STAT 535: Statistical Machine Learning **Autumn 2024**

Lecture 3: Mixture models and variational inference

Instructor: Yen-Chi Chen

3.1 Mixture models

While the parametric model is featured with a fast convergence rate and interpretability, it has a limitation that the real data is generally not so well-approximated by a parametric model. The mixture model is a popular approach that generalizes conventional parametric models to a flexible class of models that can fit to the data better.

Again, we consider random sample $X_1, \cdots, X_n \sim p$, where p is the underlying PDF. Consider a parametric model $p(x;\theta)$ with $\theta \in \Theta$. Instead of using a single parametric model, the mixture model consists of K models and considers the following PDF:

$$p(x;\eta) = \sum_{\ell=1}^{K} \pi_\ell p(x;\theta_\ell), \tag{3.1}$$

where $\theta_1, \cdots, \theta_K \in \Theta$ are K parameters and $\pi_1, \cdots, \pi_K \geq 0$ with $\sum_{\ell=1}^{K} \pi_\ell = 1$. The parameter π_ℓ is the weight/proportion of the ℓ-th component. The quantity

$$\eta = (\theta_1, \pi_1, \cdots, \theta_K, \pi_K)$$

is the set of all parameters in this model. The PDF in equation (3.1) is called a K-mixture model. As can be seen easily, the PDF $p(x;\eta)$ is indeed a mixture of K individual parametric model.

The mixture model can be generated by the following two-stage procedure. First, we generate a discrete random variable $Z \in \{1, 2, 3, \cdots, K\}$ such that $P(Z = \ell) = \pi_\ell$. Then we generate $X|Z \sim p(x;\theta_Z)$. The marginal PDF of X is $p(x;\eta)$.

Thus, a random variable following $p(x;\eta)$ can always be viewed as a random variable sampling from the above two-stage procedure and the variable Z is unobserved. Since Z is unobserved, it is often called a *latent/hidden* variable.

Example (Gaussian mixture). One of the most famous example is the Gaussian mixture model (GMM). An example of 3-GMM ($K = 3$) is the case where

$$p(x;\eta) = \pi_1 \phi(x;\mu_1,\sigma_1^2) + \pi_2 \phi(x;\mu_2,\sigma_2^2) + \pi_3 \phi(x;\mu_3,\sigma_3^2),$$

where $\phi(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$ is the PDF of $N(\mu,\sigma^2)$. In this case, the parameter

$$\eta = (\mu_1, \sigma_1^2, \pi_1, \mu_2, \sigma_2^2, \pi_2, \mu_3, \sigma_3^2, \pi_3).$$

Sampling from this Gaussian mixture can be done by first generating Z from

$$P(Z = \ell) = \pi_\ell, \qquad \ell = 1, 2, 3,$$

and then sampling

$$X|Z \sim N(\mu_Z, \sigma_Z^2).$$

Example (kernel density estimator). The kernel density estimator (KDE) is one of the most popular nonparametric density estimator. It can be viewed as a special type of mixture models under some choices of kernel functions. To see this we consider Gaussian KDE. Given observations $X_1, \cdots, X_n$, the KDE is

$$\widehat{p}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right),$$

where $K(x)$ is a smooth function known as the kernel function and $h > 0$ is the smoothing bandwidth. Suppose we choose $K(x)$ to be the Gaussian kernel, i.e., $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$. Then the KDE is

$$\begin{aligned}
\widehat{p}_h(x) &= \frac{1}{nh} \sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X_i - x}{h}\right)^2} \\
&= \sum_{i=1}^{n} \frac{1}{n} \frac{1}{\sqrt{2\pi h^2}} e^{-\frac{1}{2h^2}(X_i - x)^2} \\
&= \sum_{i=1}^{n} \frac{1}{n} \phi(x; X_i, h).
\end{aligned}$$

So it is an n-Gaussian mixture with each component's mean being $\mu_1 = X_1, \cdots, \mu_n = X_n$ and variance being h^2 and the weight being $\frac{1}{n}$. The fact that we can generate from a mixture model implies a simple approach of generating from the KDE $\widehat{p}_h(x)$. Using samples from the KDE is called the *smooth bootstrap* method.

Remark (continuous mixture). Here is an interesting way of re-expressing the mixture model. Let $H(z)$ be a probability measure that place discrete probability mass π_ℓ at value $z = \ell$. Then equation (3.1) can be written as

$$p(x; \eta) = \int p(x; \theta_z) H(dz).$$

As a result, if we rewrite the model $p(x; \theta_z)$ as $p(x|z; \theta)$, then we obtain

$$p(x; \theta) = \int p(x|z; \theta) H(dz).$$

This allows us to assume that $H(z)$ has a PDF $p(z)$, which leads to

$$p(x; \theta) = \int p(x|z; \theta) p(z) dz,$$

which is known as a continuous mixture model and is related to Bayesian inference and measurement error models.

3.2 EM algorithm

When we specify a mixture model $p(x; \eta)$, the parameter η is often unknown to us and has to be estimated from the data $X_1, \cdots, X_n$. A classical approach of estimating η is via the MLE. Namely,

$$\widehat{\eta} = \mathsf{argmax}_\eta \ell(\eta | X_1, \cdots, X_n) = \mathsf{argmax}_\eta \sum_{i=1}^{n} \ell(\eta | X_i) = \mathsf{argmax}_\eta \sum_{i=1}^{n} \log\left(\sum_{\ell=1}^{K} \pi_\ell p(X_i; \theta_\ell)\right).$$

While the theoretical analysis of this MLE follows from the usual MLE theory, the MLE generally does not have a closed-form. So we need a numerical method to compute the MLE.

The EM algorithm (expectation-maximization algorithm) is one of the most popular algorithm for computing the MLE. It is an iterative algorithm that starts with an initial guess of the MLE $\widehat{\eta}^{(0)}$ and then gradually improves until convergence.

To derive the EM algorithm, we first recall the latent variable representation of this problem. Under the mixture model, each X_i can be viewed as an outcome from the two-stage sampling procedure. So X_i is associated with a latent variable $Z_i \in \{1, 2, \cdots, K\}$. When both X_i and Z_i can be observed, the joint PDF of (x, z) is (recall $\eta = (\theta, \pi)$)

$$\begin{aligned} p(x, z; \eta) &= \sum_{\ell=1}^{K} I(z = \ell) \cdot \pi_\ell \cdot p(x; \theta_\ell) \\ &= \prod_{\ell=1}^{K} [\pi_\ell \cdot p(x; \theta_\ell)]^{I(z=\ell)}. \end{aligned}$$

Thus, the *complete-data* log-likelihood is

$$\begin{aligned} \ell_c(\eta | X_i = x_i, Z_i = z_i) &= \log p(x_i, z_i; \eta) \\ &= \sum_{\ell=1}^{K} I(z_i = \ell) \log[p(x_i; \eta | z_i = \ell) p(z_i; \eta)] \\ &= \sum_{\ell=1}^{K} I(z_i = \ell) \log\left(p(x; \theta_\ell)\right) + \sum_{\ell=1}^{K} I(z_i = \ell) \log \pi_\ell. \end{aligned} \tag{3.2}$$

Let $\widehat{\eta}^{(t)}$ be a prior guess of the parameter. Given the observation X_i, the expected complete-likelihood using this prior guess of parameter will be

$$Q(\eta | X_i; \widehat{\eta}^{(t)}) = \mathbb{E}_{Z_i \sim p(z | X_i; \widehat{\eta}^{(t)})}(\ell_c(\eta | X_i, Z_i) | X_i).$$

So with all observations $X_1, \cdots, X_n$, we define the Q-function

$$Q_n(\eta; \widehat{\eta}^{(t)}) = \sum_{i=1}^{n} Q(\eta | X_i; \widehat{\eta}^{(t)}).$$

We then improve the current estimate $\widehat{\eta}^{(t)}$ to be

$$\widehat{\eta}^{(t+1)} = \mathsf{argmax}_\eta Q_n(\eta; \widehat{\eta}^{(t)}).$$

While the above procedure may seem to be complicated, here is an elegant way to write it as a two-step updating procedure.

- **E-step.** Given X_i and a prior step parameter $\widehat{\eta}^{(t)} = (\widehat{\pi}_\ell^{(t)}, \widehat{\theta}_\ell^{(t)} : \ell = 1, \cdots, K)$, we first compute

$$\tau_{i,\ell}^{(t)} = P(Z_i = \ell | X_i; \widehat{\eta}^{(t)}) = \frac{\widehat{\pi}_\ell^{(t)} p(X_i; \widehat{\theta}_\ell^{(t)})}{\sum_{j=1}^{K} \widehat{\pi}_j^{(t)} p(X_i; \widehat{\theta}_j^{(t)})}. \tag{3.3}$$

  This is because the complete data log-likelihood in equation (3.2) only involves indicator functions of Z_i. So the conditional expectation step in each $Q(\eta | X_i; \widehat{\eta}^{(t)})$ can be done easily with the above weight. With $\tau_{i,\ell}^{(t)}$, we compute

$$Q(\eta | X_i; \widehat{\eta}^{(t)}) = \sum_{\ell=1}^{K} \tau_{i,\ell}^{(t)} [\log p(X_i; \theta_\ell) + \log \pi_\ell]$$

and

$$Q_n(\eta; \widehat{\eta}^{(t)}) = \sum_{i=1}^{n} \sum_{\ell=1}^{K} \tau_{i,\ell}^{(t)} [\log p(X_i; \theta_\ell) + \log \pi_\ell]. \tag{3.4}$$

- **M-step.** With the above Q_n function, we find the maximizer

$$\widehat{\eta}^{(t+1)} = \mathsf{argmax}_\eta Q_n(\eta; \widehat{\eta}^{(t)}).$$

While the M-step may seem to be hard, maximizing Q can be done easily if we know how to maximize the individual model $\ell(\theta_\ell | X_i) = \log p(x; \theta_\ell)$. To see this, we can rewrite equation (3.4) as

$$\begin{aligned} Q_n(\eta; \widehat{\eta}^{(t)}) &= \sum_{i=1}^{n} \sum_{\ell=1}^{K} \tau_{i,\ell}^{(t)} [\log p(X_i; \theta_\ell) + \log \pi_\ell] \\ &= \sum_{\ell=1}^{K} \underbrace{\left(\sum_{i=1}^{n} \tau_{i,\ell}^{(t)} \log p(X_i; \theta_\ell) \right)}_{A_\ell(\theta_\ell)} + \underbrace{\sum_{\ell=1}^{K} \left(\sum_{i=1}^{n} \tau_{i,\ell}^{(t)} \log \pi_\ell \right)}_{B}. \end{aligned} \tag{3.5}$$

Each quantity in $A_\ell(\theta_\ell)$ only involves parameter θ_ℓ so the maximization can be done easily *without affecting other* θ_j *with* $j \neq \ell$! Namely,

$$\widehat{\theta}_\ell^{(t+1)} = \mathsf{argmax}_{\theta_\ell} \sum_{i=1}^{n} \tau_{i,\ell}^{(t)} \log p(X_i; \theta_\ell).$$

This is essentially the MLE when we are using the single model $p(x; \theta_\ell)$ and each observation is assigned with a weight $\tau_{i,\ell}^{(t)}$.

For the component B, there is a closed-form solution to it using Lagrangian multiplier method:

$$\widehat{\pi}_\ell^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{i,\ell}^{(t)}}{n}.$$

Thus, we actually only need to compute the weight $\tau_{i,\ell}^{(t)}$ in equation (3.3) and the remaining part is just the simple MLE problem with no mixture.

Example: two-Gaussian mixture. Now we consider a simple case of two-Gaussian mixture that

$$p(x; \eta) = \rho \phi(x; \mu_1, \sigma_1^2) + (1 - \rho) \phi(x; \mu_2, \sigma_2^2),$$

where the parameter $\eta = (\rho, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$. Let $z \in \{1, 2\}$ be the class lable. Then the complete-case likelihood can be expressed as

$$p(x, z; \eta) = [\rho \cdot \phi(x; \mu_1, \sigma_1^2)]^{I(z=1)} [(1 - \rho) \cdot \phi(x; \mu_2, \sigma_2^2)]^{I(z=2)}.$$

Note that the complete-case likelihood is represented in a different form as the mixture model. With this, the log-complete likelihood will be

$$\log p(x, z; \eta) = I(z = 1) \log[\rho \cdot \phi(x; \mu_1, \sigma_1^2)] + I(z = 2) \log[(1 - \rho) \cdot \phi(x; \mu_2, \sigma_2^2)].$$

E-step. Given a parameter $\eta^{(t)} = (\rho^{(t)}, \mu_1^{(t)}, \mu_2^{(t)}, \sigma_1^{(t)2}, \sigma_2^{(t)2})$, the E-step will need

$$\tau(x; \eta^{(t)}) = P(Z = 1 | X = x; \eta^{(t)}) = \frac{\rho^{(t)} \cdot \phi(x; \mu_1^{(t)}, \sigma_1^{(t)2})}{\rho^{(t)} \cdot \phi(x; \mu_1^{(t)}, \sigma_1^{(t)2}) + (1 - \rho^{(t)}) \cdot \phi(x; \mu_2^{(t)}, \sigma_2^{(t)2})}.$$

Given observations $X_1, \cdots, X_n$, we will compute

$$\tau_i^{(t)} = \tau(X_i; \eta^{(t)})$$

in the E-step for each $i = 1, \cdots, n$.

M-step. Given the output from E-step, we will then update our parameters via

$$\begin{aligned}
\rho^{(t+1)} &= \frac{1}{n}\sum_{i=1}^{n} \tau_i^{(t)} \\
\mu_1^{(t+1)} &= \frac{\sum_{i=1}^n \tau_i^{(t)} X_i}{\sum_{j=1}^n \tau_j^{(t)}} \\
\mu_2^{(t+1)} &= \frac{\sum_{i=1}^n (1-\tau_i^{(t)}) X_i}{\sum_{j=1}^n (1-\tau_j^{(t)})} \\
\sigma_1^{(t+1)2} &= \frac{\sum_{i=1}^n \tau_i^{(t)} (X_i - \mu_1^{(t+1)})^2}{\sum_{j=1}^n \tau_j^{(t)}} \\
\sigma_2^{(t+1)2} &= \frac{\sum_{i=1}^n (1-\tau_i^{(t)}) (X_i - \mu_2^{(t+1)})^2}{\sum_{j=1}^n (1-\tau_j^{(t)})}.
\end{aligned}$$

Note that in this nice scenario, we have a closed-form in both E-step and M-step. Such elegant property may not also be there when an EM algorithm is applied.

3.2.1 The ascending property of the EM

Here is a simple derivation showing that when we update $\widehat{\eta}^{(t)}$ to $\widehat{\eta}^{(t+1)}$, the observed-data likelihood function will not decrease. Formally, we have the following result:

Proposition 3.1 *The observed-data log-likelihood value is non-decreasing during the EM updates, i.e.,*

$$\ell(\widehat{\eta}^{(t+1)} | X_1, \cdots, X_n) \geq \ell(\widehat{\eta}^{(t)} | X_1, \cdots, X_n), \tag{3.6}$$

where

$$\ell(\eta | X_1, \cdots, X_n) = \sum_{i=1}^{n} \ell_o(\eta | X_i) = \sum_{i=1}^{n} \log p(X_i; \eta).$$

Proof: For simplicity, here we consider $n = 1$ case. The proof can be easily generalized to any n.

The joint PDF

$$p(X, Z; \eta) = p(Z|X; \eta) p(X; \eta)$$

so the log-likelihood function

$$\ell(\eta | X) = \log p(X, Z; \eta) - \log p(Z|X; \eta).$$

This implies

$$\ell(\widehat{\eta}^{(t+1)}|X) - \ell(\widehat{\eta}^{(t)}|X) = [\log p(X, Z; \widehat{\eta}^{(t+1)}) - \log p(X, Z; \widehat{\eta}^{(t)})] - [\log p(Z|X; \widehat{\eta}^{(t+1)}) - \log p(Z|X; \widehat{\eta}^{(t)})].$$

Note that

$$Q_1(\eta; \widehat{\eta}^{(t)}) = \mathbb{E}(\log p(X, Z; \eta)|X; Z \sim p(\cdot|X; \widehat{\eta}^{(t)}))$$

so taking expectation in both sides of the equality with $Z \sim p(\cdot|X; \widehat{\eta}^{(t)})$ leads to

$$\begin{aligned}\ell(\widehat{\eta}^{(t+1)}|X) - \ell(\widehat{\eta}^{(t)}|X) &= Q_1(\widehat{\eta}^{(t+1)}; \widehat{\eta}^{(t)}) - Q_1(\widehat{\eta}^{(t)}; \widehat{\eta}^{(t)}) - \mathbb{E}\left(\log \frac{p(Z|X; \widehat{\eta}^{(t+1)})}{p(Z|X; \widehat{\eta}^{(t)})}|X; Z \sim p(\cdot|X; \widehat{\eta}^{(t)})\right) \\ &\geq -\mathbb{E}\left(\log \frac{p(Z|X; \widehat{\eta}^{(t+1)})}{p(Z|X; \widehat{\eta}^{(t)})}|X; Z \sim p(\cdot|X; \widehat{\eta}^{(t)})\right).\end{aligned}$$

Finally, by Jensen's inequality,

$$\begin{aligned}\mathbb{E}\left(\log \frac{p(Z|X; \widehat{\eta}^{(t+1)})}{p(Z|X; \widehat{\eta}^{(t)})}|X; Z \sim p(\cdot|X; \widehat{\eta}^{(t)})\right) &\leq \log \mathbb{E}\left(\frac{p(Z|X; \widehat{\eta}^{(t+1)})}{p(Z|X; \widehat{\eta}^{(t)})}|X; Z \sim p(\cdot|X; \widehat{\eta}^{(t)})\right) \\ &= \log \int \frac{p(z|X; \widehat{\eta}^{(t+1)})}{p(z|X; \widehat{\eta}^{(t)})} p(z|X; \widehat{\eta}^{(t)}) dz \\ &= \log \int p(z|X; \widehat{\eta}^{(t+1)}) dz = \log(1) = 0.\end{aligned}$$

Thus, we conclude that

$$\ell(\widehat{\eta}^{(t+1)}|X) - \ell(\widehat{\eta}^{(t)}|X) \geq 0,$$

which proves the desired result.

■

3.2.2 Convergence of the EM algorithm

While Proposition 3.1 shows that the EM algorithm will not decrease the observed-likelihood value, it does not imply that the EM algorithm will converge to the MLE. In fact, the EM algorithm may not converge to an MLE but could stuck at a local optimum if we do not initialize $\widehat{\eta}^{(0)}$ nicely. This spurious local modes problem is a notorious issue for EM and many other similar algorithm. Thus, in practice, we often re-initialize the algorithm multiple times and choose the convergent point with highest observed-likelihood value.

A classical convergence result of the EM algorithm is in the following paper:

> Wu, C. J. (1983). On the convergence properties of the EM algorithm. The Annals of statistics, 95-103.

It proves that the EM algorithm converges to a local mode of the observed-data log-likelihood function, as long as we do not initialize the algorithm at a saddle point or a local minimum.

The following paper provides a stronger statement on the behavior of the EM algorithm *when it is initialized around the MLE:*

> Balakrishnan, S., Wainwright, M. J., & Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. Annals of Statistics, 45(1), 77-120.

It proves the so-called linear convergence property of the EM algorithm; namely, there exists a constant $c \in (0, 1)$ such that

$$\|\widehat{\eta}^{(t)} - \widehat{\eta}\| \leq c^t \|\widehat{\eta}^{(0)} - \widehat{\eta}\|$$

when the initial point $\widehat{\eta}^{(0)} \in B(\widehat{\eta}, r)$ for some constant r and $\widehat{\eta}$ is the theoretical MLE (not computable).

3.3 Mixture of Products for Multivariate Problems

Mixture model offers an elegant solution to handle the dependency among variables of different types (some can be continuous, some are categorical, and some are discrete).

Example. Suppose each observation has three variables (X, Y, Z) such that $X \in \mathbb{R}$, $Y \in \{0, 1, 2, \cdots, \}$ is a counting number, and $Z \in \{0, 1, 2, 3, \cdots, N\}$ represents a score, where N is a known upper bound. It is known that these three variables are associated with each other. The question is: how can we model the joint distribution? The mixture model offers a simple solution. We assume that the joint PDF/PMF of (X, Y, Z) is

$$p(x, y, z) = \sum_{k=1}^{K} \pi_k \cdot \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-\frac{1}{2\sigma_k^2}(x-\mu_k)^2} \cdot \frac{\lambda_k^y}{y!} e^{-\lambda_k} \cdot \binom{N}{z} \theta_k^z (1-\theta_k)^{N-z}.$$

In this case, the parameters are

$$(\pi_k, \mu_k, \sigma_k^2, \lambda_k, \theta_k) : k = 1, \cdots, K.$$

Similar to the usual mixture model, we can estimate the parameters by the MLE and compute them via the EM algorithm. The elegancy of this model is that it is applicable to scenarios where variables can be different types and the mixture structure handles the dependency among variables.

Note: this model implies a conditional independence – let the latent variable G represent the component indicator. Namely, $G = k$ means that (X, Y, Z) is from the k-th component. Then we have $X \perp Y \perp Z | G$, which is known as a local independence assumption. While (X, Y, Z) are conditionally independent, they are unconditionally (marginally) dependent!

Formally, if we have a random vector $X_1, \cdots, X_p$, the mixture of product model is

$$p(x_1, \cdots, x_p) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{p} p(x_j; \theta_{j,k}),$$

where each $p(x_j; \theta_{j,k})$ is a model of variable X_j with parameter $\theta_{j,k}$. Similar to the previous example, we can compute the MLE via the EM algorithm.

This model can be used with missing data as well and it will offer a nature clustering structure. See the following paper for more information:

> Suen, D., & Chen, Y. C. (2023). Modeling Missing at Random Neuropsychological Test Scores Using a Mixture of Binomial Product Experts. arXiv preprint arXiv:2310.09384.

3.4 Other variants of mixture models

Here are some other variants of the mixture models. For readers who are interested in more details, I would recommend the following book chapter:

> Gormley, I. C., & Frühwirth-Schnatter, S. (2019). Mixture of experts models. Handbook of Mixture Analysis, 271-307.

Since here we may involve some concept of regression, we will change our notations a bit so that Y (rather than X) plays the central role in the mixture models. 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 θ_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 μ_k and σ_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,
$$\begin{aligned} p(x, y, z) = p(y|z)p(z|x)p(x) = p_z(y)\pi_z(x)p(x) &\Rightarrow p(y, z|x) = p_z(y)\pi_z(x) \\ &\Rightarrow p(y|x) = \sum_k p_k(y)\pi_k(x). \end{aligned}$$
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: $X \to Y$ and $Z \to Y$. This, the difference compared to the mixture of expert is that the arrow $X \to Z$ becomes $X \to Y$. In this case,
$$\begin{aligned} p(x, y, z) = p(y|x, z)p(z)p(x) = p_z(y|x)\pi_z p(x) &\Rightarrow p(y, z|x) = p_z(y|x)\pi_z \\ &\Rightarrow p(y|x) = \sum_k p_k(y|x)\pi_k. \end{aligned}$$
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 π_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: $X \to Y$, $X \to Z$, and $Z \to Y$. 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.

3.5 Remarks

There are three key issues that we need to address to use the MLE of a mixture model.

- **Identifiability.** Without additional constraints, the model may not be identifiable. For instance, when using the Gaussian mixture, swapping the sets of parameters between two components leads to the same distribution. A common approach to deal with this is to add an additional ordering constraint so that parameters are identifiable (e.g., $\mu_1 < \mu_2 < \mu_3 < \cdots$).
- **No-closed form of MLE.** In general, a mixture model will not a closed-form MLE so we need a numerical procedure to optimize the likelihood function. A common approach is the *EM algorithm*[1] but you can use a gradient ascent algorithm to find the MLE as well. However, a challenge of both EM algorithm and a gradient ascent method is that the likelihood function often has several local modes– both EM and the gradient method will stuck at the local modes. So we often have to re-initialize the algorithm several times to increase the chance of getting the MLE, although there is no guarantee of that. See,

  > Chen, Y. C. (2023). Statistical inference with local optima. Journal of the American Statistical Association, 118(543), 1940-1952.

- **Selection of the number of mixtures.** We need to choose the number of mixtures when using a mixture model. In general, how to choose this is a very difficult problem. A principled approach is to choose it via the BIC although other information criteria are also applicable. See the following paper for more details:

  > Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. The computer journal, 41(8), 578-588.

3.6 Variational inference

Variational inference (VI; also known as variational approximation) is a popular tool in machine learning. It has became more and more popular in statistics communities as well. In short, VI is a method to approximate

[1] See http://faculty.washington.edu/yenchic/18A_stat516/Lec8_EM_SGD.pdf

an intractable quantity using a tractable quantity. It can be used in both Frequentist estimation as well as Bayesian inference.

However, the fact that VI can be used in both Frequentist and Bayesian inference had made VI sometimes confusing. Here we will discuss how VI can be used in both cases and how the two problems are associated.

3.6.1 Approximating an MLE (Frequentist)

Consider a regular latent variables problem where we observe IID

$$X_1, \cdots, X_n \sim p$$

and each observation has a latent variable Z that is unobserved. Namely, the complete data should be

$$(X_1, Z_1), \cdots, (X_n, Z_n)$$

but we only observe $X_1, \cdots, X_n$.

In the latent variable problem, we often place a parametric model on the complete-data distribution:

$$p(x, z; \lambda), \quad \lambda \in \Lambda.$$

This parametric model implies the observed model

$$p(x; \lambda) = \int p(x, z; \lambda) dz$$

that generates our observations.

In this case, the MLE is defined as

$$\widehat{\lambda} = \mathsf{argmax}_\lambda \frac{1}{n} \sum_{i=1}^{n} \log p(X_i; \lambda).$$

Namely, we maximizes the *observed log-likelihood* $\ell(\lambda|x) = p(x; \lambda)$. A population version of this problem is

$$\lambda^* = \mathsf{argmax}_\lambda \mathbb{E}(\log p(X_1; \lambda)).$$

In most cases, the MLE does not have a closed-form so we need to use numerical procedure such as the EM algorithm to compute it. However, EM algorithm may not have a simple form and can still be intractable (this occurs a lot when the Q function in the EM algorithm is complicated).

The VI offers a remedy to this problem. Note that we can write the observed model as

$$\begin{aligned} p(x; \lambda) &= \int p(x, z; \lambda) dz \\ &= \int \frac{p(x, z; \lambda)}{q(z; \omega)} q(z; \omega) dz \\ &= \mathbb{E}_{Z \sim q(\cdot; \omega)} \left(\frac{p(x, Z; \lambda)}{q(Z; \omega)} \right), \end{aligned}$$

where $\omega \in \Omega$ is another class of parameters and $Q = \{q(\cdot; \omega) : \omega \in \Omega\}$ is called the variational family. The variational family often consists of parametric densities that are easy to compute and sample.

Using the Jensen's inequality, the observed log-likelihood function becomes

$$\begin{aligned}\ell(\lambda|x) &= \log p(x;\lambda) \\ &= \log \mathbb{E}_{Z\sim q(\cdot;\omega)}\left(\frac{p(x,Z;\lambda)}{q(Z;\omega)}\right) \\ &\geq \mathbb{E}_{Z\sim q(\cdot;\omega)}\left(\log\frac{p(x,Z;\lambda)}{q(Z;\omega)}\right) \\ &= \mathbb{E}_{Z\sim q(\cdot;\omega)}\log p(x,Z;\lambda) - \mathbb{E}_{Z\sim q(\cdot;\omega)}(\log q(Z;\omega)) \\ &= \mathsf{ELBO}(\omega,\lambda|x).\end{aligned}$$

The quantity $\mathsf{ELBO}(\omega,\lambda|x)$ is called the *evidence lower bound*. Note that sometime people would write it as $\mathsf{ELBO}(q)$ and abbreviate λ, x but here for completeness I would keep both of them. When we observe n observations, we write

$$\mathsf{ELBO}_n(\omega,\lambda) = \frac{1}{n}\sum_{i=1}^n \mathsf{ELBO}(\omega,\lambda|X_i).$$

Note that the ELBO is similar to something similar to the KL-divergence. Define a 'partial' KL-divergence:

$$\widetilde{\mathsf{KL}}(q(\cdot;\omega)||p(\cdot,x;\lambda)) = \mathbb{E}_{Z\sim q(\cdot;\omega)}\left(\log\frac{q(Z;\omega)}{p(x,Z;\lambda)}\right),$$

which is very similar to

$$\mathsf{ELBO}(\omega,\lambda|x) = \mathbb{E}_{Z\sim q(\cdot;\omega)}\left(\log\frac{p(x,Z;\lambda)}{q(Z;\omega)}\right).$$

We only swap the numerator and denominator. Note that $\widetilde{\mathsf{KL}}$ is not the KL-divergence because we fixed x and only consider randomness of Z.

With the ELBO, the idea of variational inference is very simple. Instead of maximizing the intractable log-likelihood, we maximize ELBO to find our estimator. Namely, our estimator is

$$(\widehat{\omega}_{\mathsf{VI}}, \widehat{\lambda}_{\mathsf{VI}}) = \mathsf{argmax}_{\omega,\lambda}\mathsf{ELBO}_n(\omega,\lambda).$$

Again, in general there is no closed-form to the above estimators. But we can solve this maximization problem numerically. The power of VI is that the variational family is designed so that evaluation and computation are easy. So the expectation $\mathbb{E}_{Z\sim q(\cdot;\omega)}(\cdots)$ that appears in the ELBO is a tractable quantity.

A famous method called *mean field variational family* is the collection of densities

$$q(z;\omega) = \prod_{j=1}^d q(z_j;\omega_j),$$

where $z = (z_1, \cdots, z_d)$. In this case, sampling of z can be decomposed into sampling each coordinate independently and the optimization of ω_j can be obtained by the *coordinate ascent variational inference* algorithm in the following book:

> Bishop, C. (2006). Pattern Recognition and Machine Learning. Springer New York.

One thing to keep in mind is that when using VI, we are no longer solving the original problem. In general, the VI estimator will not converge to the MLE (in the Frequentist setting) but instead, it will still converge to a population quantity; see the following paper:

> Chen, Y. C., Wang, Y. S., & Erosheva, E. A. (2018). On the use of bootstrap with variational inference: Theory, interpretation, and a two-sample test example. The Annals of Applied Statistics, 12(2), 846-876.

3.6.2 Density evaluation problem (Bayesian)

A good review from the Bayesian perspective is

> Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. Journal of the American statistical Association, 112(518), 859-877.

Consider a simple Bayesian setting where

$$X|\theta \sim p(x|\theta), \quad \theta \sim \pi(\theta).$$

Here, X is the random variable for the observation and θ is the underlying parameter of the distribution and π is the prior. The Bayesian inference relies heavily on the posterior distribution (density):

$$\pi(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{p(x)} \propto p(x|\theta)\pi(\theta) = p(x,\theta).$$

Here we write $p(x,\theta) = p(x|\theta)\pi(\theta)$ for simplicity. In addition to the posterior distribution, the evidence $p(x) = \int p(x,\theta)d\theta$ is also sometimes of interest.

Since $p(x,\theta) = p(x|\theta)\pi(\theta)$, it is often very easy to evaluate the value of $p(x,\theta)$ for any given x,θ. Although the joint distribution is easy to compute, both the posterior and the evidence are often intractable. Take the evidence as an example, it can be written as the integral $p(x) = \int p(x,\theta)d\theta$. When the dimension of θ is large (say a mixture model), this integration is very difficult to evaluate even if we can easily compute $p(x,\theta)$.

Here is an interesting note. The problem of evaluating the evidence and the problem of evaluating the posterior distribution are the same via the following relation:

$$\pi(\theta|x) = \frac{p(x,\theta)}{p(x)}.$$

$p(x,\theta)$ is tractable so we can easily convert the evidence and the posterior to each other.

To obtain a tractable approximation of $\pi(\theta|x)$, we use the idea of VI. The quantity θ now plays the role of latent variable z in the previous section. Let $q(\theta;\omega)$ be a computable density and

$$Q = \{q(\cdot;\omega) : \omega \in \Omega\}$$

be the variational family. We attempt to find the best density $q(;\omega^*) \in Q$ such that

$$q(\cdot;\omega^*) = \mathsf{argmin}_{q\in Q}\mathsf{KL}(q(\cdot)||\pi(\cdot|x)),$$

where KL is the Kullback-Leibler divergence. Namely,

$$\omega^* = \mathsf{argmin}_{\omega}\mathsf{KL}(q(\cdot;\omega)||\pi(\cdot|x)),$$

However, minimizing the KL divergence may again ran into the same computation problem (need to evaluate the integral). The idea of VI uses the following insight:

$$\begin{aligned}\log p(x) &= \mathsf{KL}(q(\cdot;\omega)||\pi(\cdot|x)) + \mathbb{E}_{\theta\sim q(\cdot;\omega)}(\log p(x,\theta)) - \mathbb{E}_{\theta\sim q(\cdot;\omega)}(\log q(\theta;\omega)) \\ &= \mathsf{KL}(q(\cdot;\omega)||\pi(\cdot|x)) + \mathsf{ELBO}(\omega|x).\end{aligned}$$

Therefore, minimizing the KL divergence is equivalent to maximizing ELBO.

The variational approximation chooses

$$\omega^*_{\mathsf{VI}}(x) = \mathsf{argmax}_\omega \mathsf{ELBO}(\omega|x)$$

and use $q(\theta; \omega^*_{\mathsf{VI}}(x))$ as an approximation to $\pi(\theta|x)$. This gives a good approximation of $\pi(\theta|x)$.

In the case of observing $X_1, \cdots, X_n$, we want an approximation of $\pi(\theta|X_1, \cdots, X_n)$. In this case, ELBO will be

$$\begin{aligned}\mathsf{ELBO}_n(\omega) &= \mathbb{E}_{\theta\sim q(\cdot;\omega)}(\log p(X_1, \cdots, X_n, \theta)) - \mathbb{E}_{\theta\sim q(\cdot;\omega)}(\log q(\theta;\omega)) \\ &= \mathbb{E}_{\theta\sim q(\cdot;\omega)}\left(\log \pi(\theta) + \sum_{i=1}^n \log p(X_i|\theta)\right) - \mathbb{E}_{\theta\sim q(\cdot;\omega)}(\log q(\theta;\omega))\end{aligned}$$

and we estimate ω using

$$\widehat{\omega}_{\mathsf{VI}} = \mathsf{argmax}_\omega \mathsf{ELBO}_n(\omega).$$

Again, this maximization is often tractable since the expectation is with respect to the variational distribution q, which is by design easy to compute.

The posterior is then approximated by

$$\pi(\theta|X_1, \cdots, X_n) \approx p(\theta; \widehat{\omega}_{\mathsf{VI}})$$

and the evidence is approximated by

$$p(X_1, \cdots, X_n) \approx \frac{\pi(\theta)\prod_{i=1}^n p(X_1|\theta)}{p(\theta; \widehat{\omega}_{\mathsf{VI}})}.$$

Note that the approximated evidence may depend on θ because it is an approximation rather than an exact value.