ECE1513
Tutorial 10:
Selected Exercises from Chapter 14

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

October 30, 2023

1 Example 14.1

(⋆⋆) WWW Consider a set models of the form $p(\mathbf{t}|\mathbf{x}, \mathbf{z}_h, \boldsymbol{\theta}_h, h)$ in which $\mathbf{x}$ is the input vector, $\mathbf{t}$ is the target vector, h indexes the different models, $\mathbf{z}_h$ is a latent variable for model h, and $\boldsymbol{\theta}_h$ is the set of parameters for model h. Suppose the models have prior probabilities $p(h)$ and that we are given a training set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\mathbf{T} = \{\mathbf{t}_1, \ldots, \mathbf{t}_N\}$. Write down the formulae needed to evaluate the predictive distribution $p(\mathbf{t}|\mathbf{x}, \mathbf{X}, \mathbf{T})$ in which the latent variables and the model index are marginalized out. Use these formulae to highlight the difference between Bayesian averaging of different models and the use of latent variables within a single model.

Solution

The required predictive distribution is given by

$$p(\mathbf{t}|\mathbf{x}, \mathbf{X}, \mathbf{T}) = \sum_h p(h) \sum_{\mathbf{z}_h} p(\mathbf{z}_h) \int p(\mathbf{t}|\mathbf{x}, \boldsymbol{\theta}_h, \mathbf{z}_h, h) p(\boldsymbol{\theta}_h|\mathbf{X}, \mathbf{T}, h) \, \mathrm{d}\boldsymbol{\theta}_h, \tag{356}$$

where

$$\begin{aligned} p(\boldsymbol{\theta}_h|\mathbf{X}, \mathbf{T}, h) &= \frac{p(\mathbf{T}|\mathbf{X}, \boldsymbol{\theta}_h, h) p(\boldsymbol{\theta}_h|h)}{p(\mathbf{T}|\mathbf{X}, h)} \\ &\propto p(\boldsymbol{\theta}|h) \prod_{n=1}^{N} p(\mathbf{t}_n|\mathbf{x}_n, \boldsymbol{\theta}, h) \\ &= p(\boldsymbol{\theta}|h) \prod_{n=1}^{N} \left(\sum_{\mathbf{z}_{nh}} p(\mathbf{t}_n, \mathbf{z}_{nh}|\mathbf{x}_n, \boldsymbol{\theta}, h) \right) \end{aligned} \tag{357}$$

The integrals and summations in (356) are examples of Bayesian averaging, accounting for the uncertainty about which model, h, is the correct one, the value of the corresponding parameters, $\boldsymbol{\theta}_h$, and the state of the latent variable, $\mathbf{z}_h$. The summation in (357), on the other hand, is an example of the use of latent variables, where different data points correspond to different latent variable states, although all the data are assumed to have been generated by a single model, h.

2 Example 14.3

(⋆) WWW By making use of Jensen's inequality (1.115), for the special case of the convex function $f(x) = x^2$, show that the average expected sum-of-squares error E_{AV} of the members of a simple committee model, given by (14.10), and the expected error E_{COM} of the committee itself, given by (14.11), satisfy

$$E_{\mathrm{COM}} \leqslant E_{\mathrm{AV}}. \tag{14.54}$$

Solution

We start by rearranging the r.h.s. of (14.10), by moving the factor $1/M$ inside the sum and the expectation operator outside the sum, yielding

$$\mathbb{E}_{\mathbf{x}}\left[\sum_{m=1}^{M}\frac{1}{M}\epsilon_m(\mathbf{x})^2\right].$$

If we then identify $\epsilon_m(\mathbf{x})$ and $1/M$ with x_i and λ_i in (1.115), respectively, and take $f(x) = x^2$, we see from (1.115) that

$$\left(\sum_{m=1}^{M}\frac{1}{M}\epsilon_m(\mathbf{x})\right)^2 \leqslant \sum_{m=1}^{M}\frac{1}{M}\epsilon_m(\mathbf{x})^2.$$

Since this holds for all values of $\mathbf{x}$, it must also hold for the expectation over $\mathbf{x}$, proving (14.54).

3 Example 14.5

(⋆⋆) WWW Consider a committee in which we allow unequal weighting of the constituent models, so that

$$y_{\mathrm{COM}}(\mathbf{x}) = \sum_{m=1}^{M} \alpha_m y_m(\mathbf{x}). \tag{14.55}$$

In order to ensure that the predictions $y_{\mathrm{COM}}(\mathbf{x})$ remain within sensible limits, suppose that we require that they be bounded at each value of $\mathbf{x}$ by the minimum and maximum values given by any of the members of the committee, so that

$$y_{\min}(\mathbf{x}) \leqslant y_{\mathrm{COM}}(\mathbf{x}) \leqslant y_{\max}(\mathbf{x}). \tag{14.56}$$

Show that a necessary and sufficient condition for this constraint is that the coefficients α_m satisfy

$$\alpha_m \geqslant 0, \qquad \sum_{m=1}^{M} \alpha_m = 1. \tag{14.57}$$

Solution

To prove that (14.57) is a sufficient condition for (14.56) we have to show that (14.56) follows from (14.57). To do this, consider a fixed set of $y_m(\mathbf{x})$ and imagine varying the α_m over all possible values allowed by (14.57) and consider the values taken by $y_{\mathrm{COM}}(\mathbf{x})$ as a result. The maximum value of $y_{\mathrm{COM}}(\mathbf{x})$ occurs when $\alpha_k = 1$ where $y_k(\mathbf{x}) \geqslant y_m(\mathbf{x})$ for $m \neq k$, and hence all $\alpha_m = 0$ for $m \neq k$. An analogous result holds for the minimum value. For other settings of $\boldsymbol{\alpha}$,

$$y_{\min}(\mathbf{x}) < y_{\mathrm{COM}}(\mathbf{x}) < y_{\max}(\mathbf{x}),$$

since $y_{\mathrm{COM}}(\mathbf{x})$ is a convex combination of points, $y_m(\mathbf{x})$, such that

$$\forall m : y_{\min}(\mathbf{x}) \leqslant y_m(\mathbf{x}) \leqslant y_{\max}(\mathbf{x}).$$

Thus, (14.57) is a sufficient condition for (14.56).

Showing that (14.57) is a necessary condition for (14.56) is equivalent to showing that (14.56) is a sufficient condition for (14.57). The implication here is that if (14.56) holds for any choice of values of the committee members $\{y_m(\mathbf{x})\}$ then (14.57) will be satisfied. Suppose, without loss of generality, that α_k is the smallest of the α values, i.e. $\alpha_k \leqslant \alpha_m$ for $k \neq m$. Then consider $y_k(\mathbf{x}) = 1$, together with $y_m(\mathbf{x}) = 0$ for all $m \neq k$. Then $y_{\min}(\mathbf{x}) = 0$ while $y_{\mathrm{COM}}(\mathbf{x}) = \alpha_k$ and hence from (14.56) we obtain $\alpha_k \geqslant 0$. Since α_k is the smallest of the α values it follows that all of the coefficients must satisfy $\alpha_k \geqslant 0$. Similarly, consider the case in which $y_m(\mathbf{x}) = 1$ for all m. Then $y_{\min}(\mathbf{x}) = y_{\max}(\mathbf{x}) = 1$, while $y_{\mathrm{COM}}(\mathbf{x}) = \sum_m \alpha_m$. From (14.56) it then follows that $\sum_m \alpha_m = 1$, as required.

4 Example 14.9

(⋆) WWW Show that the sequential minimization of the sum-of-squares error function for an additive model of the form (14.21) in the style of boosting simply involves fitting each new base classifier to the residual errors $t_n - f_{m-1}(\mathbf{x}_n)$ from the previous model.

Solution

The sum-of-squares error for the additive model of (14.21) is defined as

$$E = \frac{1}{2}\sum_{n=1}^{N}(t_n - f_m(\mathbf{x}_n))^2.$$

Using (14.21), we can rewrite this as

$$\frac{1}{2}\sum_{n=1}^{N}(t_n - f_{m-1}(\mathbf{x}_n) - \frac{1}{2}\alpha_m y_m(\mathbf{x}))^2,$$

where we recognize the two first terms inside the square as the residual from the $(m-1)$-th model. Minimizing this error w.r.t. $y_m(\mathbf{x})$ will be equivalent to fitting $y_m(\mathbf{x})$ to the (scaled) residuals.

5 Example 14.15

(⋆) WWW We have already noted that if we use a squared loss function in a regression problem, the corresponding optimal prediction of the target variable for a new input vector is given by the conditional mean of the predictive distribution. Show that the conditional mean for the mixture of linear regression models discussed in Section 14.5.1 is given by a linear combination of the means of each component distribution. Note that if the conditional distribution of the target data is multimodal, the conditional mean can give poor predictions.

Solution

The predictive distribution from the mixture of linear regression models for a new input feature vector, $\widehat{\boldsymbol{\phi}}$, is obtained from (14.34), with $\boldsymbol{\phi}$ replaced by $\widehat{\boldsymbol{\phi}}$. Calculating the expectation of t under this distribution, we obtain

$$\mathbb{E}[t|\widehat{\boldsymbol{\phi}}, \boldsymbol{\theta}] = \sum_{k=1}^{K} \pi_k \mathbb{E}[t|\widehat{\boldsymbol{\phi}}, \mathbf{w}_k, \beta].$$

Depending on the parameters, this expectation is potentially K-modal, with one mode for each mixture component. However, the weighted combination of these modes output by the mixture model may not be close to any single mode. For example, the combination of the two modes in the left panel of Figure 14.9 will end up in between the two modes, a region with no signicant probability mass.

6 Example 14.17

($\star\star$) WWW Consider a mixture model for a conditional distribution $p(t|\mathbf{x})$ of the form

$$p(t|\mathbf{x}) = \sum_{k=1}^{K} \pi_k \psi_k(t|\mathbf{x}) \tag{14.58}$$

in which each mixture component $\psi_k(t|\mathbf{x})$ is itself a mixture model. Show that this two-level hierarchical mixture is equivalent to a conventional single-level mixture model. Now suppose that the mixing coefficients in both levels of such a hierarchical model are arbitrary functions of $\mathbf{x}$. Again, show that this hierarchical model is again equivalent to a single-level model with $\mathbf{x}$-dependent mixing coefficients. Finally, consider the case in which the mixing coefficients at both levels of the hierarchical mixture are constrained to be linear classification (logistic or softmax) models. Show that the hierarchical mixture cannot in general be represented by a single-level mixture having linear classification models for the mixing coefficients. Hint: to do this it is sufficient to construct a single counter-example, so consider a mixture of two components in which one of those components is itself a mixture of two components, with mixing coefficients given by linear-logistic models. Show that this cannot be represented by a single-level mixture of 3 components having mixing coefficients determined by a linear-softmax model.

Solution

If we define $\psi_k(t|\mathbf{x})$ in (14.58) as

$$\psi_k(t|\mathbf{x}) = \sum_{m=1}^{M} \lambda_{mk}\phi_{mk}(t|\mathbf{x}),$$

we can rewrite (14.58) as

$$\begin{aligned} p(t|\mathbf{x}) &= \sum_{k=1}^{K} \pi_k \sum_{m=1}^{M} \lambda_{mk}\phi_{mk}(t|\mathbf{x}) \\ &= \sum_{k=1}^{K}\sum_{m=1}^{M} \pi_k\lambda_{mk}\phi_{mk}(t|\mathbf{x}). \end{aligned}$$

By changing the indexation, we can write this as

$$p(t|\mathbf{x}) = \sum_{l=1}^{L} \eta_l\phi_l(t|\mathbf{x}),$$

where $L = KM$, $l = (k-1)M + m$, $\eta_l = \pi_k\lambda_{mk}$ and $\phi_l(\cdot) = \phi_{mk}(\cdot)$. By construction, $\eta_l \geqslant 0$ and $\sum_{l=1}^{L} \eta_l = 1$.

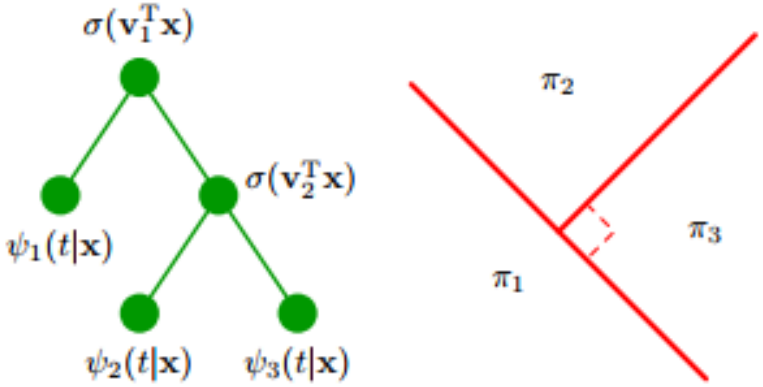
Note that this would work just as well if π_k and λ_{mk} were to be dependent on $\mathbf{x}$, as long as they both respect the constraints of being non-negative and summing to 1 for every possible value of $\mathbf{x}$.

Finally, consider a tree-structured, hierarchical mixture model, as illustrated in the left panel of Figure 12. On the top (root) level, this is a mixture with two components. The mixing coefficients are given by a linear logistic regression model and hence are input dependent. The left sub-tree correspond to a local conditional density model, $\psi_1(t|\mathbf{x})$. In the right sub-tree, the structure from the root is replicated, with the difference that both sub-trees contain local conditional density models, $\psi_2(t|\mathbf{x})$ and $\psi_3(t|\mathbf{x})$.

We can write the resulting mixture model on the form (14.58) with mixing coefficients

$$\begin{aligned} \pi_1(\mathbf{x}) &= \sigma(\mathbf{v}_1^{\mathrm{T}}\mathbf{x}) \\ \pi_2(\mathbf{x}) &= (1-\sigma(\mathbf{v}_1^{\mathrm{T}}\mathbf{x}))\sigma(\mathbf{v}_2^{\mathrm{T}}\mathbf{x}) \\ \pi_3(\mathbf{x}) &= (1-\sigma(\mathbf{v}_1^{\mathrm{T}}\mathbf{x}))(1-\sigma(\mathbf{v}_2^{\mathrm{T}}\mathbf{x})), \end{aligned}$$

Left: an illustration of a hierarchical mixture model, where the input dependent mixing coefficients are determined by linear logistic models associated with interior nodes; the leaf nodes correspond to local (conditional) density models. Right: a possible division of the input space into regions where different mixing coefficients dominate, under the model illustrated left.



where $\sigma(\cdot)$ is defined in (4.59) and $\mathbf{v}_1$ and $\mathbf{v}_2$ are the parameter vectors of the logistic regression models. Note that $\pi_1(\mathbf{x})$ is independent of the value of $\mathbf{v}_2$. This would not be the case if the mixing coefficients were modelled using a single level softmax model,

$$\pi_k(\mathbf{x}) = \frac{e^{\mathbf{u}_k^{\mathrm{T}}\mathbf{x}}}{\sum_j^3 e^{\mathbf{u}_j^{\mathrm{T}}\mathbf{x}}},$$

where the parameters $\mathbf{u}_k$, corresponding to $\pi_k(\mathbf{x})$, will also affect the other mixing coeffiecients, $\pi_{j\neq k}(\mathbf{x})$, through the denominator. This gives the hierarchical model different properties in the modelling of the mixture coefficients over the input space, as compared to a linear softmax model. An example is shown in the right panel of Figure 12, where the red lines represent borders of equal mixing coefficients in the input space. These borders are formed from two straight lines, corresponding to the two logistic units in the left panel of 12. A corresponding division of the input space by a softmax model would involve three straight lines joined at a single point, looking, e.g., something like the red lines in Figure 4.3 in PRML; note that a linear three-class softmax model could not implement the borders show in right panel of Figure 12.