Jiří Grim

Mixture of experts architectures for neural networks as a special case of conditional expectation formula

Kybernetika, Vol. 34 (1998), No. 4, [417]--422

Persistent URL: http://dml.cz/dmlcz/135225

KYBERNETIKA — VOLUME 34 (1998), NUMBER 4, PAGES 417 - 422

MIXTURE OF EXPERTS ARCHITECTURES FOR NEURAL NETWORKS AS A SPECIAL CASE OF CONDITIONAL EXPECTATION FORMULA

JIŘÍ GRIM[1]

Recently a new interesting architecture of neural networks called "mixture of experts" has been proposed as a tool of real multivariate approximation or prediction. We show that the underlying problem is closely related to approximating the joint probability density of involved variables by finite mixture. Particularly, assuming normal mixtures, we can explicitly write the conditional expectation formula which can be interpreted as a mixture-of-experts network. In this way the related optimization problem can be reduced to standard estimation of normal mixtures by means of EM algorithm. The resulting prediction is optimal in the sense of minimum dispersion if the assumed mixture model is true. It is shown that some of the recently published results can be obtained by specifying the normal components of mixtures in a special form.

1. INTRODUCTION

Mixture-of-experts architecture typically consists of two parallel feedforward networks having the same real input vector $\boldsymbol{x} \in R^N$: a network of "expert" units performing prediction of some output vector $\boldsymbol{y} \in R^K$ and a gating network which weights the outputs of expert units to form the overall output.

The original heuristic idea was to simplify e. g. a complex problem of linear regression by dividing the input space into smaller regions and solving separately the presumably less complex regression tasks within the input subsets. Thus, by proper switching between the regions, the global functioning could achieve the quality of locally optimal solutions ("local experts"). This original "divide and conquer" principle was generalized by introducing "soft" gating allowing for "soft" partitioning of the input space (cf. [5, 8, 13, 14]). The resulting prediction can be expressed by Eq.

$$\bar{\boldsymbol{y}}(\boldsymbol{x}) = \sum_{m=1}^{M} g_m(\boldsymbol{x})\, \hat{\boldsymbol{y}}(\boldsymbol{x}, \theta_m) \tag{1.1}$$

[1] Supported by the Grant of the Academy of Sciences No. A2075703, by the Grant of the Ministry of Education No. VS 96063 and partially by the Complex Research Project of the Academy of Sciences of the Czech Republic No. K1075601.

where $g_m(\boldsymbol{x})$ represent the M gating (weight-) functions and $\hat{\boldsymbol{y}}(\boldsymbol{x}, \theta_m)$ the corresponding "locally" optimal predictions. The composite prediction formula (1.1) has further been generalized by introducing hierarchical structures (cf. [6, 7]).

Optimization of the mixture-of-experts networks is a difficult problem. Roughly speaking, the recently published techniques combine EM algorithm with sofisticated gradient- and regression methods. In practical problems the reported computational results appear to be satisfactory (cf. e. g. [5, 8]).

From a statistical point of view the underlying problem can be formulated as a prediction of a real random vector $\mathbf{Y}$ given the value $\boldsymbol{x} \in R^N$ of a random vector $\mathbf{X}$. If the joint probability density functions $P(\boldsymbol{x}, \boldsymbol{y})$ is known then we can write the optimal minimum-dispersion prediction formula in terms of conditional expectation

$$\hat{\boldsymbol{y}}(\boldsymbol{x}) = E[\mathbf{Y}|\boldsymbol{x}] = \int \boldsymbol{y} P(\boldsymbol{y}|\boldsymbol{x})\, \mathrm{d}\boldsymbol{y}, \tag{1.2}$$

$$P(\boldsymbol{y}|\boldsymbol{x}) = \frac{P(\boldsymbol{x}, \boldsymbol{y})}{P(\boldsymbol{x})}, \quad P(\boldsymbol{x}) = \int P(\boldsymbol{x}, \boldsymbol{y})\, \mathrm{d}\boldsymbol{y}, \quad \boldsymbol{x} \in R^N. \tag{1.3}$$

It appears that the optimization methods for mixture-of-experts architectures locally approximate the conditional expectation (1.2) by means of different regression or gradient techniques. An alternative possibility is to approximate the unknown probability density function $P(\boldsymbol{x}, \boldsymbol{y})$ by a mixture. The corresponding conditional expectation can be expressed explicitely by Eq. similar to the composite prediction formula (1.1). Thus, if the assumed model is true, the involved parameters are optimal in the sense of minimum dispersion and can be easily derived from that of the estimated mixture. In this way the complex optimization problem can be reduced to standard estimation of mixtures by means of EM algorithm (cf [1, 2]). We show that some of the recently published mixture-of-experts architectures realize prediction equations which can be obtained as a special case of conditional expectation formula for normal mixtures.

The present approach can also be viewed as a modification of probabilistic neural networks based on distribution mixtures (cf. [3, 4]).

2. PREDICTION BASED ON NORMAL MIXTURES

We denote $\boldsymbol{z} = (\boldsymbol{x}^T, \boldsymbol{y}^T)^T$ the compound $(N + K)$-dimensional column vector and assume that the unknown probability density function $P(\boldsymbol{z})$ can be approximated by a normal mixture

$$P(\boldsymbol{z}) = \sum_{m \in \mathcal{M}} w_m F(\boldsymbol{z}|\mu_m, \Sigma_m), \quad \boldsymbol{z} \in R^{(N+K)}, \tag{2.1}$$

where $F(\boldsymbol{z}|\mu_m, \Sigma_m)$ are normal densities with the means μ_m and covariance matrices Σ_m, $m \in \mathcal{M}$, $\mathcal{M} = \{1, 2, \ldots, M\}$.

Considering the following partition of μ_m and Σ_m in accordance with the component vectors $\boldsymbol{x}$ and $\boldsymbol{y}$

$$\mu_m = \begin{pmatrix} \boldsymbol{c}_m \\ \boldsymbol{d}_m \end{pmatrix}, \quad \Sigma_m = \begin{pmatrix} A_m & V_m^T \\ V_m & B_m \end{pmatrix}, \tag{2.2}$$

we can easily verify the well known formula for the marginal density

$$P(\boldsymbol{x}) = \sum_{m \in \mathcal{M}} w_m G(\boldsymbol{x}|\boldsymbol{c}_m, A_m), \tag{2.3}$$

$$G(\boldsymbol{x}|\boldsymbol{c}_m, A_m) = \frac{1}{\sqrt{((2\pi)^N \det A_m)}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{c}_m)^T A_m^{-1}(\boldsymbol{x} - \boldsymbol{c}_m)\right\}$$

and for the conditional probability density

$$P(\boldsymbol{y}|\boldsymbol{x}) = \sum_{m \in \mathcal{M}} \gamma_m(\boldsymbol{x}) H(\boldsymbol{y}|\boldsymbol{u}_m, U_m), \tag{2.4}$$

$$H(\boldsymbol{y}|\boldsymbol{u}_m, U_m) = \frac{1}{\sqrt{((2\pi)^K \det U_m)}} \exp\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{u}_m)^T U_m^{-1}(\boldsymbol{y} - \boldsymbol{u}_m)\}.$$

Here $\boldsymbol{u}_m$ and U_m denote means and covariance matrices respectively

$$\boldsymbol{u}_m = \boldsymbol{d}_m + V_m A_m^{-1}(\boldsymbol{x} - \boldsymbol{c}_m), \quad U_m = B_m - V_m A_m^{-1} V_m^T \tag{2.5}$$

and $\gamma_m(\boldsymbol{x})$ are the conditional weights

$$\gamma_m(\boldsymbol{x}) = \frac{w_m G(\boldsymbol{x}|\boldsymbol{c}_m, A_m)}{\sum_{j \in \mathcal{M}} w_j G(\boldsymbol{x}|\boldsymbol{c}_j, A_j)}. \tag{2.6}$$

Making substitution in the prediction formula (1.2) we obtain

$$\hat{\boldsymbol{y}}(\boldsymbol{x}) = \sum_{m \in \mathcal{M}} \gamma_m(\boldsymbol{x}) \int \boldsymbol{y} H(\boldsymbol{y}|\boldsymbol{u}_m, U_m)\, \mathrm{d}\boldsymbol{y}$$

$$= \sum_{m \in \mathcal{M}} \gamma_m(\boldsymbol{x})[\boldsymbol{d}_m + V_m A_m^{-1}(\boldsymbol{x} - \boldsymbol{c}_m)]. \tag{2.7}$$

Let us recall that, in view of the composite prediction formula (1.1), the parenthesized linear expression corresponds to the locally optimal output of the mth expert unit which is weighted by the expression $\gamma_m(\boldsymbol{x})$ produced by a gating network. In terms of the original heuristic idea the "soft" weights $\gamma_m(\boldsymbol{x})$ divide the input space into "soft" hyperellipsoids to simplify the local regression tasks.

Let us note that the parameters of the prediction formula (2.7) directly follow from the estimated normal mixture (2.1) while other optimization methods (cf. e. g. [5, 8, 13, 14]) usually have to solve (weighted) least squares problem for each expert unit separately. In practical problems, the quality of results may be different because of the different criteria. However, it should be emphasized that in case of a true mixture model (2.1) Eq. (2.7) represents the optimal minimum-dispersion solution of the underlying problem.

3. COMPARISON WITH OTHER APPROACHES

Comparison of the conditional expectation formula (2.7) with some recently published results reveals considerable similarity. We show that the heuristically motivated solutions can be obtained usually as a special case of the Eq. (2.7) by choosing the components of the joint mixture (2.1) in a special form.

Thus e. g. Jacobs et al [5] (see also Xu and Jordan [13]) consider a mixture-of-experts architecture characterized by probabilistic weights $g_m(\boldsymbol{x})$ and by a constant local prediction θ_m:

$$\bar{\boldsymbol{y}}(\boldsymbol{x}) = \sum_{m=1}^{M} g_m(\boldsymbol{x})\theta_m, \qquad g_m(\boldsymbol{x}) = \frac{p_m(\boldsymbol{x})}{\sum_{i=1}^{M} p_i(\boldsymbol{x})}. \tag{3.1}$$

The parameters are optimized by applying gradient ascent to the criterion

$$\log\left[\sum_{m=1}^{M} g_m(\boldsymbol{x}) \exp\left\{-\frac{1}{2}\|\boldsymbol{y} - \theta_m\|^2\right\}\right]. \tag{3.2}$$

It can be seen that the prediction formula (3.1) corresponds to conditional expectation of the joint mixture

$$P(\boldsymbol{x}, \boldsymbol{y}) = \sum_{m=1}^{M} \frac{1}{M} p_m(\boldsymbol{x}) F(\boldsymbol{y}|\theta_m, \mathrm{II}) \tag{3.3}$$

having uniform weights and product components, whereby the densities $p_m(\boldsymbol{x})$ are not specified and $F(\boldsymbol{y}|\theta_m, \mathrm{II})$ denotes normal density $\mathcal{N}(\theta_m, \mathrm{II})$ with the mean θ_m and unity covariance matrix II.

Obviously the mixture model (3.3) is rather restrictive in comparison with the general normal mixture (2.1). Nevertheless, if appropriate and correctly estimated, it implies the optimal minimum-dispersion parameters of the prediction formula (3.1).

Xu et al [14] (see also Ramamurti and Ghosh [8]) consider a mixture-of-experts architecture with normal gating functions $g_m(\boldsymbol{x})$ and a linear local prediction formula $\hat{\boldsymbol{y}}(\boldsymbol{x}, \Theta_m) = \Theta_m \boldsymbol{x}$, i. e.

$$\bar{\boldsymbol{y}}(\boldsymbol{x}) = \sum_{m=1}^{M} g_m(\boldsymbol{x})(\Theta_m \boldsymbol{x}), \qquad g_m(\boldsymbol{x}) = \frac{w_m P(\boldsymbol{x}|\boldsymbol{c}_m, A_m)}{\sum_{j=1}^{M} w_j P(\boldsymbol{x}|\boldsymbol{c}_j, A_j)}. \tag{3.4}$$

Here $P(\cdot|\boldsymbol{c}_j, A_j) \approx \mathcal{N}(\boldsymbol{c}_j, A_j)$ are normal densities and Θ_m is a matrix. In [14] the optimization procedure is based on EM algorithm combined with iteratively reweighted least squares. On the other hand relation (3.4) can be obtained as conditional expectation from the general normal mixture

$$P(\boldsymbol{x}, \boldsymbol{y}) = \sum_{m=1}^{M} w_m P(\boldsymbol{x}, \boldsymbol{y}|\mu_m, \Sigma_m) \tag{3.5}$$

where

$$\Sigma_m = \begin{pmatrix} A_m & V_m^T \\ V_m & U_m \end{pmatrix}, \quad \mu_m = \begin{pmatrix} \boldsymbol{c}_m \\ \boldsymbol{d}_m \end{pmatrix}, \quad \Theta_m = V_m A_m^{-1}, \tag{3.6}$$

provided that

$$\boldsymbol{d}_m = V_m A_m^{-1} \boldsymbol{c}_m \quad \text{for all} \quad m = 1, 2 \ldots, M. \tag{3.7}$$

As the last conditions are generally not satisfied the composite prediction formula (3.4) would not be optimal because of the missing constant prediction term.

Let us remark further that, by grouping the components of the mixture $P(\boldsymbol{y}|\boldsymbol{x})$ (cf. (2.4)), we obtain prediction formula which corresponds to the hierarchical mixture-of-experts architecture. Indeed, considering a partition of the index set $\mathcal{M}$

$$\mathcal{M} = \bigcup_{j \in \mathcal{J}} \mathcal{M}_j, \quad \mathcal{J} = \{1, 2, \ldots, J\} \tag{3.8}$$

we may define the first- and second level weights $g_{jm}(\boldsymbol{x})$ and $g_j(\boldsymbol{x})$

$$g_{jm}(\boldsymbol{x}) = \frac{\gamma_m(\boldsymbol{x})}{g_j(\boldsymbol{x})}, \quad g_j(\boldsymbol{x}) = \sum_{m \in \mathcal{M}_j} \gamma_m(\boldsymbol{x}), \quad j \in \mathcal{J} \tag{3.9}$$

and finally we obtain the prediction formula

$$\widehat{\boldsymbol{y}}(\boldsymbol{x}) = \sum_{j \in \mathcal{J}} g_j(\boldsymbol{x}) \sum_{m \in \mathcal{M}_j} g_{jm}(\boldsymbol{x}) \left[\boldsymbol{d}_m + V_m A_m^{-1}(\boldsymbol{x} - \boldsymbol{c}_m)\right] \tag{3.10}$$

which is similar to that of Jordan et al (cf. [6], p. 185 and also [7]) and even more general in certain sense. However, in our case, this formula is equivalent to that of the nonhierarchical mixture of experts (2.7). From this point of view the usefulness of hierarchical mixtures of experts becomes questionable.

4. CONCLUDING REMARKS

Let us recall that all the parameters involved in the prediction formula (2.7) can be derived from the mixture (2.1) and therefore the optimization procedure can be reduced to estimating normal mixtures by means of EM algorithm (cf. [1, 2, 11, 12]).

Let us note that, in case of general covariance matrices, we can obtain ill conditioned matrices Σ_m in the iterative equations of EM algorithm. To avoid numerical problems we can remove the singular components or regularize the obtained matrices, e. g. by adding small positive constants to eigenvalues of matrices (to preserve the covariance structure of data). However, any such manipulation may violate the monotone convergence of EM algorithm in the immediately following iteration [2].

Another computational problem relates to estimation of the initial components of mixtures. In practical situations a method based on successive adding of components could be useful: For a given M we iterate the EM algorithm until reasonable convergence. Then we add a new sufficiently "flat" randomly placed component with a relatively high initial weight (e. g. $w_{M+1} = 0.5$). Continuing computation we obtain again a monotonely converging sequence of values of the log-likelihood function for

the new enlarged mixture. In this way there is a chance to find out the data regions not sufficiently covered by the previous set of component densities. The increased initial weight w_{M+1} helps the new component to "survive" in competition with the old well "fitted" components. The adding of components may be continued until the weight of the new component is repeatedly suppressed despite the increased initial value.

(Received December 18, 1997.)

REFERENCES

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B *39* (1977), 1–38.

[2] J. Grim: On numerical evaluation of maximum–likelihood estimates for finite mixtures of distributions. Kybernetika *18* (1982), 3, 173–190.

[3] J. Grim: Maximum likelihood design of layered neural networks. In: IEEE Proceedings of the 13th International Conference on Pattern Recognition, IEEE Press 1996, pp. 85–89.

[4] J. Grim: Design of multilayer neural networks by information preserving transforms. In: Proc. 3rd Systems Science European Congress (E. Pessa, M. B. Penna and A. Montesanto, eds.), Edizzioni Kappa, Roma 1996, pp. 977–982.

[5] R. A. Jacobs, M. I. Jordan, S. J. Nowlan and G. E. Hinton: Adaptive mixtures of local experts. Neural Comp. *3* (1991), 79–87.

[6] M. I. Jordan and R. A. Jacobs: Hierarchical mixtures of experts and the EM algorithm. Neural Comp. *6* (1994), 181–214.

[7] Ke Chen, Dahong Xie and Huisheng Chi: A modified HME architecture for text-dependent speaker identification. IEEE Trans. Neural Networks 7 (1996), 1309–1313.

[8] V. Ramamurti and J. Ghosh: Structural adaptation in mixtures of experts. In: IEEE Proceedings of the 13th International Conference on Pattern Recognition, IEEE Press, 1996, pp. 704–708.

[9] D. M. Titterington, A. F. M. Smith and U. E. Makov: Statistical Analysis of Finite Mixture Distributions. Wiley, Chichester – Singapore – New York 1985.

[10] I. Vajda: Theory of Statistical Inference and Information. Kluwer, Boston 1992.

[11] C. F. J. Wu: On the convergence properties of the EM algorithm. Ann. Statist. *11* (1983), 95–103.

[12] L. Xu and M. I. Jordan: On convergence properties of the EM algorithm for Gaussian mixtures. Neural Comp. *8* (1996), 129–151.

[13] L. Xu, M. I. Jordan and G. E. Hinton: A modified gating network for the mixtures of experts architecture. In: Proc. WCNN'94, San Diego 1994, Vol. 2, pp. 405–410.

[14] L. Xu, M. I. Jordan and G. E. Hinton: An alternative model for mixture of experts. In: Advances in Neural Information Processing Systems (G. Tesauro, D. S. Touretzky and T. K. Leen, eds.), MIT Press 1995, Vol. 7. pp. 633–640,

Ing. Jiří Grim, CSc., Institute of Information Theory and Automation – Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Praha 8. Czech Republic. e-mail: grim@utia.cas.cz