ECE1513
Tutorial 4:
Selected Exercises from Chapter 5

TA: Faye Pourghasem, fateme.pourghasem@mail.utoronto.ca

September 22, 2023

1 Example 5.2

5.2 (⋆) WWW Show that maximizing the likelihood function under the conditional distribution (5.16) for a multioutput neural network is equivalent to minimizing the sum-of-squares error function (5.11).

Solution

5.2 The likelihood function for an i.i.d. data set, $\{(\mathbf{x}_1, \mathbf{t}_1), \ldots, (\mathbf{x}_N, \mathbf{t}_N)\}$, under the conditional distribution (5.16) is given by

$$\prod_{n=1}^{N} \mathcal{N}\left(\mathbf{t}_n | \mathbf{y}(\mathbf{x}_n, \mathbf{w}), \beta^{-1}\mathbf{I}\right).$$

If we take the logarithm of this, using (2.43), we get

$$\begin{aligned}
&\sum_{n=1}^{N} \ln \mathcal{N}\left(\mathbf{t}_n | \mathbf{y}(\mathbf{x}_n, \mathbf{w}), \beta^{-1}\mathbf{I}\right) \\
&= -\frac{1}{2}\sum_{n=1}^{N} \left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right)^{\mathrm{T}} (\beta\mathbf{I}) \left(\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\right) + \text{const} \\
&= -\frac{\beta}{2}\sum_{n=1}^{N} \|\mathbf{t}_n - \mathbf{y}(\mathbf{x}_n, \mathbf{w})\|^2 + \text{const},
\end{aligned}$$

where 'const' comprises terms which are independent of $\mathbf{w}$. The first term on the right hand side is proportional to the negative of (5.11) and hence maximizing the log-likelihood is equivalent to minimizing the sum-of-squares error.

2 Example 5.9

5.9 ($\star$) **WWW** The error function (5.21) for binary classification problems was derived for a network having a logistic-sigmoid output activation function, so that $0 \leqslant y(\mathbf{x}, \mathbf{w}) \leqslant 1$, and data having target values $t \in \{0, 1\}$. Derive the corresponding error function if we consider a network having an output $-1 \leqslant y(\mathbf{x}, \mathbf{w}) \leqslant 1$ and target values $t = 1$ for class $\mathcal{C}_1$ and $t = -1$ for class $\mathcal{C}_2$. What would be the appropriate choice of output unit activation function?

Solution

5.9 This simply corresponds to a scaling and shifting of the binary outputs, which directly gives the activation function, using the notation from (5.19), in the form

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

The corresponding error function can be constructed from (5.21) by applying the inverse transform to y_n and t_n, yielding

$$\begin{aligned} E(\mathbf{w}) &= -\sum_{n=1}^{N} \frac{1+t_n}{2} \ln \frac{1+y_n}{2} + \left(1 - \frac{1+t_n}{2}\right) \ln \left(1 - \frac{1+y_n}{2}\right) \\ &= -\frac{1}{2} \sum_{n=1}^{N} \{(1+t_n) \ln(1+y_n) + (1-t_n) \ln(1-y_n)\} + N \ln 2 \end{aligned}$$

where the last term can be dropped, since it is independent of $\mathbf{w}$.

To find the corresponding activation function we simply apply the linear transformation to the logistic sigmoid given by (5.19), which gives

$$\begin{aligned} y(a) &= 2\sigma(a) - 1 = \frac{2}{1+e^{-a}} - 1 \\ &= \frac{1-e^{-a}}{1+e^{-a}} = \frac{e^{a/2} - e^{-a/2}}{e^{a/2} + e^{-a/2}} \\ &= \tanh(a/2). \end{aligned}$$

3 Example 5.10

5.10 (⋆) WWW Consider a Hessian matrix $\mathbf{H}$ with eigenvector equation (5.33). By setting the vector $\mathbf{v}$ in (5.39) equal to each of the eigenvectors $\mathbf{u}_i$ in turn, show that $\mathbf{H}$ is positive definite if, and only if, all of its eigenvalues are positive.

Solution

5.10 From (5.33) and (5.35) we have

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{H}\mathbf{u}_i = \mathbf{u}_i^{\mathrm{T}}\lambda_i\mathbf{u}_i = \lambda_i.$$

Assume that $\mathbf{H}$ is positive definite, so that (5.37) holds. Then by setting $\mathbf{v} = \mathbf{u}_i$ it follows that

$$\lambda_i = \mathbf{u}_i^{\mathrm{T}}\mathbf{H}\mathbf{u}_i > 0 \tag{174}$$

for all values of i. Thus, if $\mathbf{H}$ is positive definite, all of its eigenvalues will be positive.

Conversely, assume that (174) holds. Then, for any vector, $\mathbf{v}$, we can make use of (5.38) to give

$$\begin{aligned}
\mathbf{v}^{\mathrm{T}}\mathbf{H}\mathbf{v} &= \left(\sum_i c_i\mathbf{u}_i\right)^{\mathrm{T}}\mathbf{H}\left(\sum_j c_j\mathbf{u}_j\right) \\
&= \left(\sum_i c_i\mathbf{u}_i\right)^{\mathrm{T}}\left(\sum_j \lambda_j c_j\mathbf{u}_j\right) \\
&= \sum_i \lambda_i c_i^2 > 0
\end{aligned}$$

where we have used (5.33) and (5.34) along with (174). Thus, if all of the eigenvalues are positive, the Hessian matrix will be positive definite.

4 Example 5.25

5.25 (⋆⋆⋆) **www** Consider a quadratic error function of the form

$$E = E_0 + \frac{1}{2}(\mathbf{w} - \mathbf{w}^\star)^{\mathrm{T}}\mathbf{H}(\mathbf{w} - \mathbf{w}^\star) \tag{5.195}$$

where $\mathbf{w}^\star$ represents the minimum, and the Hessian matrix $\mathbf{H}$ is positive definite and constant. Suppose the initial weight vector $\mathbf{w}^{(0)}$ is chosen to be at the origin and is updated using simple gradient descent

$$\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau-1)} - \rho\nabla E \tag{5.196}$$

where τ denotes the step number, and ρ is the learning rate (which is assumed to be small). Show that, after τ steps, the components of the weight vector parallel to the eigenvectors of $\mathbf{H}$ can be written

$$w_j^{(\tau)} = \{1 - (1 - \rho\eta_j)^\tau\}\, w_j^\star \tag{5.197}$$

where $w_j = \mathbf{w}^{\mathrm{T}}\mathbf{u}_j$, and $\mathbf{u}_j$ and η_j are the eigenvectors and eigenvalues, respectively, of $\mathbf{H}$ so that

$$\mathbf{H}\mathbf{u}_j = \eta_j\mathbf{u}_j. \tag{5.198}$$

Show that as $\tau \to \infty$, this gives $\mathbf{w}^{(\tau)} \to \mathbf{w}^\star$ as expected, provided $|1 - \rho\eta_j| < 1$. Now suppose that training is halted after a finite number τ of steps. Show that the components of the weight vector parallel to the eigenvectors of the Hessian satisfy

$$w_j^{(\tau)} \simeq w_j^\star \quad \text{when} \quad \eta_j \gg (\rho\tau)^{-1} \tag{5.199}$$

$$|w_j^{(\tau)}| \ll |w_j^\star| \quad \text{when} \quad \eta_j \ll (\rho\tau)^{-1}. \tag{5.200}$$

Compare this result with the discussion in Section 3.5.3 of regularization with simple weight decay, and hence show that $(\rho\tau)^{-1}$ is analogous to the regularization parameter λ. The above results also show that the effective number of parameters in the network, as defined by (3.91), grows as the training progresses.

Solution

5.25 The gradient of (5.195) is given

$$\nabla E = \mathbf{H}(\mathbf{w} - \mathbf{w}^\star)$$

and hence update formula (5.196) becomes

$$\mathbf{w}^{(\tau)} = \mathbf{w}^{(\tau-1)} - \rho\mathbf{H}(\mathbf{w}^{(\tau-1)} - \mathbf{w}^\star).$$

Pre-multiplying both sides with $\mathbf{u}_j^{\mathrm{T}}$ we get

$$\begin{aligned} w_j^{(\tau)} &= \mathbf{u}_j^{\mathrm{T}}\mathbf{w}^{(\tau)} && (186)\\ &= \mathbf{u}_j^{\mathrm{T}}\mathbf{w}^{(\tau-1)} - \rho\mathbf{u}_j^{\mathrm{T}}\mathbf{H}(\mathbf{w}^{(\tau-1)} - \mathbf{w}^\star) \\ &= w_j^{(\tau-1)} - \rho\eta_j\mathbf{u}_j^{\mathrm{T}}(\mathbf{w} - \mathbf{w}^\star) \\ &= w_j^{(\tau-1)} - \rho\eta_j(w_j^{(\tau-1)} - w_j^\star), && (187) \end{aligned}$$

where we have used (5.198). To show that

$$w_j^{(\tau)} = \{1 - (1 - \rho\eta_j)^\tau\}\, w_j^\star$$

for $\tau = 1, 2, \ldots$, we can use proof by induction. For $\tau = 1$, we recall that $\mathbf{w}^{(0)} = \mathbf{0}$ and insert this into (187), giving

$$\begin{aligned} w_j^{(1)} &= w_j^{(0)} - \rho\eta_j(w_j^{(0)} - w_j^\star) \\ &= \rho\eta_j w_j^\star \\ &= \{1 - (1 - \rho\eta_j)\}\, w_j^\star. \end{aligned}$$

Now we assume that the result holds for $\tau = N - 1$ and then make use of (187)

$$\begin{aligned} w_j^{(N)} &= w_j^{(N-1)} - \rho\eta_j(w_j^{(N-1)} - w_j^\star) \\ &= w_j^{(N-1)}(1 - \rho\eta_j) + \rho\eta_j w_j^\star \\ &= \left\{1 - (1 - \rho\eta_j)^{N-1}\right\} w_j^\star(1 - \rho\eta_j) + \rho\eta_j w_j^\star \\ &= \left\{(1 - \rho\eta_j) - (1 - \rho\eta_j)^N\right\} w_j^\star + \rho\eta_j w_j^\star \\ &= \left\{1 - (1 - \rho\eta_j)^N\right\} w_j^\star \end{aligned}$$

as required.

Provided that $|1 - \rho\eta_j| < 1$ then we have $(1 - \rho\eta_j)^\tau \to 0$ as $\tau \to \infty$, and hence $\left\{1 - (1 - \rho\eta_j)^N\right\} \to 1$ and $\mathbf{w}^{(\tau)} \to \mathbf{w}^\star$.

If τ is finite but $\eta_j \gg (\rho\tau)^{-1}$, τ must still be large, since $\eta_j\rho\tau \gg 1$, even though $|1 - \rho\eta_j| < 1$. If τ is large, it follows from the argument above that $w_j^{(\tau)} \simeq w_j^\star$.

If, on the other hand, $\eta_j \ll (\rho\tau)^{-1}$, this means that $\rho\eta_j$ must be small, since $\rho\eta_j\tau \ll 1$ and τ is an integer greater than or equal to one. If we expand,

$$(1 - \rho\eta_j)^\tau = 1 - \tau\rho\eta_j + O(\rho\eta_j^2)$$

and insert this into (5.197), we get

$$\begin{aligned} |w_j^{(\tau)}| &= |\{1-(1-\rho\eta_j)^\tau\}\, w_j^\star| \\ &= |\left\{1-(1-\tau\rho\eta_j+O(\rho\eta_j^2))\right\} w_j^\star| \\ &\simeq \tau\rho\eta_j|w_j^\star| \ll |w_j^\star| \end{aligned}$$

Recall that in Section 3.5.3 we showed that when the regularization parameter (called α in that section) is much larger than one of the eigenvalues (called λ_j in that section) then the corresponding parameter value w_i will be close to zero. Conversely, when α is much smaller than λ_i then w_i will be close to its maximum likelihood value. Thus α is playing an analogous role to $\rho\tau$.

5 Example 5.26

5.26 ($\star\star$) Consider a multilayer perceptron with arbitrary feed-forward topology, which is to be trained by minimizing the *tangent propagation* error function (5.127) in which the regularizing function is given by (5.128). Show that the regularization term Ω can be written as a sum over patterns of terms of the form

$$\Omega_n = \frac{1}{2}\sum_k (\mathcal{G}y_k)^2 \tag{5.201}$$

where $\mathcal{G}$ is a differential operator defined by

$$\mathcal{G} \equiv \sum_i \tau_i \frac{\partial}{\partial x_i}. \tag{5.202}$$

By acting on the forward propagation equations

$$z_j = h(a_j), \qquad a_j = \sum_i w_{ji} z_i \tag{5.203}$$

with the operator $\mathcal{G}$, show that Ω_n can be evaluated by forward propagation using the following equations:

$$\alpha_j = h'(a_j)\beta_j, \qquad \beta_j = \sum_i w_{ji}\alpha_i. \tag{5.204}$$

where we have defined the new variables

$$\alpha_j \equiv \mathcal{G}z_j, \qquad \beta_j \equiv \mathcal{G}a_j. \tag{5.205}$$

Now show that the derivatives of Ω_n with respect to a weight w_{rs} in the network can be written in the form

$$\frac{\partial \Omega_n}{\partial w_{rs}} = \sum_k \alpha_k \{\phi_{kr} z_s + \delta_{kr}\alpha_s\} \tag{5.206}$$

where we have defined

$$\delta_{kr} \equiv \frac{\partial y_k}{\partial a_r}, \qquad \phi_{kr} \equiv \mathcal{G}\delta_{kr}. \tag{5.207}$$

Write down the backpropagation equations for δ_{kr}, and hence derive a set of back-propagation equations for the evaluation of the ϕ_{kr}.

Solution

5.26 **NOTE**: In PRML, equation (5.201) should read

$$\Omega_n = \frac{1}{2}\sum_k (\mathcal{G}y_k)^2\Big|_{\mathbf{x}_n}.$$

In this solution, we will indicate dependency on $\mathbf{x}_n$ with a subscript n on relevant symbols.

Substituting the r.h.s. of (5.202) into (5.201) and then using (5.70), we get

$$\Omega_n = \frac{1}{2}\sum_k \left(\sum_i \tau_{ni}\frac{\partial y_{nk}}{\partial x_{ni}}\right)^2 \tag{188}$$

$$= \frac{1}{2}\sum_k \left(\sum_i \tau_{ni} J_{nki}\right)^2 \tag{189}$$

where J_{nki} denoted J_{ki} evaluated at $\mathbf{x}_n$. Summing (189) over n, we get (5.128).

By applying $\mathcal{G}$ from (5.202) to the equations in (5.203) and making use of (5.205) we obtain (5.204). From this, we see that β_{nl} can be written in terms of α_{ni}, which in turn can be written as functions of β_{ni} from the previous layer. For the input layer, using (5.204) and (5.205), we get

$$\begin{aligned}
\beta_{nj} &= \sum_i w_{ji}\alpha_{ni} \\
&= \sum_i w_{ji}\mathcal{G}x_{ni} \\
&= \sum_i w_{ji}\sum_{i'}\tau_{ni'}\frac{\partial x_{ni}}{\partial x_{ni'}} \\
&= \sum_i w_{ji}\tau_{ni}. \qquad (190)
\end{aligned}$$

Thus we see that, starting from (190), τ_n is propagated forward by subsequent application of the equations in (5.204), yielding the β_{nl} for the output layer, from which Ω_n can be computed using (5.201),

$$\Omega_n = \frac{1}{2}\sum_k (\mathcal{G}y_{nk})^2 = \frac{1}{2}\sum_k \alpha_{nk}^2.$$

Considering $\partial\Omega_n/\partial w_{rs}$, we start from (5.201) and make use of the chain rule, together with (5.52), (5.205) and (5.207), to obtain

$$\begin{aligned}\frac{\partial \Omega_n}{\partial w_{rs}} &= \sum_k (\mathcal{G}y_{nk})\,\mathcal{G}\left(\delta_{nkr} z_{ns}\right) \\ &= \sum_k \alpha_{nk}\left(\phi_{nkr} z_{ns} + \delta_{nkr}\alpha_{ns}\right).\end{aligned}$$

The backpropagation formula for computing δ_{nkr} follows from (5.74), which is used in computing the Jacobian matrix, and is given by

$$\delta_{nkr} = h'(a_{nr})\sum_l w_{lr}\delta_{nkl}.$$

Using this together with (5.205) and (5.207), we can obtain backpropagation equations for ϕ_{nkr},

$$\begin{aligned}\phi_{nkr} &= \mathcal{G}\delta_{nkr} \\ &= \mathcal{G}\left(h'(a_{nr})\sum_l w_{lr}\delta_{nkl}\right) \\ &= h''(a_{nr})\beta_{nr}\sum_l w_{lr}\delta_{nkl} + h'(a_{nr})\sum_l w_{lr}\phi_{nkl}.\end{aligned}$$

6 Example 5.29

5.29 (⋆) **WWW** Verify the result (5.141).

Solution

5.29 This is easily verified by taking the derivative of (5.138), using (1.46) and standard derivatives, yielding

$$\frac{\partial \Omega}{\partial w_i} = \frac{1}{\sum_k \pi_k \mathcal{N}(w_i|\mu_k,\sigma_k^2)}\sum_j \pi_j \mathcal{N}(w_i|\mu_j,\sigma_j^2)\frac{(w_i-\mu_j)}{\sigma^2}.$$

Combining this with (5.139) and (5.140), we immediately obtain the second term of (5.141).

7 Example 5.34

5.34 (⋆) WWW Derive the result (5.155) for the derivative of the error function with respect to the network output activations controlling the mixing coefficients in the mixture density network.

Solution

5.34 **NOTE**: In the 1st printing of PRML, the l.h.s. of (5.154) should be replaced with $\gamma_{nk} = \gamma_k(\mathbf{t}_n|\mathbf{x}_n)$. Accordingly, in (5.155) and (5.156), γ_k should be replaced by γ_{nk} and in (5.156), t_l should be t_{nl}.

We start by using the chain rule to write

$$\frac{\partial E_n}{\partial a_k^\pi} = \sum_{j=1}^{K} \frac{\partial E_n}{\partial \pi_j} \frac{\partial \pi_j}{\partial a_k^\pi}. \tag{193}$$

Note that because of the coupling between outputs caused by the softmax activation function, the dependence on the activation of a single output unit involves all the output units.

For the first factor inside the sum on the r.h.s. of (193), standard derivatives applied to the n^{th} term of (5.153) gives

$$\frac{\partial E_n}{\partial \pi_j} = -\frac{\mathcal{N}_{nj}}{\sum_{l=1}^{K} \pi_l \mathcal{N}_{nl}} = -\frac{\gamma_{nj}}{\pi_j}. \tag{194}$$

For the for the second factor, we have from (4.106) that

$$\frac{\partial \pi_j}{\partial a_k^\pi} = \pi_j(I_{jk} - \pi_k). \tag{195}$$

Combining (193), (194) and (195), we get

$$\begin{aligned} \frac{\partial E_n}{\partial a_k^\pi} &= -\sum_{j=1}^{K} \frac{\gamma_{nj}}{\pi_j} \pi_j(I_{jk} - \pi_k) \\ &= -\sum_{j=1}^{K} \gamma_{nj}(I_{jk} - \pi_k) = -\gamma_{nk} + \sum_{j=1}^{K} \gamma_{nj}\pi_k = \pi_k - \gamma_{nk}, \end{aligned}$$

where we have used the fact that, by (5.154), $\sum_{j=1}^{K} \gamma_{nj} = 1$ for all n.