Approximation of conditional densities by smooth mixtures of regressions

Andriy Norets, Department of Economics, Princeton University

Abstract

This paper explores approximation properties of finite smooth mixtures of normal regressions as flexible models for conditional densities. These models are a special case of mixtures of experts (ME) introduced by Jacobs et al. (1991). ME have become increasingly popular in statistical literature since they are very flexible, easy to interpret, and reasonably easy to estimate. This paper contributes to the literature that provides a theoretical explanation of the success of ME models in applications. In particular, I show that large classes of conditional densities can be approximated in the Kullback–Leibler (KL) distance by finite smooth mixtures of normal regressions. Approximation results are obtained in the KL distance for the following reason. If a data generating density is in the KL closure of a class of models then this density can be consistently estimated from data by these models under weak regularity conditions. The results can be useful for establishing posterior consistency of certain Bayesian nonparametric models for conditional distributions.

Consider a joint probability distribution F on a product space $Y \times X$, $Y \subset R^d$ and $X \subset R^{d_x}$. Assume the conditional distribution $F(y|x)$ has a density $f(y|x)$ with respect to the Lebesgue measure. The marginal density of x with respect to some generic measure is denoted by $f(x)$. A model $\mathcal{M}$ for the conditional density $f(y|x)$ is described by $p(y|x,\mathcal{M})$. The KL distance between $f(y|x)f(x)$ and $p(y|x,\mathcal{M})f(x)$ is defined by

$$d_{KL}(F,\mathcal{M}) = \int \log \frac{f(y|x)}{p(y|x,\mathcal{M})} F(dy,dx).$$

This distance can also be interpreted as the expected KL distance between the conditional distributions. Either way, this is the distance useful for obtaining estimation consistency results. Also, convergence in the KL distance implies convergence in the total variation distance. In the paper, I consider several different specifications of mixture of normal regressions models, $p(y|x,\mathcal{M})$, and provide conditions on F under which $d_{KL}(F,\mathcal{M})$ can be made arbitrarily small. I also derive rates of convergence and easy to interpret bounds for $d_{KL}(F,\mathcal{M})$.

In general, a finite mixture of normal regressions model can be written as

$$p(y|x,\mathcal{M}) = \sum_{j=1}^{m} \alpha_j^m(x)\phi(y,\mu_j^m(x),\sigma_j^m(x))$$

where mixing probabilities satisfy $\alpha_j^m(x) \in [0,1]$ and $\sum_j \alpha_j^m(x) = 1$ and $\phi(y,\mu,\sigma)$ is a normal density with mean μ and standard deviation σ evaluated at y (if y is multidimensional then the variance-covariance matrix is diagonal $\sigma^2 I$). Most of the results obtained in the paper can be easily extended to models in which general location scale densities $\sigma^{-d}K((y-\mu)/\sigma)$ are mixed instead of the normal densities $\phi(y,\mu,\sigma)$. Models, in which the mixing weights depend on x, are referred in the paper as smooth mixtures. In practice, $\alpha_j^m(x)$'s are often modelled by a multinomial choice model, e.g., multinomial logit (Peng et al. (1996)) or probit (Geweke and Keane (2007)) or it might not depend on x. The mean $\mu_j^m(x)$ can be constant, linear or flexible, e.g., polynomial, in x. An exponentiated polynomial or spline in x can be used for modeling the standard deviation $\sigma_j^m(x)$ (Villani et al. (2009)).

To the best of my knowledge, previous literature on smooth mixtures of regressions (or experts) does not provide a theory on what specifications for α_j^m, μ_j^m, and σ_j^m deliver a model that can approximate and consistently estimate large nonparametric classes of densities F. There are theoretical results on approximation of smooth functions and estimation of conditional expectations by ME, see Zeevi et al. (1998) and Maiorov and Meir (1998). The only paper on approximation of conditional densities by ME seems to be Jiang and Tanner (1999) who develop approximation and estimation results for target densities from a single parameter exponential family, in which the parameter is a smooth function of covariates. In this paper, I do not restrict the functional form of $f(y|x)$ and use weak regularity conditions to describe a class of F that can be approximated. Conditions on approximable classes of $f(y|x)$ and $f(x)$ that are common for different model specifications include bounded support for $f(x)$, continuity of $f(y|x)$ in (y,x), finite expectation of a change of $\log f(y|x)$ in a neighborhood of y, and existence of the second moments of y. The latter restriction can be weakened by adding densities with fat tails to the mixtures in addition to normal densities.

Considerable flexibility is already attained when α_j^m's are modeled by multinomial logit with linear indices in x and (μ_j^m,σ_j^m) are independent of x. Using polynomials in the logit specification reduces the number of mixture components m required to achieve a specified approximation precision. Models for univariate response y in which the mixing probabilities and the variances of the mixed normals are independent of x and the means are flexible, e.g., polynomial in x, can

approximate large classes of $f(y|x)$. Differences in quantiles of $f(y|x)$ from these classes have to be bounded above and below uniformly in x. These restrictions on $f(y|x)$ can be weakened if the variances of the mixed normals are modeled by flexible functions of x.

Obtained approximation error bounds and convergences rates suggest that models with flexible mixing probabilities might perform better in practice than models with flexible means of the mixed normals and constant mixing probabilities. They also suggest that estimating conditional distributions directly is better than estimating the joint distributions first and then extracting the conditional distributions of interest.

Overall, the paper provides a number of encouraging approximation results for (smooth) mixtures of densities or experts, which might stimulate more theoretical and applied work in this area of research.

References

Geweke, J. and Keane, M. Smoothly mixing regressions. *Journal of Econometrics*, 138, 2007.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. Adaptive mixtures of local experts. *Neural Comput.*, 3(1):79–87, 1991. ISSN 0899-7667. doi: http://dx.doi.org/10.1162/neco.1991.3.1.79.

Jiang, W. and Tanner, M. Hierarchical mixtures-of-experts for exponential family regression models: Approximation and maximum likelihood estimation. *Annals of Statistics*, 27:987–1011, 1999.

Maiorov, V. and Meir, R. Approximation bounds for smooth functions in c(rd) by neural and mixture networks. *Neural Networks, IEEE Transactions on*, 9(5):969–978, 1998. ISSN 1045-9227. doi:10.1109/72.712173.

Peng, F., Jacobs, R. A., and Tanner, M. A. Bayesian inference in mixtures-of-experts and hierarchical mixtures-of-experts models with an application to speech recognition. *Journal of the American Statistical Association*, 91(435):953–960, 1996. ISSN 01621459.

Villani, M., Kohn, R., and Giordani, P. Regression density estimation using smooth adaptive gaussian mixtures. *Journal of Econometrics*, 153(2):155 – 173, 2009.

Zeevi, A., Meir, R., and Maiorov, V. Error bounds for functional approximation and estimation using mixtures of experts. *Information Theory, IEEE Transactions on*, 44(3):1010–1025, 1998. ISSN 0018-9448. doi:10.1109/18.669150.