A short note on the mixture of experts

Yen-Chi Chen
University of Washington
April 8, 2020

The mixture of expert is a popular approach in statistics and machine learning. It is similar but different from the usual mixture model (and the mixture of regression). Here we give a gentle introduction about this idea. For readers who are interested in more details, I would recommend the following book chapter:


Let \( Y \in \mathbb{R} \) be a continuous random variable that is our primary response variable and \( Z \in \{1, 2, \cdots, K\} \) be a discrete/categorical variable and \( X \in \mathbb{R}^d \) be a multivariate covariate. We only observe \((X, Y)\) and \( Z \) is unobserved; here \( Z \) is often refers to the latent class label or the label of an expert. In mixture models or mixture of experts, we often use a parametric form of the conditional densities. Depending on the relation among \( X, Y, Z \), there are 4 popular mixture-type models:

- **Mixture model.** In the usual mixture model, there is no covariate \( X \) so we only observe \( Y \). The mixture model can be written as a graphical model with a direct arrow \( Z \to Y \). Suppose we observe both \((Y, Z)\), then

\[
p(y, z) = p(y|z)p(z) = p_z(y)\pi_z \Rightarrow p(y) = \sum_k p_k(y)\pi_k,
\]

where \( p_k(y) \) is the conditional distribution of \( Y \) given \( Z = k \) and \( \pi_k = P(Z = k) \) is the proportion of the \( k \)-th component. Let \( \theta_k \) be the parameter of \( p_k(y) \), then the marginal distribution is

\[
p(y; \theta) = \sum_k p(y; \theta_k)\pi_k,
\]

which is the usual mixture model. The Gaussian mixture model is that each \( p(y; \theta_k) \) is a Gaussian, i.e., \( p(y; \theta_k) = p(y; \mu_k, \sigma_k^2) \), where \( \mu_k \) and \( \sigma_k^2 \) is the mean and variance of \( k \)-th component.

- **Mixture of expert.** In the mixture of expert, the model can be expressed as a graphical model with two arrows \( X \to Z \) and \( Z \to Y \). Note that \( Z \) is unobserved—we only observe \( X, Y \). In this case,

\[
p(x, y, z) = p(y|z)p(z|x)p(x) = p_z(y)p(x) \Rightarrow p(y, z|x) = p_z(y)\pi_z(x) \Rightarrow p(y|x) = \sum_k p_k(y)\pi_k(x).
\]

Namely, in the mixture of expert, the density of \( Y \) at each component remains the same across different \( X \). What changes with respect to \( X \) is the proportion \( \pi_k(x) \).

In this case, we need parameters for both \( p_k(y) \) and \( \pi_k(x) \), which leads to

\[
p(y|x; \theta, \eta) = \sum_k p(y; \theta_k)\pi_k(x; \eta).
\]
A popular model is place a Gaussian model over \( p(y; \theta_k) \) and a logistic model of \( \pi_k(x; \eta) \), i.e.,

\[
\pi_k(x; \eta) = \frac{\exp(\eta_{0,k} + \eta_{1,k}^T x)}{\sum_m \exp(\eta_{0,m} + \eta_{1,m}^T x)}.
\]

- **Mixture of regression.** The mixture of regression (a.k.a. regression mixture) looks very similar to the mixture of expert from a graphical perspective. The mixture of regression has two arrows: \( X \rightarrow Y \) and \( Z \rightarrow Y \). This, the difference compared to the mixture of expert is that the arrow \( X \rightarrow Z \) becomes \( X \rightarrow Y \). In this case,

\[
p(x, y, z) = p(y|x, z)p(z)p(x) = p_z(y|x)\pi_z(x) \Rightarrow p(y|x) = p_z(y|x)\pi_z(x)
\]

\[
\Rightarrow p(y|x) = \sum_k p_k(y|x)\pi_k.
\]

In particular, the conditional mean (regression function) becomes

\[
m(x) = \mathbb{E}(Y|X = x) = \int \sum_k p_k(y|x)\pi_k dy = \sum_k \pi_k \cdot m_k(x),
\]

where \( m_k(x) = \mathbb{E}(Y|Z = k, X = x) \) is the regression function of the \( k \)-th component. So the regression function is written as a mixture of several regression function. Note that the proportion \( \pi_k \) is independent of \( X \).

- **Mixture of expert regression.** The mixture of expert and the mixture of regression can be combined into the mixture of expert regression. It corresponds to the graph with three arrows: \( X \rightarrow Y \), \( X \rightarrow Z \), and \( Z \rightarrow Y \). In this case,

\[
p(x, y, z) = p(y|x, z)p(z|x)p(x) = p_z(y|x)\pi_z(x)p(x) \Rightarrow p(y,z|x) = p_z(y|x)\pi_z(x)
\]

\[
\Rightarrow p(y|x) = \sum_k p_k(y|x)\pi_k(x).
\]

The conditional mean (regression function) is

\[
m(x) = \mathbb{E}(Y|X = x) = \int \sum_k p_k(y|x)\pi_k(x)dy = \sum_k \pi_k(x) \cdot m_k(x).
\]

So it is the mixture of regression with the proportion \( \pi_k(x) \) being allowed to change with respect to \( x \).

1 **Mixture of expert**

In what follows, we will focus on the mixture of expert. Recall that in the mixture of expert,

\[
p(y|x; \theta, \eta) = \sum_k p(y; \theta_k)p_k(x; \eta)
\]

and what we observe is

\[(X_1, Y_1), \cdots, (X_n, Y_n).\]
The goal is to estimate $\theta$ and $\eta$ from the observed data.

A simple way to estimate these parameters is based on the maximum likelihood (ML) approach. For a single observation $X_i, Y_i$, the likelihood function is

$$ L(\theta, \eta|X_i, Y_i) = \sum_k p(Y_i; \theta_k) \pi_k(X_i; \eta) $$

and the log-likelihood is

$$ \ell(\theta, \eta|X_i, Y_i) = \log \left( \sum_k p(Y_i; \theta_k) \pi_k(X_i; \eta) \right) $$

The MLE (maximum likelihood estimator) is

$$ \hat{\theta}, \hat{\eta} = \arg\max_{\theta, \eta} \frac{1}{n} \sum_{i=1}^n \ell(\theta, \eta|X_i, Y_i). $$

## 2 EM algorithm

Although the MLE is well-defined, it is often hard to compute due to the fact that it does not have a closed-form in general. So we often need to use the EM-algorithm to numerically find the MLE. An introduction on the procedure of the EM is given in: [http://faculty.washington.edu/yenchic/19A_stat535/Lec13_EM_SGD.pdf](http://faculty.washington.edu/yenchic/19A_stat535/Lec13_EM_SGD.pdf). Starting with an initial guess $(\theta^{(0)}, \eta^{(0)})$, the EM algorithm creates a sequence of parameters

$$(\theta^{(0)}, \eta^{(0)}), (\theta^{(1)}, \eta^{(1)}), \ldots, (\theta^{(t)}, \eta^{(t)}), (\theta^{(t+1)}, \eta^{(t+1)}), \ldots$$

such that the likelihood function

$$ \sum_{i=1}^n \ell(\theta^{(t+1)}, \eta^{(t+1)}|X_i, Y_i) \geq \sum_{i=1}^n \ell(\theta^{(t)}, \eta^{(t)}|X_i, Y_i). $$

A key quantity in the EM algorithm is the complete-data likelihood—the likelihood function when the latent variable $Z$ is also observed:

$$ L_{\text{comp}}(\theta, \eta|X,Y,Z) = \prod_k [p(Y; \theta_k) \pi_k(X; \eta)]^{I(Z=k)} $$

and $\ell_{\text{comp}}(\theta, \eta|X,Y,Z) = \log L_{\text{comp}}(\theta, \eta|X,Y,Z)$. Given a complete-data likelihood and a previous parameter $(\theta^{(t)}, \eta^{(t)})$, we define the $Q$ function in the EM algorithm:

$$ Q(\theta, \eta; \theta^{(t)}, \eta^{(t)}|X,Y) = \mathbb{E}(\ell_{\text{comp}}(\theta, \eta|X,Y,Z)|X,Y; \theta^{(t)}, \eta^{(t)}) $$

$$ = \mathbb{E} \left( \sum_k I(Z = k) \log[p(Y; \theta_k) \pi_k(X; \eta)] \bigg| X,Y; \theta^{(t)}, \eta^{(t)} \right) $$

$$ = \sum_k \omega_k(X,Y; \theta^{(t)}, \eta^{(t)}) \log[p(Y; \theta_k) \pi_k(X; \eta)], $$

$$ \omega_k(X,Y; \theta^{(t)}, \eta^{(t)}) = \frac{p(Z = k|X,Y; \theta^{(t)}, \eta^{(t)})}{\sum_m p(Y; \theta_m^{(t)}) \pi_m(X; \eta^{(t)})}. $$
With this, we can write down the E-step and the M-step in the algorithm:

- **E-step.** Compute
  \[
  \omega_k(X_i; Y_i; \theta^{(t)}, \eta^{(t)}) = \frac{p(Y_i; \theta_k^{(t)}) \pi_k(X_i; \eta^{(t)})}{\sum_m p(Y_i; \theta_m^{(t)}) \pi_m(X_i; \eta^{(t)})}
  \]
  and
  \[
  Q_n(\theta, \eta; \theta^{(t)}, \eta^{(t)}) = \frac{1}{n} \sum_{i=1}^n Q(\theta, \eta; \theta^{(t)}, \eta^{(t)}|X_i, Y_i)
  \]
  \[
  Q(\theta, \eta; \theta^{(t)}, \eta^{(t)}|X_i, Y_i) = \sum_k \omega_k(X_i, Y_i; \theta^{(t)}, \eta^{(t)}) \left( \log p(Y_i; \theta_k) + \log \pi_k(X_i; \eta) \right).
  \]

- **M-step.** We update \(\theta, \eta\) using
  \[
  \theta^{(t+1)}, \eta^{(t+1)} = \arg\max_{\theta, \eta} Q_n(\theta, \eta; \theta^{(t)}, \eta^{(t)}).
  \]
  A nice property of this maximization is that \(\theta\) and \(\eta\) can be maximized separately and each componentwise parameter \(\theta_k\) can also be maximized individually:
  \[
  \theta_k^{(t+1)} = \arg\max_{\theta_k} Q_{k,n}(\theta_k; \theta^{(t)}, \eta^{(t)})
  \]
  \[
  Q_{k,n}(\theta_k; \theta^{(t)}, \eta^{(t)}) = \frac{1}{n} \sum_{i=1}^n \omega_k(X_i, Y_i; \theta^{(t)}, \eta^{(t)}) \log p(Y_i; \theta_k)
  \]
  and
  \[
  \eta^{(t+1)} = \arg\max_{\eta} Q_n(\eta; \theta^{(t)}, \eta^{(t)})
  \]
  \[
  Q_n(\eta; \theta^{(t)}, \eta^{(t)}) = \frac{1}{n} \sum_{i=1}^n \sum_k \omega_k(X_i, Y_i; \theta^{(t)}, \eta^{(t)}) \log \pi_k(Y_i; \eta).
  \]

Note that the EM algorithm suffers from the same problem of being stuck in a local maximum. Thus, multiple random initializations are often needed to increase the chance of getting the MLE.

### 3 Remarks

Here are some remarks about the mixture of expert method.

- **Common choice of the parametric model.** A popular choice is \(p(y; \theta_k) = \phi(y; \mu_k, \sigma_k^2)\), where \(\phi(y; \mu, \sigma^2)\) is the normal density with mean \(\mu\) and variance \(\sigma^2\) and \(\pi_k(x; \eta) = \frac{\exp(x^T \eta_k)}{\sum_{\eta} \exp(x^T \eta)}\), where \(x = (1, x) \in \mathbb{R}^{d+1}\) is the augmented covariate with the interception term. Note that although here we assume a univariate response \(Y \in \mathbb{R}\), the whole model can be easily generalized to multivariate response \(Y \in \mathbb{R}^p\).
• **Identifiability.** Model identifiability is often a problem in the mixture model and so is the mixture of experts. The label switching would occur if we do not place constraint over parameters $\theta_1, \cdots, \theta_k$. Consider a simple mixture of experts model with two experts:

$$p(y|x; \theta, \eta) = p(y; \theta_1)\pi_1(x; \eta) + p(y; \theta_2)\pi_2(x; \eta) = p(y; \theta'_1)\pi_1(x; \eta) + p(y; \theta'_2)\pi_2(x; \eta)$$

if we choose $\theta'_1 = \theta_2$ and $\theta'_2 = \theta_1$. Thus, $(\theta', \eta) \neq (\theta, \eta)$ but the probability model is the same. This also implies that the MLE will not be unique (since we can permute the parameters). A common approach to resolve this is to enforce some ordering among parameters.

• **Asymptotic theory.** The asymptotic theory follows from the regular MLE theory and we can construct confidence intervals using either a sandwich estimator of the underlying variance or a bootstrap approach.

• **Choice of number of experts $K$.** In general, the choice of number of experts is similar to the problem of choosing the number of mixture components in a mixture model. Common approaches such as AIC, BIC are often used. Note that if the problem is written as a prediction problem (given $X$, we use mixture of expert to predict $Y$), we may also use the cross-validation approach.

• **Bayesian approach and variational inference.** It is possible to use a Bayesian approach in the mixture of experts. The following paper discussed this idea along with variational inference: