frac{e^{-a}e^{a}}{(1+e^{-a})e^{a}} \\
&(multiplying\ divisor\ and\ dividend\ simultaneously\ by\ e^{a}\ doesn't\ change\ function\ value) \\
&= \frac{1}{1+e^{a}} \\
&= \sigma(-a) \; (according\ to\ (4.59))
\end{aligned}$$

Let $y = \sigma(a) = \frac{1}{1+e^{-a}}$. Then, $1 + e^{-a} = \frac{1}{y} \Rightarrow e^{-a} = \frac{1}{y} - 1 \Rightarrow -a = ln^{\frac{1}{y}-1} \Rightarrow a = -ln^{\frac{1-y}{y}} \Rightarrow a = ln^{(\frac{1-y}{y})^{-1}} = ln^{\frac{y}{1-y}} = \sigma^{-1}(y)$

6. Ex 4.12 answer here

$$\begin{aligned}
\frac{d\sigma}{da} &= \frac{d(\frac{1}{1+e^{-a}})}{da} \ (according\ to\ (4.59)) \\
&= \frac{1 \cdot e^{-a} - 0 \cdot (1+e^{-a})}{(1+e^{-a})^2} \ (division\ rule) \\
&= \frac{e^{-a}}{(1+e^{-a})^2} \\
&= \frac{1}{1+e^{-a}} \frac{e^{-a}}{1+e^{-a}} \\
&= \sigma(a) \frac{e^{-a}}{1+e^{-a}} \ (according\ to\ (4.59)) \\
&= \sigma(a)(\frac{1+e^{-a}}{1+e^{-a}} - \frac{1}{1+e^{-a}}) \\
&= \sigma(a)(1-\sigma(a)) \ (according\ to\ (4.59))
\end{aligned}$$

7. Ex 4.13 answer here

Let $z_n = t_n ln^{y_n} + (1-t_n) ln^{1-y_n}$

$$\begin{aligned}
\frac{\partial z_n}{\partial y_n} &= \frac{\partial(t_n ln^{y_n} + (1-t_n) ln^{1-y_n})}{\partial y_n} \ (substitute\ z_n) \\
&= \frac{t_n}{y_n} - \frac{1-t_n}{1-y_n} \ (chain\ rule) \\
&= \frac{t_n(1-y_n) - y_n(1-t_n)}{y_n(1-y_n)} \\
&= \frac{t_n - t_n y_n - y_n + t_n y_n}{y_n(1-y_n)} \\
&= \frac{t_n - y_n}{y_n(1-y_n)}
\end{aligned}$$

$\frac{\partial y_n}{\partial a_n} = \sigma(a_n)(1-\sigma(a_n))$ (according to (4.88), since $y_n = \sigma(a_n)$).

$\frac{\partial a_n}{\partial w} = \frac{\partial(w^T \phi_n)}{\partial w} = \phi_n$

Therefore,

$$\begin{aligned}
\nabla E(w) &= \nabla(-\sum_{n=1}^{N} t_n ln^{y_n} + (1-t_n) ln^{1-y_n}) \\
&= -\sum_{n=1}^{N} \nabla(t_n ln^{y_n} + (1-t_n) ln^{1-y_n}) \\
&= -\sum_{n=1}^{N} (\frac{\partial z_n}{\partial y_n} \frac{\partial y_n}{\partial a_n} \frac{\partial a_n}{\partial w}) \\
&= -\sum_{n=1}^{N} (\frac{t_n - y_n}{y_n(1-y_n)} \sigma(a_n)(1-\sigma(a_n))\phi_n) \ (substitute\ all\ equations\ above) \\
&= -\sum_{n=1}^{N} (\frac{t_n - y_n}{y_n(1-y_n)} y_n(1-y_n)\phi_n) \ (since\ y_n = \sigma(a_n)) \\
&= -\sum_{n=1}^{N} (t_n - y_n)\phi_n \\
&= \sum_{n=1}^{N} (y_n - t_n)\phi_n
\end{aligned}$$

8. Ex 4.14 answer here The likelihood function for the logistic regression model is given by (4.89), which is $p(t|w) = \prod_{n=1}^{N} y_n^{t_n}(1-y_n)^{1-t_n}$. To get the maximum likelihood solution, we want to minimize the error function which is the negative of natural log of the likelihood function. $E(w) = -ln^{p(t|w)}$. From Ex 4.13, we get the gradient of the error function with respect to w, which is $\nabla E(w) = \sum_{n=1}^{N}(y_n - t_n)\phi_n$. Therefore, the gradient is optimized when $y_n = t_n$.

   Since the data set is linearly separable, we can find a w such that $w^T\phi(x_m) > 0$ for $x_m \in C_1$ and $w^T\phi(x_l) < 0$ for $x_l \in C_2$. Therefore, $w^T\phi(x) = 0$ is the decision boundary, which separates two classes. $y_n = t_n$ happens when the decision boundary can classify all points correctly. Therefore, $p(C_1|\phi) = 1 \Rightarrow \sigma(w^T\phi(x_m)) = 1$. Due to the property of σ, $w \to \infty$ makes $p(C_1|\phi(x_m)) \to 1$. Similarly, $w \to -\infty \Rightarrow \sigma(w^T\phi(x_l) \to 0 \Rightarrow p(C_2|\phi(x_l)) = 1 - \sigma(w^T\phi(x_l)) \to 1$. Therefore, finding a vector w whose decision boundary $w^T\phi(x) = 0$ that separates the classes and taking the magnitude of w to infinity gets the maximum likelihood solution for the logistic regression.

9. Ex 4.17 answer here

When $j = k$,

$$
\begin{aligned}
\frac{\partial y_k}{\partial a_j} &= \frac{\partial y_k}{\partial a_k} \\
&= \frac{\partial \frac{e^{a_k}}{\sum_i e^{a_i}}}{\partial a_k} \ (\textit{substitute } y_k) \\
&= \frac{e^{a_k}(\sum_i e^{a_i}) - e^{a_k} e^{a_k}}{(\sum_i e^{a_i})^2} \ (\textit{division rule}) \\
&(e^{a_k} \textit{ is the only term in } \sum_i e^{a_i} \textit{ that is dependent on } a_k) \\
&= \frac{e^{a_k}(\sum_i e^{a_i} - e^{a_k})}{(\sum_i e^{a_i})^2} \\
&= \frac{e^{a_k}}{\sum_i e^{a_i}} \frac{\sum_i e^{a_i} - e^{a_k}}{\sum_i e^{a_i}} \\
&= y_k(1 - y_k) \ (\textit{substitute } y_k \textit{ back})
\end{aligned}
$$

When $j \neq k$,

$$
\begin{aligned}
\frac{\partial y_k}{\partial a_j} &= \frac{\partial \frac{e^{a_k}}{\sum_i e^{a_i}}}{\partial a_j} \ (\textit{substitute } y_k) \\
&= \frac{0 \cdot \sum_i e^{a_i} - e^{a_k} e^{a_j}}{(\sum_i e^{a_i})^2} \ (\textit{division rule}) \\
&(\textit{since } e^{a_j} \textit{ is the only term in } \sum_i e^{a_i} \textit{ that is dependent on } a_j \textit{ and } e^{a_k} \textit{ is independent on } a_j) \\
&= -\frac{e^{a_k} e^{a_j}}{(\sum_i e^{a_i})^2} \\
&= -\frac{e^{a_k}}{\sum_i e^{a_i}} \frac{e^{a_j}}{\sum_i e^{a_i}} \\
&= -y_k y_j \ (\textit{substitute } y_k \textit{ and } y_j \textit{ back})
\end{aligned}
$$

Since $I_{kj} = 0$, when $k \neq j$ and $I_{kj} = 1$, when $k = j$, where I is an identity matrix, we can get $\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j)$ (combining two equations we got in two different scenarios).

10. Ex 4.18 answer here

    From Ex 4.17, we get that $\frac{\partial y_k}{\partial a_j} = y_k(I_{kj} - y_j)$. Therefore, $\frac{\partial y_{nk}}{\partial a_j} = y_{nk}(I_{kj} - y_{nj})$

Therefore,

$$
\begin{aligned}
\nabla_{w_j} E(w_1, ..., w_K) &= \frac{\partial - \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} ln^{y_{nk}}}{\partial w_j} \ (4.108) \\
&= -\sum_{n=1}^{N} \frac{\partial \sum_{k=1}^{K} t_{nk} ln^{y_{nk}}}{\partial w_j} \ (take\ summation\ out\ since\ n\ is\ independent\ of\ w_j) \\
&= -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \frac{\partial ln^{y_{nk}}}{\partial a_j} \frac{\partial a_j}{\partial w_j} \ (take\ summation\ and\ t_{nk}\ out\ for\ the\ same\ reason) \\
&= -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \frac{1}{y_{nk}} \frac{\partial y_{nk}}{\partial a_j} \frac{\partial a_j}{\partial w_j} \ (chain\ rule) \\
&= -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \frac{1}{y_{nk}} y_{nk} (I_{kj} - y_{nj}) \phi_n (substitute\ \frac{\partial y_{nk}}{\partial a_j}\ and\ \frac{\partial a_j}{\partial w_j}) \\
&= -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} (I_{kj} - y_{nj}) \phi_n \\
&= \sum_{n=1}^{N} (\sum_{k=1}^{K} t_{nk} y_{nj} \phi_n - \sum_{k=1}^{K} t_{nk} I_{kj} \phi_n) \\
&= \sum_{n=1}^{N} (y_{nj} \phi_n - t_{nj} \phi_n) \ (since\ \sum_{k=1}^{K} t_{nk} = 1\ and\ I_{kj} = 0\ when\ k \neq j) \\
&= \sum_{n=1}^{N} (y_{nj} - t_{nj}) \phi_n
\end{aligned}
$$