1: AML Pg. 79, Ex. 3.1

Answer:

PLA will never stop updating if the data is not linearly separable because PLA only stops on the condition that there are no misclassified datapoints. If, however, the data is not linearly separable, any PLA-generated decision boundary, by being linear, will result in at least one misclassified datapoint.

2: AML Pg. 99, Ex. 3.7

Answer: a) Showing

$$-\frac{1}{N}\sum_{n=1}^{N}\frac{y_n\boldsymbol{x_n}}{1+\exp\left(y_n\boldsymbol{w^T x_n}\right)} = \frac{1}{N}\sum_{n=1}^{N} -y_n\boldsymbol{x_n}\theta(-y_n\boldsymbol{w^T x_n})$$

is the same as showing

$$\frac{1}{1+\exp\left(y_n\boldsymbol{w^T}\right)\boldsymbol{x_n}} = \theta(-y_n\boldsymbol{w^T x_n}) = \frac{\exp\left(-y_n\boldsymbol{w^T x_n}\right)}{1+\exp\left(-y_n\boldsymbol{w^T x_n}\right)}$$

Define $\alpha = y_n\boldsymbol{w^T x_n}$

$$\frac{1}{1+\exp\alpha} * \frac{\exp -\alpha}{\exp -\alpha} = \frac{\exp -\alpha}{1+\exp -\alpha}$$

Substituting in α...

$$\frac{1}{1+\exp\left(y_n\boldsymbol{w^T x_n}\right)} = \frac{\exp\left(-y_n\boldsymbol{w^T x_n}\right)}{1+\exp\left(-y_n\boldsymbol{w^T x_n}\right)}$$

b) If an example is misclassified, α is negative, so $\exp\alpha < 1$. The denominator of the error gradient of this example is therefore smaller than if the example were correctly classified, meaning the gradient is larger. A misclassified example therefore contributed more to a change in weights.

3: AML Pg. 94, Ex. 3.8

Answer:

The unit vector v-hat only gives the largest E_{in} decrease for small η because v-hat is defined by $\precapprox(\nvDash)$. Large η will adjust weights to be far from these initial weights, at which point the unit vector measuring $_{in}$ slope will have changed from v-hat. These new weights may therefore overshoot the minimum E_{in}. a)
$h(\boldsymbol{x}(t)) = \boldsymbol{w^T}(t)\boldsymbol{x}(t)$
For misclassification, $y(t) = -h(\boldsymbol{x}(t))$.
Proof by cases:
Case 1:
If $y(t) = +1$, $\boldsymbol{w^T}(t)\boldsymbol{x}(t) = -1$, so $y(t)\boldsymbol{w^T}(t)\boldsymbol{x}(t) = -1 < 0$
Case 2:

If $y(t) = -1$, $\boldsymbol{w^T}(t)\boldsymbol{x}(t) = +1$, so $y(t)\boldsymbol{w^T}(t)\boldsymbol{x}(t) = -1 < 0$

b) By eq. 1.3:

$$y(t)\boldsymbol{w^T}(t+1)\boldsymbol{x}(t) = y(t)[\boldsymbol{w}(t) + y(t)\boldsymbol{x}(t)]^T\boldsymbol{x}(t)$$
$$= y(t)\boldsymbol{w^T}(t)\boldsymbol{x}(t) + y^2(t)\boldsymbol{x^T}(t)\boldsymbol{x}(t)$$
$$\therefore y(t)\boldsymbol{w^T}(t+1)\boldsymbol{x}(t) > y(t)\boldsymbol{w^T}(t)\boldsymbol{x}(t)$$

holds if and only if

$$y^2(t)\boldsymbol{x^T}(t)\boldsymbol{x}(t) > 0$$

Both $y^2(t)$ and $\boldsymbol{x^T}(t)\boldsymbol{x}(t)$ are positive – the former because any scalar squared is positive, and the latter because $\boldsymbol{x^T}(t)\boldsymbol{x}(t)$ is a sum of the squared scalars within $\boldsymbol{x}(t)$. Both $y^2(t)\boldsymbol{x^T}(t)\boldsymbol{x}(t) > 0$ and subsequently $y(t)\boldsymbol{w^T}(t+1)\boldsymbol{x}(t) > y(t)\boldsymbol{w^T}(t)\boldsymbol{x}(t)$ are therefore true.

c) If $h(\boldsymbol{x}(t)) = -1$ and the point $\boldsymbol{x}(t)$ was misclassified, $y(t) = +1$. The update rule $\boldsymbol{w}(t+1) = \boldsymbol{w}(t) + y(t)\boldsymbol{x}(t)$ will therefore make the new decision boundary more positive. That is, for any t, $\boldsymbol{w}^T(t+1)\boldsymbol{x}(t)$ is more likely to be positive than $\boldsymbol{w}^T(t)\boldsymbol{x}(t)$. A future point similar to $\boldsymbol{x}(t)$ is therefore less likely to be misclassified as negative. The same argument holds with $h(\boldsymbol{x}(t)) = +1$ and $y(t) = -1$, where in this case the decision boundary is made more negative. Each misclassified point causes an adjustment that makes nearby points less likely to be misclassified in the same way.

4: AML Pg. 97, Ex. 3.9

Answer:

a)
See code notebook

b)
As one can see in the notebook graph, $e_{class}(s, y)$ makes its closest approach to $e_{sq}(s, y)$ when $s = 0$ and when $s = 1$. In the former case, $\lim_{s\to 0^+} e_{class}(s, 1) = 1$, and $e_sq(0, 1) = 1$. For $s < 0$, $e_{class} = 1$ and e_{sq} is decreasing, so for all $s \leq 0$, $e_{class} \leq e_{sq}$.
At $s = 1$, $e_{class}(1, 1) = 0$ and $e_{sq}(1, 1) = 1$. For $0 < s < 1$, e_{sq} is decreasing and $e_{class} = 0$, so $e_{class} < e_{sq}$. Lastly, for $s > 1$, $e_{class} = 0$ and e_{sq} is increasing from 0.
For all s, therefore, $e_{class} < e_{sq}$.

c)
Unlike $e_{sq}(s, 1)$, $e_{log}(s, 1)$ monotonically decreases with s. This is because $\frac{de_{log}(s,1)}{ds} = \frac{1}{1+\exp(-ys)} * \exp(-ys) * (-y)$, which is negative for all s. This monotonicity is also evident in the graph. Indeed, one can infer from the graph that $\lim_{s\to\infty} e_{log}(s, 1) = 0$. This holds mathematically as,

$$\lim_{s\to\infty} e_{log}(s, 1) = \lim_{s\to\infty} \ln(1 + \exp(-ys)) = \ln(1 + \exp(-\infty)) = \ln(1) = 0$$

5: PRML Pg. 134, Ex. 2.43

Answer: First, note that $\exp(\frac{-|x|^q}{2\sigma^2})$ is even, such that

$$\int_0^\infty \exp(\frac{-x^q}{2\sigma^2})dx = \frac{1}{2}\int_{-\infty}^\infty \exp(\frac{-|x|^q}{2\sigma^2})dx$$

So...

$$\frac{q}{2(2\sigma^2)^{1/q}\Gamma(\frac{1}{q})}\int_{-\infty}^\infty \exp(\frac{-|x|^q}{2\sigma^2})dx = \frac{q}{(2\sigma^2)^{1/q}\Gamma(\frac{1}{q})}\int_0^\infty \exp(\frac{-x^q}{2\sigma^2})dx$$

If we define $u = \frac{x^q}{2\sigma^2}$, then substituting in we get

$$\frac{q}{(2\sigma^2)^{1/q}\Gamma(\frac{1}{q})}\int_0^\infty \exp(\frac{-x^q}{2\sigma^2})dx = \frac{q}{(2\sigma^2)^{1/q}\Gamma(\frac{1}{q})}\int_0^\infty \frac{(2\sigma^2)^{1/q}}{q}u^{1/q}\exp(-u)du$$

$$= (\frac{q}{(2\sigma^2)^{1/q}\Gamma(\frac{1}{q})}) * (\frac{(2\sigma^2)^{1/q}}{q}\Gamma(\frac{1}{q})) = 1$$

As the integral evaluates to zero, this probability distribution is normalized.
For q =2, we get

$$p(x|\sigma^2, 2) = \frac{1}{(\sigma^2)^{1/2}\Gamma(\frac{1}{2})}\exp(-\frac{|x|^2}{2\sigma^2})$$

and since $\Gamma(1/2) = \sqrt{2\pi}$...

$$p(x|\sigma^2, 2) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{|x|^2}{2\sigma^2})$$

This is the Gaussian distribution.
The target variable can be expressed as $t = y(\boldsymbol{x}, \boldsymbol{w}) + \epsilon$, and therefore $-\epsilon$, a quantity we would like to maximize, is $y(\boldsymbol{x}, \boldsymbol{w}) - t$. The probability of achieving a target outcome is the sum of log likelihoods, where likelihood is the probability of $-\epsilon$. We therefore have

$$\ln p(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{w}, \boldsymbol{\sigma^2}) = \sum_{n=1}^N \ln p(y(\boldsymbol{x_n}, \boldsymbol{w}) - t_n)|\sigma^2, q)$$

$$= \frac{-1}{2\sigma^2}\sum_{n=1}^N |y(\boldsymbol{x_n}, \boldsymbol{w}) - t_n|^q + N\ln\frac{q}{2(2\sigma^2)^{\frac{1}{q}}\Gamma(\frac{1}{q})}$$

$$= \frac{-1}{2\sigma^2}\sum_{n=1}^N |y(\boldsymbol{x_n}, \boldsymbol{w}) - t_n|^q - \frac{N\ln 2\sigma^2}{q} +$$

6: PRML Pg. 173, Ex. 3.1

Answer:

We know that $\tanh a = \frac{e^a - e^{-a}}{e^a + e^{-a}}$ Let's start from the following:

$$2\sigma(2a) - 1 = \frac{2}{1 + e^{-2a}} - 1$$

$$= \frac{2}{1+e^{-2a}} * \frac{e^{2a}}{e^{2a}} - 1 = \frac{2e^{2a}}{1+e^{2a}} - 1$$
$$= \frac{2e^{2a}}{1+e^{2a}} - \frac{1+e^{2a}}{1+e^{2a}} = \frac{2e^{2a} - (1+e^{2a})}{1+e^{2a}}$$
$$= \frac{e^{2a}-1}{1+e^{2a}} = \frac{e^{2a}-1}{1+e^{2a}} * \frac{e^{-a}}{e^{-a}}$$
$$= \frac{e^{a}-e^{-a}}{e^{a}+e^{-a}}$$

That is, $2\sigma(2a) - 1 = \tanh a$.
Now start with $y(x, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M} \boldsymbol{w}_j \sigma(2a_j)$ where $a_j = \frac{x-\mu_j}{2s}$

$$y(x, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M} \frac{1}{2} \boldsymbol{w}_j (2\sigma(2a_j) + 1 - 1)$$

By the result $2\sigma(2a) - 1 = \tanh a$...

$$y(x, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M} \frac{1}{2} \boldsymbol{w}_j (\tanh a_j + 1)$$

Defining $u_j = \frac{\boldsymbol{w_j}}{2}$...

$$= w_0 + \sum_{j=1}^{M} u_j \tanh a_j + \sum_{j=1}^{M} \frac{\boldsymbol{w}_j}{2}$$
$$= u_0 + \sum_{j=1}^{M} u_j \tanh a_j$$

where I define $u_0 = w_0 + \sum_{j=1}^{M} \frac{\boldsymbol{w}_j}{2}$.

7: PRML Pg. 174, Ex. 3.2

Answer: Suppose we have a vector $\boldsymbol{v}$ that can be expressed by a sum of its orthogonal components as $\boldsymbol{v} = \boldsymbol{v_1} + \boldsymbol{v_2}$, where $\boldsymbol{v_1} \in \Phi$ and $\boldsymbol{v_2} \in \Phi^{\perp}$. There then exists a vector $\boldsymbol{w}$ such that $\Phi w = \boldsymbol{v_1}$. Let's operate on $\boldsymbol{v_1}$ as follows:

$$\Phi(\Phi^T\Phi)^{-1}\Phi^T \boldsymbol{v_1} = \Phi(\Phi^T\Phi)^{-1}\Phi^T\Phi \boldsymbol{w}$$

But we know that $(\Phi^T\Phi)^{-1}\Phi^T\Phi = 1$, so

$$\Phi(\Phi^T\Phi)^{-1}\Phi^T \boldsymbol{v_1} = \Phi \boldsymbol{w}$$
$$= \boldsymbol{v_1}$$

The operator $\Phi(\Phi^T\Phi)^{-1}\Phi^T$, then, returns $\boldsymbol{v_1}$ when it operates on $\boldsymbol{v_1}$.
By contrast, take $\boldsymbol{v_2}$, for which we know (by the definition of orthogonality) that $\Phi^T \boldsymbol{v_2} = 0$.
Therefore...

$$\Phi(\Phi^T\Phi)^{-1}\Phi^T \boldsymbol{v_2} = \Phi(\Phi^T\Phi)^{-1} 0 = 0$$

In conclusion, we can take any vector $\boldsymbol{v}$ and operate on it as follows:

$$\Phi(\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{v} = \Phi(\Phi^T\Phi)^{-1}\Phi^T(\boldsymbol{v_1} + \boldsymbol{v_2})$$
$$= \Phi(\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{v_1}$$
$$= \boldsymbol{v_1}$$

The vector $\boldsymbol{v}$, therefore, is projected onto the columns of Φ by $\Phi(\Phi^T\Phi)^{-1}\Phi^T$, as $\boldsymbol{v_1} \in \Phi$.

If we consider Φ to be our matrix of features and the vector $\boldsymbol{v}$ to be our target values $\boldsymbol{t}$, then $\Phi(\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{t}$ projects $\boldsymbol{t}$ onto the columns of our features, which is a the manifold $\mathcal{S}$ of dimensionality M (the number of features).

8: PRML Pg. 220, Ex. 4.1

Answer:

We can define an intersection of convex holds to be the point $\boldsymbol{z} = \sum_n \alpha_n \boldsymbol{x_n} = \sum_n \alpha_n \boldsymbol{y_n}$ for two sets of points $\{\boldsymbol{x_n}\}$ and $\{\boldsymbol{y_n}\}$.
For separability, we require a $\boldsymbol{w^T}$ and w_0 such that $\boldsymbol{w^T x_n} + w_0 > 0$ for all n and $\boldsymbol{w^T y_n} + w_0 < 0$ for all n. Consider, however, the point of intersection...

$$\boldsymbol{w^T z} + w_0 = \boldsymbol{w^T} \sum_n \alpha_n \boldsymbol{x_n} + w_0$$

Since, by the definition of α_n, $\sum_n \alpha_n = 1$...

$$\boldsymbol{w^T z} + w_0 = \boldsymbol{w^T} \sum_n \alpha_n \boldsymbol{x_n} + \sum_n \alpha_n w_0$$

Rearranging and acknowledging that $\boldsymbol{w^T}$ is independent of n...

$$\boldsymbol{w^T z} + w_0 = \sum_n \alpha_n (\boldsymbol{w^T x_n} + w_0)$$

However, by the same procedure, we also find the following:

$$\boldsymbol{w^T z} + w_0 = \sum_n \alpha_n (\boldsymbol{w^T y_n} + w_0)$$

We have found that $\sum_n \alpha_n(\boldsymbol{w^T x_n} + w_0) = \sum_n \alpha_n(\boldsymbol{w^T y_n} + w_0)$. Therefore, if $\boldsymbol{w^T x_n} + w_0$ is positive, nega-
tive, or zero, $\boldsymbol{w^T y_n} + w_0$ also is. This means $\{\boldsymbol{x_n}\}$ and $\{\boldsymbol{y_n}\}$ are not linearly separable, by definition.

Assuming linear separability gives us the following:

$$\boldsymbol{w^T x_n} + w_0 > 0$$

and

$$\boldsymbol{w^T y_n} + w_0 < 0$$

or vice versa. From before, we know that

$$\boldsymbol{w^T z_1} + w_0 = \sum_n \alpha_n (\boldsymbol{w^T x_n} + w_0)$$

$$\boldsymbol{w^T z_2} + w_0 = \sum_n \alpha_n (\boldsymbol{w^T y_n} + w_0)$$

where

$$\boldsymbol{z_1} = \sum_n \alpha_n x_n$$

$$\boldsymbol{z_2} = \sum_n \alpha_n y_n$$

This means that

$$\boldsymbol{w^T z_1} + w_0 \neq \boldsymbol{w^T z_2} + w_0$$

$$\boldsymbol{z_1} \neq \boldsymbol{z_2}$$

Since $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ are the convex hulls of $\boldsymbol{x_n}$ and $\boldsymbol{y_n}$, respectively, the hulls do not intersect.

9: PRML Pg. 221, Ex. 4.7

Answer:

Show $\sigma(-a) = 1 - \sigma(a)$.

$$\begin{aligned}\sigma(-a) &= \frac{1}{1+e^a} \\ &= \frac{e^{-a}}{1+e^{-a}} \\ &= \frac{1+e^{-a}-1}{1+e^{-a}} \\ &= 1 - \frac{1}{1+e^{-a}} \\ &= 1 - \sigma(a)\end{aligned}$$

Show $\sigma^{-1}(y) = \ln[\frac{y}{1-y}]$

$$\sigma^{-1}(y) \rightarrow y = \sigma(a) = \frac{1}{1+e^{-a}}$$

$$1 + e^{-a} = \frac{1}{y}$$

$$\sigma^{-1}(y) = a = -\ln[\frac{1}{y} - 1]$$

By the rules of logarithms...

$$\begin{aligned}\sigma^{-1}(y) &= \ln[\frac{1}{y} - 1]^{-1} \\ &= \ln[\frac{1}{\frac{1}{y} - 1}]\end{aligned}$$

$$= \ln[\frac{1}{\frac{1}{y} - 1} * \frac{y}{y}]$$
$$= \ln[\frac{y}{1-y}]$$

10: PRML Pg. 222, Ex. 4.12

Answer:

$$\frac{d\sigma}{da} = \frac{d}{da}(\frac{1}{1+e^{-a}})$$
$$= \frac{d}{da}(1+e^{-a})^{-1}$$
$$= -1 * (1+e^{-a})^{-2} * e^{-a} * -1$$
$$= \frac{e^{-a}}{(1+e^{-a})^2}$$
$$= \frac{1}{1+e^{-a}} * \frac{e^{-a}}{1+e^{-a}} * \frac{e^a}{e^a}$$
$$= \sigma(a) * \frac{1}{1+e^a}$$
$$= \sigma(a) * \sigma(-a)$$

And by the result of PRML Ex. 4.7...

$$\frac{d\sigma}{da} = \sigma(a)(1-\sigma(a))$$

11: PRML Pg. 222, Ex. 4.13

Answer:

$$E(\boldsymbol{w}) = -\sum_{n=1}^{N}[t_n \ln y_n + (1-t_n)\ln(1-y_n)]$$

where $y_n = \sigma(\boldsymbol{w}^T\phi_n)$ We can distribute the gradient into the sum as follows:

$$\nabla_w E = -\sum_{n=1}^{N}[t_n \nabla_w \ln y_n + (1-t_n)\nabla_w \ln(1-y_n)]$$

since the target is independent of the weights. By the result of PRML Ex. 10 and the chain rule...

$$\nabla_w E = -\sum_{n=1}^{N}[t_n(\frac{1}{\sigma(\boldsymbol{w}^T\phi_n)} * \sigma(\boldsymbol{w}^T\phi_n) * (1-\sigma(\boldsymbol{w}^T\phi_n)) * \phi_n) + (1-t_n)(\frac{1}{1-\sigma(\boldsymbol{w}^T\phi_n)} * -\sigma(\boldsymbol{w}^T\phi_n) * (1-\sigma(\boldsymbol{w}^T\phi_n) * \phi_n)]$$

Distributing...

$$\nabla_w E = -\sum_{n=1}^{N}[(t_n\phi_n - t_n\phi_n\sigma(\boldsymbol{w}^T\phi_n)) + (-\phi_n\sigma(\boldsymbol{w}^T\phi_n) + t_n\phi_n\sigma(\boldsymbol{w}^T\phi_n))]$$

The second and fourth terms cancel, leaving the following:

$$\nabla_w E = -\sum_{n=1}^{N}\phi_n(t_n - \sigma(\boldsymbol{w}^T\phi_n))$$

12: PRML Pg. 222, Ex. 4.14

Answer:

Assume there are two classes C_1 and C_2 where there exists a point $\boldsymbol{x_i} \rightarrow \boldsymbol{w}^T\phi(\boldsymbol{x_i}) > 0$ such that $\boldsymbol{x_i}$ is in C_1 and there exists a different point $\boldsymbol{x_j} \rightarrow \boldsymbol{w}^T\phi(\boldsymbol{x_j}) < 0$ such that $\boldsymbol{x_j}$ is in C_2.

Intuitively, we can see that the decision boundary for the weight vector $\boldsymbol{w}$ occurs at $\boldsymbol{w}^T\phi(\boldsymbol{x}) = 0$

For any point $\boldsymbol{x_i}$ in C_1, $|w| \rightarrow \infty$ causes

$$p(C_1|\phi(x_i)) = \sigma(\boldsymbol{w}^T\phi(\boldsymbol{x_i})) \rightarrow 1$$

This is because $\sigma(\boldsymbol{w}^T\phi(\boldsymbol{x_i})) = \frac{1}{1+e^{-\boldsymbol{w}^T\phi(\boldsymbol{x_i})}}$, where large weights will cause the exponent in the denominator to become very negative, rendering that term effectively zero and leaving $\sigma(\boldsymbol{w}^T\phi(\boldsymbol{x_i})) = \frac{1}{1}$.

For any point $\boldsymbol{x_j}$ in C_2, $|w| \rightarrow \infty$ causes

$$p(C_2|\phi(x_j)) \rightarrow 1$$

This is because

$$\begin{aligned} p(C_2|\phi(x_j)) &= 1 - p(C_1|\phi(x_j)) \\ &= 1 - \sigma(\boldsymbol{w}^T\phi(\boldsymbol{x_j})) \\ &= 1 - \frac{1}{1 + e^{-\boldsymbol{w}^T\phi(\boldsymbol{x_j})}} \end{aligned}$$

But $\boldsymbol{w}^T\phi(\boldsymbol{x_j})$ is negative (C_2 is represented by predictions less than 0) and has a large magnitude, so we get

$$\begin{aligned} p(C_2|\phi(x_j)) &= 1 - \frac{1}{1+\infty} \\ &= 1 \end{aligned}$$

We therefore see that $|w| \rightarrow \infty$ maximizes the probability of correct classification for both classes – a reason that hard-boundary error functions are more prone to overfitting than soft-boundary ones.

13: PRML Pg. 223, Ex. 4.17

Answer:

Note that $y_k = \frac{e^{a_k}}{\sum_j e^{a_j}}$

For the case $k \neq j$...

$$\frac{\partial y_k}{\partial a_j} = \frac{\partial}{\partial a_j} \frac{e^{a_k}}{\sum_j e^{a_j}}$$

$$= e^{a_k} * (-1) * (\sum_j e^{a_j})^{-2} * e^{a_j}$$

$$= -\frac{e^{a_k}}{\sum_j e^{a_j}} * \frac{e^{a_j}}{\sum_j e^{a_j}}$$

$$= -y_k y_j$$

More generally, we must use the product rule to account for a_j dependency in both the numerator and denominator...

$$\frac{\partial y_k}{\partial a_j} = \frac{e^{a_k}}{\sum_j e^{a_j}} + e^{a_k} * (-1) * (\sum_j e^{a_j})^{-2} * e^{a_k}$$

$$= \frac{e^{a_k}}{\sum_j e^{a_j}} - (\frac{e^{a_k}}{\sum_j e^{a_j}})^2$$

$$= y_k(1 - y_k)$$

But also note that we may have a vector of outputs $\boldsymbol{y_k}$, so in general..

$$\frac{\partial \boldsymbol{y_k}}{\partial a_j} = \boldsymbol{y_k}(I_{kj} - \boldsymbol{y_k})$$

where I_{kj} is the identity matrix.

14: PRML Pg. 223, Ex. 4.18

Answer: Distribute the gradient into the error function sums, noting that the target values do not depend on the weights

$$\nabla_{w_j} E(\boldsymbol{w_1}, ..., \boldsymbol{w_K}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \nabla_{w_j} \ln y_{nk}$$

Using the chain rule...

$$\nabla_{w_j} E(\boldsymbol{w_1}, ..., \boldsymbol{w_K}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \frac{1}{y_{nk}} \nabla_{w_j} y_{nk}$$

From PRML Ex. 4.17, and using the chain rule, we can express this as follows:

$$\nabla_{w_j} E(\boldsymbol{w_1}, ..., \boldsymbol{w_K}) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \frac{1}{y_{nk}} y_{nk}(I_{kj} - y_{nj}) \phi_n$$

$$= -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}(I_{kj} - y_{nj})\phi_n$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}y_{nj}\phi_n - \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}I_{kj}\phi_n$$

We can eliminate the k-dependence in the second-term by acknowledging that anything times the identiy matrix is itself projected onto the columns of the identity matrix.

$$\nabla_{w_j} E(\boldsymbol{w_1}, ..., \boldsymbol{w_K}) = \sum_{n=1}^{N}(\sum_{k=1}^{K} t_{nk})y_{nj}\phi_n - \sum_{n=1}^{N} t_{nj}\phi_n$$

For a classification problem, the sum of probabilities of classes for any target example is 1.

$$\nabla_{w_j} E(\boldsymbol{w_1}, ..., \boldsymbol{w_K}) = \sum_{n=1}^{N} y_{nj}\phi_n - \sum_{n=1}^{N} t_{nj}\phi_n$$

$$= \sum_{n=1}^{N}(y_{nj} - t_{nj})\phi_n$$