ECE1513
Tutorial 3:
Selected Exercises from Chapter 4

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

September 21, 2023

1 Example 4.2

Consider the minimization of a sum-of-squares error function (4.15), and suppose that all of the target vectors in the training set satisfy a linear constraint

$$a^T t_n + b = 0$$

where t_n corresponds to the nth row of the matrix T in (4.15). Show that as a consequence of this constraint, the elements of the model prediction y(x) given by the least-squares solution (4.17) also satisfy this constraint, so that

$$a^T y(x) + b = 0$$

To do so, assume that one of the basis functions ϕ_0(x)=1 so that the corresponding parameter w_0 plays the role of a bias.

Solution

For the purpose of this exercise, we make the contribution of the bias weights explicit in (4.15), giving

$$E_D(\widetilde{\mathbf{W}}) = \frac{1}{2}\mathrm{Tr}\left\{(\mathbf{XW} + \mathbf{1}\mathbf{w}_0^{\mathrm{T}} - \mathbf{T})^{\mathrm{T}}(\mathbf{XW} + \mathbf{1}\mathbf{w}_0^{\mathrm{T}} - \mathbf{T})\right\}$$

We can take the derivative w.r.t. w0, giving

$$2N\mathbf{w}_0 + 2(\mathbf{XW} - \mathbf{T})^{\mathrm{T}}\mathbf{1}$$

Setting this to zero, and solving for w0, we obtain

$$\mathbf{w}_0 = \bar{\mathbf{t}} - \mathbf{W}^{\mathrm{T}}\bar{\mathbf{x}}$$

where

$$\bar{\mathbf{t}} = \frac{1}{N}\mathbf{T}^{\mathrm{T}}\mathbf{1} \quad \text{and} \quad \bar{\mathbf{x}} = \frac{1}{N}\mathbf{X}^{\mathrm{T}}\mathbf{1}.$$

If we substitute into E_D we get

$$E_D(\mathbf{W}) = \frac{1}{2}\mathrm{Tr}\left\{(\mathbf{X}\mathbf{W} + \overline{\mathbf{T}} - \overline{\mathbf{X}}\mathbf{W} - \mathbf{T})^{\mathrm{T}}(\mathbf{X}\mathbf{W} + \overline{\mathbf{T}} - \overline{\mathbf{X}}\mathbf{W} - \mathbf{T})\right\}$$

Setting the derivative of this w.r.t. W to zero we get

$$\mathbf{W} = (\widehat{\mathbf{X}}^{\mathrm{T}}\widehat{\mathbf{X}})^{-1}\widehat{\mathbf{X}}^{\mathrm{T}}\widehat{\mathbf{T}} = \widehat{\mathbf{X}}^{\dagger}\widehat{\mathbf{T}}$$

where we have defined $\widehat{\mathbf{X}} = \mathbf{X} - \overline{\mathbf{X}}$ and $\widehat{\mathbf{T}} = \mathbf{T} - \overline{\mathbf{T}}$.
Now consider the prediction for a new input vector $\mathbf{x}^\star$,

$$\begin{aligned}
\mathbf{y}(\mathbf{x}^\star) &= \mathbf{W}^{\mathrm{T}}\mathbf{x}^\star + \mathbf{w}_0 \\
&= \mathbf{W}^{\mathrm{T}}\mathbf{x}^\star + \bar{\mathbf{t}} - \mathbf{W}^{\mathrm{T}}\bar{\mathbf{x}} \\
&= \bar{\mathbf{t}} - \widehat{\mathbf{T}}^{\mathrm{T}}\left(\widehat{\mathbf{X}}^{\dagger}\right)^{\mathrm{T}}(\mathbf{x}^\star - \bar{\mathbf{x}}).
\end{aligned}$$

Then,

$$\mathbf{a}^{\mathrm{T}}\bar{\mathbf{t}} = \frac{1}{N}\mathbf{a}^{\mathrm{T}}\mathbf{T}^{\mathrm{T}}\mathbf{1} = -b.$$

Finally we can follow,

$$\begin{aligned}
\mathbf{a}^{\mathrm{T}}\mathbf{y}(\mathbf{x}^\star) &= \mathbf{a}^{\mathrm{T}}\bar{\mathbf{t}} + \mathbf{a}^{\mathrm{T}}\widehat{\mathbf{T}}^{\mathrm{T}}\left(\widehat{\mathbf{X}}^{\dagger}\right)^{\mathrm{T}}(\mathbf{x}^\star - \bar{\mathbf{x}}) \\
&= \mathbf{a}^{\mathrm{T}}\bar{\mathbf{t}} = -b,
\end{aligned}$$

since $\mathbf{a}^{\mathrm{T}}\widehat{\mathbf{T}}^{\mathrm{T}} = \mathbf{a}^{\mathrm{T}}(\mathbf{T} - \overline{\mathbf{T}})^{\mathrm{T}} = b(\mathbf{1} - \mathbf{1})^{\mathrm{T}} = \mathbf{0}^{\mathrm{T}}$.

2 Example 4.7

Show that the logistic sigmoid function (4.59) satisfies the property $\sigma(-a) = 1 - \sigma(a)$ and that its inverse is given by $\sigma^{-1}(y) = ln(\frac{y}{1-y})$

Solution

From (4.59) we have,

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

The inverse of the logistic sigmoid is easily found as follows

$$\begin{aligned} y = \sigma(a) &= \frac{1}{1+e^{-a}} \\ \Rightarrow \quad \frac{1}{y} - 1 &= e^{-a} \\ \Rightarrow \quad \ln\left\{\frac{1-y}{y}\right\} &= -a \\ \Rightarrow \quad \ln\left\{\frac{y}{1-y}\right\} &= a = \sigma^{-1}(y). \end{aligned}$$

3 Example 4.9

Consider a generative classification model for K classes defined by prior class probabilities p(C_k) = π_k and general class-conditional densities p($\phi \backslash C_k$) where ϕ is the input feature vector. Suppose we are given a training data set (ϕ_n, t_n) where n = 1,...,N, and t_n is a binary target vector of length K that uses the 1-of-K coding scheme, so that it has components $t_{nj} = I_{jk}$ if pattern n is from class C_k.
Assuming that the data points are drawn independently from this model, show that the maximum-likelihood solution for the prior probabilities is given by

$$\pi_k = \frac{N_k}{N}$$

where N_k is the number of data points assigned to class C_k.

Solution

The likelihood function is given by,

$$p\left(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}\right) = \prod_{n=1}^{N}\prod_{k=1}^{K}\{p(\phi_n|C_k)\pi_k\}^{t_{nk}}$$

Hence we can obtain the expression for the logarithm likelihood:

$$\ln p = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left[\ln\pi_k + \ln p(\boldsymbol{\phi_n}|C_k)\right] \propto \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln\pi_k$$

Since there is a constraint on π_k, so we need to add a Lagrange Multiplier to the expression, which becomes:

$$L = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\ln\pi_k + \lambda(\sum_{k=1}^{K}\pi_k - 1)$$

We calculate the derivative of the expression above with regard to π_k:

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^{N}\frac{t_{nk}}{\pi_k} + \lambda$$

And if we set the derivative equal to 0, we can obtain:

$$\pi_k = -(\sum_{n=1}^{N} t_{nk})/\lambda = -\frac{N_k}{\lambda} \qquad (*)$$

And if we preform summation on both sides with regard to k, we can see that:

$$1 = -(\sum_{k=1}^{K} N_k)/\lambda = -\frac{N}{\lambda}$$

Which gives $\lambda = -N$, and substitute it into $(*)$, we can obtain $\pi_k = N_k/N$.

4 Example 4.10

4.10 ($\star\star$) Consider the classification model of Exercise 4.9 and now suppose that the class-conditional densities are given by Gaussian distributions with a shared covariance matrix, so that

$$p(\boldsymbol{\phi}|\mathcal{C}_k) = \mathcal{N}(\boldsymbol{\phi}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}). \tag{4.160}$$

Show that the maximum likelihood solution for the mean of the Gaussian distribution for class $\mathcal{C}_k$ is given by

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} \boldsymbol{\phi}_n \tag{4.161}$$

which represents the mean of those feature vectors assigned to class $\mathcal{C}_k$. Similarly, show that the maximum likelihood solution for the shared covariance matrix is given by

$$\boldsymbol{\Sigma} = \sum_{k=1}^{K} \frac{N_k}{N} \mathbf{S}_k \tag{4.162}$$

where

$$\mathbf{S}_k = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} (\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}. \tag{4.163}$$

Thus $\boldsymbol{\Sigma}$ is given by a weighted average of the covariances of the data associated with each class, in which the weighting coefficients are given by the prior probabilities of the classes.

Solution

4.10 If we substitute (4.160) into (155) and then use the definition of the multivariate Gaussian, (2.43), we obtain

$$\ln p\left(\{\phi_n, \mathbf{t}_n\}|\{\pi_k\}\right) = \\ -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left\{\ln|\boldsymbol{\Sigma}| + (\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi} - \boldsymbol{\mu})\right\}, \quad (157)$$

where we have dropped terms independent of $\{\boldsymbol{\mu}_k\}$ and $\boldsymbol{\Sigma}$.

Setting the derivative of the r.h.s. of (157) w.r.t. $\boldsymbol{\mu}_k$, obtained by using (C.19), to zero, we get

$$\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k) = 0.$$

Making use of (156), we can re-arrange this to obtain (4.161).

Rewriting the r.h.s. of (157) as

$$-\frac{1}{2}b\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\left\{\ln|\boldsymbol{\Sigma}| + \mathrm{Tr}\left[\boldsymbol{\Sigma}^{-1}(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)(\boldsymbol{\phi} - \boldsymbol{\mu}_k)^{\mathrm{T}}\right]\right\},$$

we can use (C.24) and (C.28) to calculate the derivative w.r.t. $\boldsymbol{\Sigma}^{-1}$. Setting this to zero we obtain

$$\frac{1}{2}\sum_{n=1}^{N}\sum_{k}^{\mathrm{T}} t_{nk}\left\{\boldsymbol{\Sigma} - (\boldsymbol{\phi}_n - \boldsymbol{\mu}_n)(\boldsymbol{\phi}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}\right\} = 0.$$

Again making use of (156), we can re-arrange this to obtain (4.162), with $\mathbf{S}_k$ given by (4.163).

Note that, as in Exercise 2.34, we do not enforce that $\boldsymbol{\Sigma}$ should be symmetric, but simply note that the solution is automatically symmetric.

5 Example 4.17

Show that the derivatives of the softmax activation function (4.104), where the a_k are defined by (4.105), are given by (4.106).

Solution

We should discuss in two situations separately, namely $j = k$ and $j \neq k$. When $j \neq k$, we have:

$$\frac{\partial y_k}{\partial a_j} = \frac{-exp(a_k) \cdot exp(a_j)}{[\sum_j exp(a_j)]^2} = -y_k \cdot y_j$$

And when $j = k$, we have:

$$\frac{\partial y_k}{\partial a_k} = \frac{exp(a_k)\sum_j exp(a_j) - exp(a_k)exp(a_k)}{[\sum_j exp(a_j)]^2} = y_k - y_k^2 = y_k(1 - y_k)$$

Therefore, we can obtain:

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

Where I_{kj} is the elements of the indentity matrix.

6 Example 4.19

Write down expressions for the gradient of the log likelihood, as well as the corresponding Hessian matrix, for the probit regression model defined in Section 4.3.5. These are the quantities that would be required to train such a model using IRLS.

Solution

We write down the log likelihood.

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

Therefore, we can obtain:

$$\begin{aligned} \nabla_{\mathbf{w}} \ln p &= \frac{\partial \ln p}{\partial y_n} \cdot \frac{\partial y_n}{\partial a_n} \cdot \frac{\partial a_n}{\partial \mathbf{w}} \\ &= \sum_{n=1}^{N} (\frac{t_n}{y_n} - \frac{1 - t_n}{1 - y_n}) \Phi'(a_n) \boldsymbol{\phi_n} \\ &= \sum_{n=1}^{N} \frac{y_n - t_n}{y_n(1 - y_n)} \Phi'(a_n) \boldsymbol{\phi_n} \end{aligned}$$

Where we have used $y = p(t=1|a) = \Phi(a)$ and $a_n = \mathbf{w}^T\boldsymbol{\phi_n}$. According to (4.114), we can obtain:

$$\Phi'(a) = \mathcal{N}(\theta|0,1)\big|_{\theta=a} = \frac{1}{\sqrt{2\pi}}exp(-\frac{1}{2}a^2)$$

Hence, we can obtain:

$$\nabla_{\mathbf{w}} \ln p = \sum_{n=1}^{N} \frac{y_n - t_n}{y_n(1-y_n)} \frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\boldsymbol{\phi_n}$$

To calculate the Hessian Matrix, we need to first evaluate several derivatives.

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}}\{\frac{y_n - t_n}{y_n(1-y_n)}\} &= \frac{\partial}{\partial y_n}\{\frac{y_n - t_n}{y_n(1-y_n)}\} \cdot \frac{\partial y_n}{\partial a_n} \cdot \frac{\partial a_n}{\partial \mathbf{w}} \\
&= \frac{y_n(1-y_n) - (y_n - t_n)(1-2y_n)}{[y_n(1-y_n)]^2}\Phi'(a_n)\boldsymbol{\phi_n} \\
&= \frac{y_n^2 + t_n - 2y_n t_n}{y_n^2(1-y_n)^2}\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\boldsymbol{\phi_n}
\end{aligned}$$

And

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}}\{\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\} &= \frac{\partial}{\partial a_n}\{\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\}\frac{\partial a_n}{\partial \mathbf{w}} \\
&= -\frac{a_n}{\sqrt{2\pi}}exp(-\frac{a_n^2}{2})\boldsymbol{\phi_n}
\end{aligned}$$

Therefore, using the chain rule, we can obtain:

$$\begin{aligned}
\frac{\partial}{\partial \mathbf{w}}\{\frac{y_n - t_n}{y_n(1-y_n)}\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\} &= \frac{\partial}{\partial \mathbf{w}}\{\frac{y_n - t_n}{y_n(1-y_n)}\}\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}} + \frac{y_n - t_n}{y_n(1-y_n)}\frac{\partial}{\partial \mathbf{w}}\{\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\} \\
&= [\frac{y_n^2 + t_n - 2y_n t_n}{y_n(1-y_n)}\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}} - a_n(y_n - t_n)]\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}y_n(1-y_n)}\boldsymbol{\phi_n}
\end{aligned}$$

Finally if we perform summation over n, we can obtain the Hessian Matrix:

$$\begin{aligned}
\mathbf{H} &= \nabla\nabla_{\mathbf{w}} \ln p \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{w}}\{\frac{y_n - t_n}{y_n(1-y_n)}\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}}\} \cdot \boldsymbol{\phi_n} \\
&= \sum_{n=1}^{N} [\frac{y_n^2 + t_n - 2y_n t_n}{y_n(1-y_n)}\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}} - a_n(y_n - t_n)]\frac{exp(-\frac{a_n^2}{2})}{\sqrt{2\pi}y_n(1-y_n)}\boldsymbol{\phi_n}\boldsymbol{\phi_n}^T
\end{aligned}$$

7 Example 4.23

4.23 ($\star\star$) **www** In this exercise, we derive the BIC result (4.139) starting from the Laplace approximation to the model evidence given by (4.137). Show that if the prior over parameters is Gaussian of the form $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}, \mathbf{V}_0)$, the log model evidence under the Laplace approximation takes the form

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m})^{\mathrm{T}}\mathbf{V}_0^{-1}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m}) - \frac{1}{2}\ln|\mathbf{H}| + \text{const}$$

where $\mathbf{H}$ is the matrix of second derivatives of the log likelihood $\ln p(\mathcal{D}|\boldsymbol{\theta})$ evaluated at $\boldsymbol{\theta}_{\mathrm{MAP}}$. Now assume that the prior is broad so that $\mathbf{V}_0^{-1}$ is small and the second term on the right-hand side above can be neglected. Furthermore, consider the case of independent, identically distributed data so that $\mathbf{H}$ is the sum of terms one for each data point. Show that the log model evidence can then be written approximately in the form of the BIC expression (4.139).

Solution

4.23 **NOTE**: In the 1st printing of PRML, the text of the exercise contains a typographical error. Following the equation, it should say that $\mathbf{H}$ is the matrix of second derivatives of the *negative* log likelihood.

The BIC approximation can be viewed as a large N approximation to the log model evidence. From (4.138), we have

$$\begin{aligned}\mathbf{A} &= -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})p(\boldsymbol{\theta}_{\mathrm{MAP}}) \\ &= \mathbf{H} - \nabla\nabla \ln p(\boldsymbol{\theta}_{\mathrm{MAP}})\end{aligned}$$

and if $p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\mathbf{m}, \mathbf{V}_0)$, this becomes

$$\mathbf{A} = \mathbf{H} + \mathbf{V}_0^{-1}.$$

If we assume that the prior is broad, or equivalently that the number of data points is large, we can neglect the term $\mathbf{V}_0^{-1}$ compared to $\mathbf{H}$. Using this result, (4.137) can be rewritten in the form

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}}) - \frac{1}{2}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m})\mathbf{V}_0^{-1}(\boldsymbol{\theta}_{\mathrm{MAP}} - \mathbf{m}) - \frac{1}{2}\ln|\mathbf{H}| + \text{const} \tag{169}$$

as required. Note that the phrasing of the question is misleading, since the assumption of a broad prior, or of large N, is required in order to derive this form, as well as in the subsequent simplification.

We now again invoke the broad prior assumption, allowing us to neglect the second term on the right hand side of (169) relative to the first term.

Since we assume i.i.d. data, $\mathbf{H} = -\nabla\nabla \ln p(\mathcal{D}|\boldsymbol{\theta}_{\mathrm{MAP}})$ consists of a sum of terms, one term for each datum, and we can consider the following approximation:

$$\mathbf{H} = \sum_{n=1}^{N} \mathbf{H}_n = N\widehat{\mathbf{H}}$$

where $\mathbf{H}_n$ is the contribution from the n^{th} data point and

$$\widehat{\mathbf{H}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{H}_n.$$

Combining this with the properties of the determinant, we have

$$\ln|\mathbf{H}| = \ln|N\widehat{\mathbf{H}}| = \ln\left(N^M|\widehat{\mathbf{H}}|\right) = M\ln N + \ln|\widehat{\mathbf{H}}|$$

where M is the dimensionality of $\boldsymbol{\theta}$. Note that we are assuming that $\widehat{\mathbf{H}}$ has full rank M. Finally, using this result together (169), we obtain (4.139) by dropping the $\ln|\widehat{\mathbf{H}}|$ since this $O(1)$ compared to $\ln N$.

8 Example 4.25

4.25 ($\star\star$) Suppose we wish to approximate the logistic sigmoid $\sigma(a)$ defined by (4.59) by a scaled probit function $\Phi(\lambda a)$, where $\Phi(a)$ is defined by (4.114). Show that if λ is chosen so that the derivatives of the two functions are equal at $a = 0$, then $\lambda^2 = \pi/8$.

Solution

4.25 From (4.88) we have that

$$\begin{aligned} \left.\frac{\mathrm{d}\sigma}{\mathrm{d}a}\right|_{a=0} &= \sigma(0)(1-\sigma(0)) \\ &= \frac{1}{2}\left(1-\frac{1}{2}\right) = \frac{1}{4}. \end{aligned} \tag{170}$$

Since the derivative of a cumulative distribution function is simply the corresponding density function, (4.114) gives

$$\begin{aligned} \left.\frac{\mathrm{d}\Phi(\lambda a)}{\mathrm{d}a}\right|_{a=0} &= \lambda \mathcal{N}(0|0,1) \\ &= \lambda \frac{1}{\sqrt{2\pi}}. \end{aligned}$$

Setting this equal to (170), we see that

$$\lambda = \frac{\sqrt{2\pi}}{4} \quad \text{or equivalently} \quad \lambda^2 = \frac{\pi}{8}.$$

This is illustrated in Figure 4.9.