Yiru Wang, 930891925
ywang24@email.wm.edu
CSCI 420-02 Neural Networks for Machine Learning, Spring 2019

Homework 3, Due March 19, 2019

AML Ex:3.1:

Will PLA ever stop updating if the data is not linearly separable?

Solution:

No. PLA will not stop updating even if the data is not linearly separable because there always exist some examples that are misclassified.

AML Ex:3.7:

For logistic regression, **show** that

$$\begin{aligned}\nabla E_{\text{in}}(\mathbf{w}) &= -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n\mathbf{x}_n}{1+e^{y_n\mathbf{w}^T\mathbf{x}_n}} \\ &= \frac{1}{N}\sum_{n=1}^{N}-y_n\mathbf{x}_n\theta(-y_n\mathbf{w}^T\mathbf{x}_n)\end{aligned}$$

Argue that a 'misclassified' example contributes more to the gradient than a correctly classified one.

Solution:

(a) The E_{in} is as follows.

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N}\ln\left(1+e^{-y_n\mathbf{w}^T\mathbf{x}_n}\right)$$

Thus,

$$\begin{aligned}\frac{\mathrm{d}E_{\text{in}}(\mathbf{w})}{\mathrm{d}w_i} &= \frac{1}{N}\sum_{n=1}^{N}\frac{\mathrm{d}\ln\left(1+e^{-y_n\mathbf{w}^T\mathbf{x}_n}\right)}{\mathrm{d}w_i} \\ &= \frac{1}{N}\sum_{n=1}^{N}\frac{1}{1+e^{-y_n\mathbf{w}^T\mathbf{x}_n}}e^{-y\mathbf{w}^T\mathbf{x}}(-yx_i) \\ &= \frac{1}{N}\sum_{n=1}^{N}\frac{-yx_i}{1+e^{y_n\mathbf{w}^T\mathbf{x}_n}}\end{aligned}$$

Therefore,

$$\nabla E_{\text{in}}(\mathbf{w}) = -\frac{1}{N}\sum_{n=1}^{N}\frac{y_n\mathbf{x}_n}{1+e^{y_n\mathbf{w}^T\mathbf{x}_n}} = \frac{1}{N}\sum_{n=1}^{N}-y_n\mathbf{x}_n\theta(-y_n\mathbf{w}^T\mathbf{x}_n)$$

since $\theta(s) = \frac{e^s}{1+e^s}$ and $\theta(-s) = 1-\theta(s)$.

(b) From part (a), it is to see that the gradient that each example contributes to is decided by the value of $\theta(\cdot)$.

The value of 'misclassified' example is $y\mathbf{w}^T\mathbf{x} < 0.5$ and that of the correctly classified example is $y\mathbf{w}^T\mathbf{x} > 0.5$. In addition, $\theta(-s) = 1-\theta(s)$.

Therefore $\theta(\text{incorrect-example}) > \theta(\text{correct-example})$, i.e. the 'misclassified' examples contribute more to the gradient than correctly classified ones.

AML Ex:3.8:

The claim that $\hat{\mathbf{v}}$ is the direction which gives largest decrease in E_{in} only holds for small η. Why?

Solution:

Because the Taylor expansion is used in the above conclusions, this theorem is only established when η is relatively small; when η is too large, the Taylor expansion cannot be applied, and the conclusion may not be established.

AML Ex:3.9:

Consider pointwise error measures $\mathrm{e}_{\text{class}}(s,y) = [\![y \neq \text{sign}(s)]\!]$, $\mathrm{e}_{\text{sq}}(s,y) = (y-s)^2$, and $\mathrm{e}_{\log}(s,y) = \ln(1+\exp(-ys))$, where the signal $s = \mathbf{w}^T\mathbf{x}$.

(a) For $y = +1$, plot $\mathrm{e}_{\text{class}}$, e_{sq} and $\frac{1}{\ln 2}\mathrm{e}_{\log}$ versus s, on the same plot.

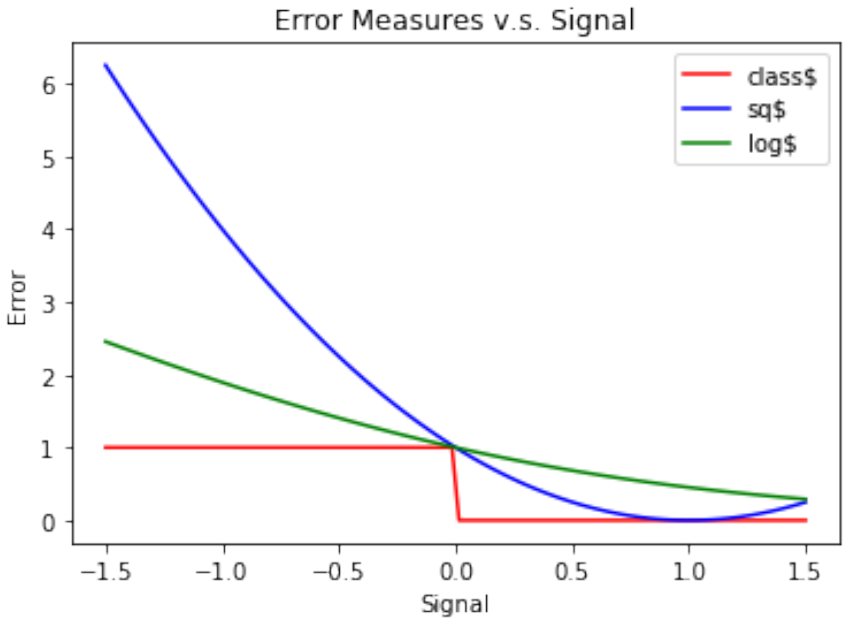
(b) Show that $\mathrm{e}_{\text{class}}(s,y) \leq \mathrm{e}_{\text{sq}}(s,y)$, and hence that the classification error is upper bounded by the squared error.

(c) Show that $\mathrm{e}_{\text{class}}(s,y) \leq \frac{1}{\ln 2}\mathrm{e}_{\mathbf{log}}(s,y)$, and, as in part (b), get an upper bound (up to a constant factor) using the logistic regression error.

These bounds indicate that minimizing the squared or logistic regression error should also decrease the classification error, which justifies using the weights returned by linear or logistic regression as approximations for classification.

Solution:

(a) Please refer to the attachment (Jupyter notebook – AML-plot.ipynb) for detailed code. The produced figure is as follows.



(b) We discuss in two cases.
Case $y = +1$: when $s > 0$, $e_{\text{class}} = 0$; and $e_{\text{sq}} \geq 0$, so $e_{\text{class}} \leq e_{\text{sq}}$ at this time.
When $s \leq 0$, $e_{\text{class}} = +1$; and $e_{\text{sq}} \geq +1$ (the equal sign is true when $s = 0$), so $e_{\text{class}} \leq e_{\text{sq}}$ at this time.
In summary, when $y = +1$, $e_{\text{class}} \leq e_{\text{sq}}$.

Similarly, when $y = -1$, it is easy to verify $e_{\text{class}} \leq e_{\text{sq}}$.

Therefore, the classification error is upper bounded by the squared error.

(c) We discuss in two cases.
Case $y = +1$: when $s > 0$, $e_{\text{class}} = 0$; and $\frac{1}{\ln 2} e_{\text{log}} \geq 0$, so $e_{\text{class}} \leq e_{\text{log}}$ at this time.
When $s \leq 0$, $e_{\text{class}} = +1$; and $\frac{1}{\ln 2} e_{\text{sq}} \geq +1$ (the equal sign is true when $s = 0$), so $e_{\text{class}} \leq e_{\text{log}}$ at this time.
In summary, when $y = +1$, $e_{\text{class}} \leq e_{\text{log}}$.

Similarly, when $y = -1$, it is easy to verify $e_{\text{class}} \leq e_{\text{log}}$.

Therefore, the classification error is upper bounded by the log error (scaled by $\ln 2$).

PRML 2.43:

The following distribution

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

where

$$\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u) \mathrm{d}u$$

is a generalization of the univariate Gaussian distribution. **Show** that this distribution is normalized so that

$$\int_{-\infty}^{+\infty} p(x|\sigma^2, q) \mathrm{d}x = 1$$

and that it reduces to the Gaussian when $q = 2$. Consider a regression model in which the target variable is given by $t = y(\mathbf{x}, \mathbf{w}) + \epsilon$ and ϵ is a random noise variable drawn from the distribution (2.293). **Show** that the log likelihood function over $\mathbf{w}$ and σ^2, for an observed data set of input vectors $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and corresponding target variables $\mathbf{t} = (t_1, \ldots, t_N)^T$, is given by

$$\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} |y(\mathbf{x}_n, \mathbf{w}) - t_n|^q - \frac{N}{q} \ln(2\sigma^2) + \text{const}$$

where 'const' denotes terms independent of both $\mathbf{w}$ and σ^2. Note that, as a function of $\mathbf{w}$, this is the L_q error function considered in Section 1.5.5.

Solution:

(a) Make the change of variable

$$u = \frac{x^q}{2\sigma^2}$$

and

$$x = (2\sigma^2 u)^{1/q}$$
$$\mathrm{d}x = \frac{(2\sigma^2)^{1/q} u^{1/q-1}}{q} \mathrm{d}u.$$

Then, we obtain:

$$\begin{aligned}
\int_{-\infty}^{+\infty} p(x|\sigma^2, q)\mathrm{d}x &= \int_{-\infty}^{+\infty} \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)} \exp(-\frac{|x|^q}{2\sigma^2})\mathrm{d}x \\
&= \int_{0}^{+\infty} \frac{q}{(2\sigma^2)^{1/q}\Gamma(1/q)} \exp(-\frac{|x|^q}{2\sigma^2})\mathrm{d}x \\
&= \int_{0}^{+\infty} \frac{q}{(2\sigma^2)^{1/q}\Gamma(1/q)} \exp(-u) \cdot \frac{(2\sigma^2)^{1/q} u^{\frac{1}{q}-1}}{q}\mathrm{d}u \\
&= \int_{0}^{+\infty} \frac{\exp(-u)u^{\frac{1}{q}-1}}{\Gamma(1/q)}\mathrm{d}u \\
&= \frac{\int_0^{+\infty} \exp(-u)u^{\frac{1}{q}-1}\mathrm{d}u}{\int_0^{+\infty} \exp(-x)x^{\frac{1}{q}-1}\mathrm{d}x} \\
&= 1
\end{aligned}$$

Hence, this distribution is normailzed.

(b) Because

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon$$

and

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

the likelihood

$$p(t|\mathbf{x}, \mathbf{w}, \sigma^2) = \frac{q}{2(2\sigma^2)^{1/q}\Gamma(1/q)} \exp\Big(-\frac{|t - y(\mathbf{x}, \mathbf{w})|^q}{2\sigma^2}\Big).$$

and

$$\ln p(t|\mathbf{x}, \mathbf{w}, \sigma^2) = -\frac{1}{q}\ln(2\sigma^2) - \frac{|t - y(\mathbf{x}, \mathbf{w})|^q}{2\sigma^2} + \ln\frac{q}{2\Gamma(1/q)}.$$

Hence,

$$\begin{aligned}
\ln p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2) &= \ln \prod_{n=1}^{N} p(t|\mathbf{x}, \mathbf{w}, \sigma^2) \\
&= \sum_{n=1}^{N} \ln p(t|\mathbf{x}, \mathbf{w}, \sigma^2) \\
&= -\frac{1}{2\sigma^2}\sum_{n=1}^{N} |y(\mathbf{x}_n, \mathbf{w}) - t_n|^q - \frac{N}{q}\ln(2\sigma^2) + \text{const}
\end{aligned}$$

PRML 3.1:

Show that the 'tanh' function and the logistic sigmoid function (3.6)

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

are related by

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

Hence show that a general linear combination of logistic sigmoid functions of the form

$$y(x, \mathbf{w}) = w_0 + \sum_{j=1}^{M} w_j \sigma\Big(\frac{x - \mu_j}{s}\Big)$$

is equivalent to a linear combination of 'tanh' functions of the form

$$y(x, \mathbf{u}) = u_0 + \sum_{j=1}^{M} u_j \tanh\Big(\frac{x - \mu_j}{s}\Big)$$

and **find** expressions to related the new parameters $\{u_1, \ldots, u_M\}$ to the original parameters $\{w_1, \ldots, w_M\}$.

Solution:

(a)

$$\begin{aligned}
\text{RHS} &= 2\sigma(2a) - 1 \\
&= \frac{2}{1 + \exp(-2a)} - 1 \\
&= \frac{1 - \exp(-2a)}{1 + \exp(-2a)} \\
&= \frac{1 - \exp(-2a)}{1 + \exp(-2a)} \cdot \frac{\exp(a)}{\exp(a)} \\
&= \frac{\exp(a) - \exp(-a)}{\exp(a) + \exp(-a)} \\
&= \tanh(a) \\
&= \text{LHS}
\end{aligned}$$

(b) If take $a_j = (x - \mu_j)/2s$, then

$$\begin{aligned} y(x, \mathbf{w}) &= w_0 + \sum_{j=1}^{M} w_j \sigma(2a_j) \\ &= w_0 + \sum_{j=1}^{M} \frac{w_j}{2}(2\sigma(2a_j) - 1 + 1) \\ &= \Big(w_0 + \sum_{j=1}^{M} \frac{w_j}{2}\Big) + \sum_{j=1}^{M} \frac{w_j}{2} \tanh(a_j) \\ &= u_0 + \sum_{j=1}^{M} u_j \tanh(a_j) \end{aligned}$$

Hence, $u_j = w_j/2$, for $j = 1, \ldots, M$, and $u_0 = w_0 + \sum_{j=1}^{M} w_j/2$.

PRML 3.2:

Show that the matrix

$$\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T$$

takes any vector $\mathbf{v}$ and projects it onto the space spanned by the columns of $\mathbf{\Phi}$. Use this result to **show** that the least-square solution (3.15)

$$\mathbf{w}_{\mathrm{ML}} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{t}$$

corresponds to an orthogonal projection of the vector $\mathbf{t}$ onto the manifold $\mathcal{S}$ as shown in Figure 3.2.

Solution:

We first write

$$\begin{aligned} \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{v} &= \mathbf{\Phi}\tilde{\mathbf{v}} \\ &= \varphi_1\tilde{v}^{(1)} + \cdots + \varphi_M\tilde{v}^{(M)} \end{aligned}$$

where φ_m is the m-th column of $\mathbf{\Phi}$ and $\tilde{\mathbf{v}} = (\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{v}$. By comparing this with the least squares solution, we can see that

$$\mathbf{\Phi}\mathbf{w}_{\mathrm{ML}} = \mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{t}$$

corresponds to a projection of $\mathbf{t}$ onto the space spanned by the columns of $\mathbf{\Phi}$. To see that this is indeed an orthogonal projection, we first note that for any column of $\mathbf{\Phi}$, φ_j,

$$\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\varphi_{\mathbf{j}} = [\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T\mathbf{\Phi}]_j = \varphi_j$$

and therefore

$$(\mathbf{y} - \mathbf{t})^T\varphi_j = (\mathbf{\Phi}\mathbf{w}_{\mathrm{ML}} - \mathbf{t})^T\varphi_j = \mathbf{t}^T(\mathbf{\Phi}(\mathbf{\Phi}^T\mathbf{\Phi})^{-1}\mathbf{\Phi}^T - \mathbf{I})^T\varphi_j = 0$$

and thus $(\mathbf{y} - \mathbf{t})$ is ortogonal to every column of $\mathbf{\Phi}$ and hence is orthogonal to $\mathcal{S}$.

PRML 4.1:

Given a set of data points $\{x_n\}$, we can define the *convex hull* to be the set of all points $\mathbf{x}$ given by

$$\mathbf{x} = \sum_n \alpha_n \mathbf{x}_n$$

where $\alpha_n \geq 0$ and $\sum_n \alpha_n = 1$. Consider a second set of points $\{\mathbf{y}_n\}$ together with their corresponding convex hull. By definition, the wto sets of points will be linearly separable if there exists a vector $\hat{\mathbf{w}}$ and a scalar w_0 such that $\hat{\mathbf{w}}^T\mathbf{x}_n + w_0 > 0$ for all $\mathbf{x}_n$, and $\hat{\mathbf{w}}^T\mathbf{y}_n + w_0 < 0$ for all $\mathbf{y}_n$. **Show** that if their convex hulls intersect, the two sets of points cannot be linearly separable, and conversely that if they are linearly separable, their convex hulls do not intersect.

Solution:

Assume that the convex hulls of $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_m\}$ intersect. Then there exist a point $\mathbf{z}$ such that

$$\mathbf{z} = \sum_n \alpha_n \mathbf{x}_n = \sum_m \beta_m \mathbf{y}_m$$

where $\beta_m \geq 0$ for all m and $\sum_m \beta_m = 1$.
If $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_m\}$ also were to be linearly separable, we would have that

$$\hat{\mathbf{w}}^T\mathbf{z} + w_0 = \sum_n \alpha_n \hat{\mathbf{w}}^T\mathbf{x}_n + w_0$$

since $\hat{\mathbf{w}}^T\mathbf{x}_n + w_0 > 0$ and the $\{\alpha_n\}$ are all non-negative and sum to 1, but by the corresponding argument

$$\hat{\mathbf{w}}^T\mathbf{z} + w_0 = \sum_m \beta_m \hat{\mathbf{w}}^T\mathbf{y}_m + w_0 = \sum_m \beta_m(\hat{\mathbf{w}}^T\mathbf{y}_m + w_0) < 0,$$

which is a contradiction and hence $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_m\}$ cannot be linearly separable if their convex hulls intersect.
If we instead assume that $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_m\}$ are linearly separable and consider a point $\mathbf{z}$ in the intersection of their convex hulls, the same contradiction arise. Thus no such point can exist and the intersection of the convex hulls of $\{\mathbf{x}_n\}$ and $\{\mathbf{y}_m\}$ must be empty.

PRML 4.7:

Show that the logitic sigmoid function (4.59)

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

satisfies the property $\sigma(-a) = 1 - \sigma(a)$ and that its inverse is given by $\sigma^{-1}(y) = \ln\{y/(1-y)\}$.

Solution:

(a)

$$\text{LHS} = \sigma(-a) = \frac{1}{1 + \exp(a)}$$

$$\text{RHS} = 1 - \sigma(a) = \frac{\exp(-a)}{1 + \exp(-a)} = \frac{\exp(-a)}{1 + \exp(-a)} \cdot \frac{\exp(a)}{\exp(a)} = \frac{1}{1 + \exp(a)}$$

Thus, $\sigma(-a) = 1 - \sigma(a)$.

(b)

$$\ln\{\sigma(a)/(1-\sigma(a))\} = \ln \frac{1}{\exp(-a)} = a$$

Hence, the inverse of σ is $\sigma^{-1}(y) = \ln\{y/(1-y)\}$.

PRML 4.12:

Verify the relation (4.88)

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

for the derivative of the logistic sigmoid function defined by (4.59).

$$\sigma(a) = \frac{1}{1+\exp(-a)}$$

Solution:

$$\begin{aligned}
\frac{\mathrm{d}\sigma}{\mathrm{d}a} &= \frac{\mathrm{d}(1+\exp(-a))^{-1}}{\mathrm{d}(1+\exp(-a))} \cdot \frac{\mathrm{d}(1+\exp(-a))}{\mathrm{dexp}(-a)} \cdot \frac{\mathrm{dexp}(-a)}{\mathrm{d}a} \\
&= -(1+\exp(-a))^{-2} \cdot 1 \cdot \exp(-a) \cdot -1 \\
&= \frac{1}{1+\exp(-a)} \cdot \frac{\exp(-a)}{1+\exp(-a)} \\
&= \sigma(1-\sigma)
\end{aligned}$$

PRML 4.13:

By making use of the result (4.88)

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

for the derivative of the logistic sigmoid, **show** that the derivative of the error function (4.90)

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N}\{t_n \ln y_n + (1-t_n)\ln(1-y_n)\}$$

where $y_n = \sigma(a_n)$ and $a_n = \mathbf{w}^T\phi_n$ for the logistic regression models is given by (4.91). Taking the gradient of the error function with respect to $\mathbf{w}$,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\phi_n$$

Solution:

$$\begin{aligned}
\nabla E(\mathbf{w}) &= -\nabla \ln p(\mathbf{t}|\mathbf{w}) \\
&= -\sum_{n=1}^{N}\{t_n \nabla \ln y_n + (1-t_n)\nabla \ln(1-y_n)\} \\
&= -\sum_{n=1}^{N}\{\frac{t_n}{y_n}\cdot y_n(1-y_n)\phi_n - \frac{1-t_n}{1-y_n}\cdot y_n(1-y_n)\phi_n\} \\
&= -\sum_{n=1}^{N}(y_n - t_n)\phi_n
\end{aligned}$$

PRML 4.14:

Show that for a linearly separable dataset, the maximum likelihood solution for the logistic regression model is obtained by finding a vector $\mathbf{w}$ whose decision boundary $\mathbf{w}^T\phi(x) = 0$ separates the classes and then taking the magnitude of $\mathbf{w}$ to infinity.

Solution:

If the data set is linearly separable, any decision boundary separating the two classes will have the property

$$\mathbf{w}^T\phi_n = \begin{cases} \geq 0 & \text{if } t_n = 1, \\ < 0 & \text{otherwise.} \end{cases}$$

Moreover, from

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

we see that the negative log-likelihood will be minimized when $y_n = \sigma(\mathbf{w}^T\phi_n) = t_n$ for all n. This will be the case when the sigmoid function is saturated, which occurs when its argument, $\mathbf{w}^T\phi$, goes to $\pm\infty$, i.e., when the magnitude of $\mathbf{w}$ goes to infinity.

PRML 4.17:

Show that the derivatives of the softmax activation function (4.104),

$$p(\mathcal{C}_k|\phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

where the a_k are defined by (4.105),

$$a_k = \mathbf{w}_k^T \phi$$

are given by (4.106).

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

where I_{kj} are the elements of the identify matrix.

Solution:

From

$$p(\mathcal{C}_k|\phi) = y_k(\phi) = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

we have

$$\begin{aligned}
\frac{\partial y_k}{\partial a_k} &= \frac{e^{a_k}}{\sum_i e^{a_i}} - \left(\frac{e^{a_k}}{\sum_i e^{a_i}}\right)^2 &&= y_k(1 - y_k), \\
\frac{\partial y_k}{\partial a_j} &= -\frac{e^{a_k} e^{a_j}}{\left(\sum_i e^{a_i}\right)^2} &&= -y_k y_j, \qquad j \neq k.
\end{aligned}$$

Therefore, we can obtain

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

PRML 4.18:
Using the result (4.91) Taking the gradient of the error function with respect to $\mathbf{w}$,

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N}(y_n - t_n)\phi_n$$

for the derivatives of the softmax activation function, **show** that the gradients of the cross-entropy error (4.108)

$$E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{w}_1, \ldots, \mathbf{w}_k) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}$$

are given by (4.109). We now take the gradient of the error function with respect to one of the parameter vectors $\mathbf{w}_j$,

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_k) = \sum_{n=1}^{N}(y_{nj} - t_{nj})\phi_n.$$

Solution:
From

$$E(\mathbf{w}_1, \ldots, \mathbf{w}_K) = -\ln p(\mathbf{T}|\mathbf{w}_1, \ldots, \mathbf{w}_k) = -\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \ln y_{nk}$$

we have

$$\frac{\partial E}{\partial y_n k} = -\frac{t_{nk}}{y_{nk}}.$$

Then we combine this with (4.106) using the chain rule, we get

$$\begin{aligned}
\frac{\partial E}{\partial a_{nj}} &= \sum_{k=1}^{K} \frac{\partial E}{\partial y_{nk}} \frac{\partial y_{nk}}{\partial a_{nj}} \\
&= -\sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}} y_{nk}(I_{kj} - y_{nj}) \\
&= y_{nj} - t_{nj}.
\end{aligned}$$

where we have used that $\forall n : \sum_k t_{nk} = 1$.
If we combine this with (solution PRML 4.13), again using the chain rule, we can obtain

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \ldots, \mathbf{w}_k) = \sum_{n=1}^{N}(y_{nj} - t_{nj})\phi_n.$$