Bayesian Continual Learning

Jifei Luo

October 28, 2022

Streaming Bayes[1]

Consider data $x_1, x_2, \ldots$ generated iid according to a distribution $p(x|\Theta)$ given parameters Θ. The Bayes theorem gives us the posterior distribution of Θ given a collection of S_1 data points, $C_1 := (x_1, \ldots, x_{S_1})$:

$$p(\Theta|C_1) = p(C_1)^{-1} p(C_1|\Theta) p(\Theta)$$

where $p(C_1|\Theta) = p(x_1, \ldots, x_{S_1}|\Theta) = \prod_{i=1}^{S_1} p(x_i|\Theta)$

Suppose we have seen and processed $k-1$ minibatchs of data. Given the posterior $p(\Theta|C_1, \ldots C_{k-1})$, we can calculate the posterior after the k-th minibatch:

$$p(\Theta|C_1, \ldots C_k) \propto p(C_k|\Theta) p(\Theta|C_1, \ldots C_{k-1})$$

[1] Tamara Broderick et al. "Streaming Variational Bayes". In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper/2013/file/51ef186e18dc00c2d31982567235c559-Paper.pdf.

Incremental Moment Matching[2]

Lee proposed two types of incremental moment matching (IMM) methods for overcoming catastrophic forgetting, they are so called Mean-Incremental Moment Matching (mean-IMM) and Mode-Incremental Moment Matching (mode-IMM). IMM can be interpreted in Bayesian perspectives.

For the sake of uniform representation, we use a Gaussian distribution $f_i \sim N(\mu_i, \Sigma_i)$ to approximate the posterior distribution of parameters in i-th stage, what we want to find is a Gaussian distribution $g \sim N(\mu^*, \Sigma^*)$ that can approximate the whole posterior of the previous k stages.

[2] Sang-Woo Lee et al. "Overcoming Catastrophic Forgetting by Incremental Moment Matching". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/f708f064faaf32a43e4d3c784e6af9ea-Paper.pdf.

Incremental Moment Matching

Mean-IMM: minimize local KL-divergence

$$\mu^*, \Sigma^* = \arg\min_{\mu,\Sigma} \sum_{i=1}^{k} \alpha_i KL(f_i||g)$$

$$\mu^* = \sum_{i=1}^{k} \alpha_i \mu_i$$

$$\Sigma^* = \sum_{i=1}^{k} \alpha_i (\Sigma_i + (\mu_i - \mu)(\mu_i - \mu)^T)$$

Mode-IMM: find a mode of mixture of local posteriors

$$\mu^*, \Sigma^* = \arg\max_{\mu} \sum_{i=1}^{k} \alpha_i \log f_i$$

$$\mu^* = \Sigma^* \sum_{i=1}^{k} \alpha_i \Sigma_i^{-1} \mu_i$$

$$\Sigma^* = (\sum_{i=1}^{k} \alpha_i \Sigma_i^{-1})^{-1}$$

Incremental Moment Matching

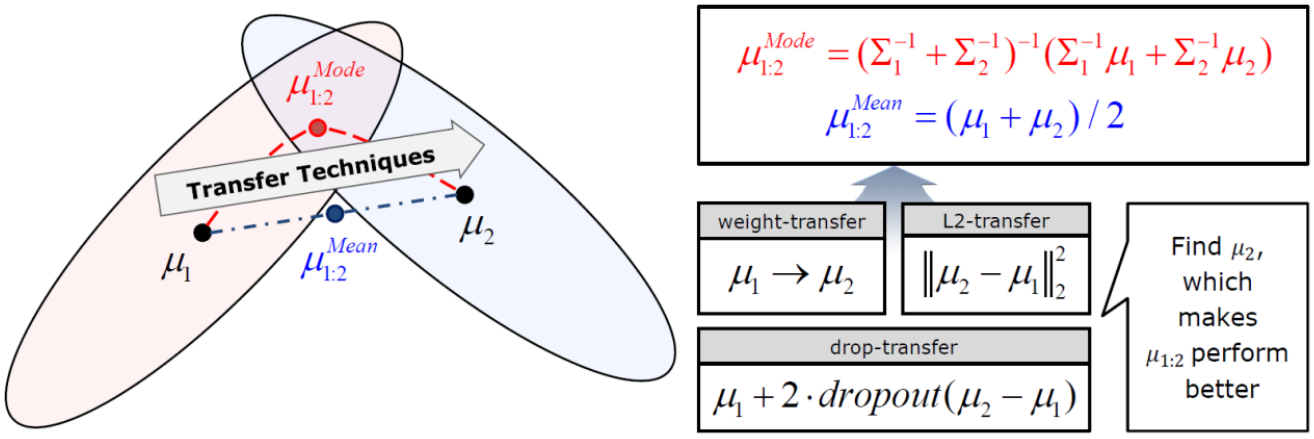

Figure: Geometric illustration of incremental moment matching (IMM)

Hierarchical Clustering of a Mixture Model[3]

In this work, the author proposed a useful clustering algorithm, suppose we are given a mixture density f composed of k d-dimensional Gaussian components:

$$f(\theta) = \sum_{i=1}^{k} \alpha_i N(\theta; \mu_i, \Sigma_i) = \sum_{i=1}^{k} \alpha_i f_i(\theta)$$

We want to cluster the components of f into a reduced mixture of $m < k$ components. This work introduced a distance measure between $f = \sum_{i=0}^{k} \alpha_i f_i$ and $g = \sum_{j=0}^{m} \beta_j g_j$ which can be analytically computed.

$$d(f, g) = \sum_{i=1}^{k} \alpha_i \min_{j=1}^{m} KL(f_i || g_j)$$

[3] Jacob Goldberger and Sam Roweis. "Hierarchical Clustering of a Mixture Model". In: *Advances in Neural Information Processing Systems*. Ed. by L. Saul, Y. Weiss, and L. Bottou. Vol. 17. MIT Press, 2004. URL: https://proceedings.neurips.cc/paper/2004/file/36e729ec173b94133d8fa552e4029f8b-Paper.pdf.

Proof

The KL-divergence between two Gaussian distributions $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ is shown as below, see proof here.

$$\frac{1}{2}(\log \frac{|\Sigma_2|}{|\Sigma_1|} + tr(\Sigma_2^{-1}\Sigma_1) + (\mu_1 - \mu_2)^T \Sigma_2^{-1}(\mu_1 - \mu_2) - c)$$

For simplicity, we consider reducing the components of f into only one mixture component, denote as $g \sim N(\mu, \sigma)$, the target function becomes:

$$\arg\min_{\mu,\Sigma} d(f, g) = \arg\min_{\mu,\Sigma} \sum_{i=1}^{k} \alpha_i KL(f_i||g)$$

Proof

The ideal Gaussian parameters μ and Σ need to satisfy the following equations respectively.

$$\frac{\partial d(f,g)}{\partial \mu} = 0, \quad \frac{\partial d(f,g)}{\partial \Sigma} = 0$$

For the first part,

$$\begin{aligned}\frac{\partial d(f,g)}{\partial \mu} &= \sum_{i=1}^{k} \alpha_i (\frac{\partial (\mu_i - \mu)}{\partial \mu})^T \frac{\partial (\mu_i - \mu)^T \Sigma^{-1} (\mu_i - \mu)}{\partial (\mu_i - \mu)} \\ &= \sum_{i=1}^{k} -\alpha_i 2I\Sigma^{-1}(\mu - \mu_i) = 0\end{aligned}$$

we can get:

$$\mu = \sum_{i=1}^{k} \alpha_i \mu_i$$

Proof

For the second one, we can apply the following Jacobian matrix into our proof.

$$\frac{\partial \log |X|}{\partial X} = X^{-1}$$
$$\frac{\partial tr(AX^{-1})}{\partial X} = \frac{\partial tr(X^{-1}A)}{\partial X} = -X^{-1}AX^{-1}$$
$$\frac{\partial tr(AX^{-1}B)}{\partial X} = -X^{-1}BAX^{-1}$$

The second equation can be formulated as:

$$\frac{\partial d(f,g)}{\partial \Sigma} = \frac{1}{2}\sum_{i=1}^{k} \alpha_i(\Sigma^{-1} - \Sigma^{-1}\Sigma_i\Sigma^{-1} - \Sigma^{-1}(\mu_i - \mu)(\mu_i - \mu)^T\Sigma^{-1}) = 0$$

it will be easy to get:

$$\Sigma = \sum_{i=1}^{k} \alpha_i(\Sigma_i + (\mu_i - \mu)(\mu_i - \mu)^T) \qquad \square$$

The Topography of Multivariate Normal Mixtures[4]

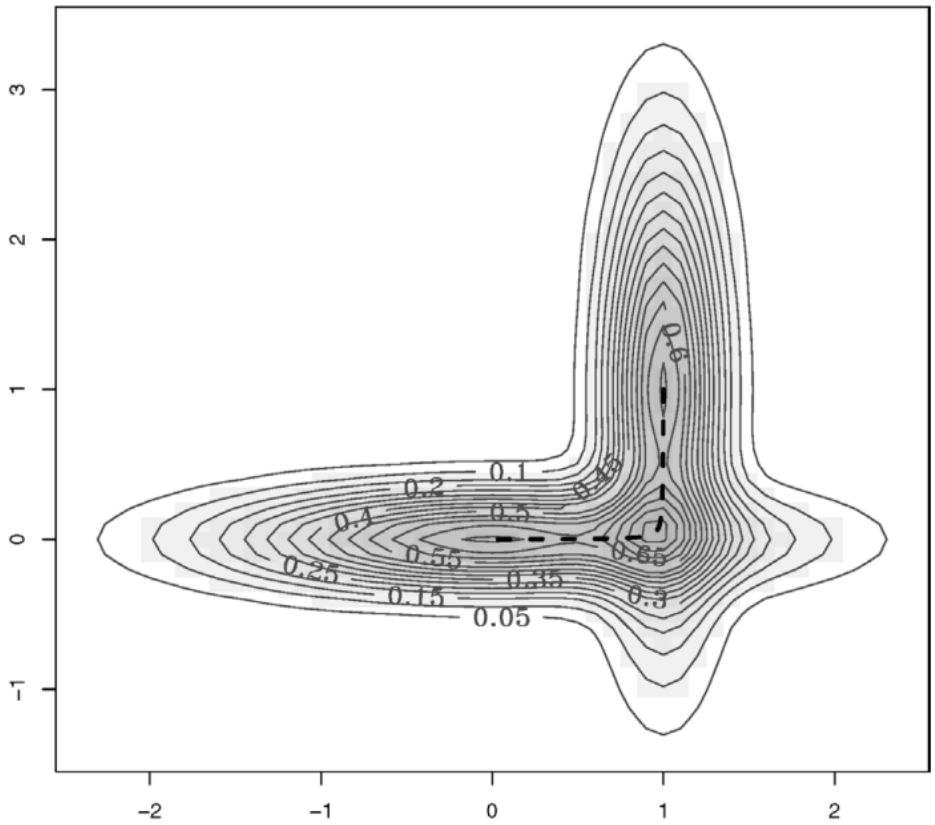

Figure: Contour plot and ridgeline curve for the mixture density

[4] Surajit Ray and Bruce G. Lindsay. "The Topography of Multivariate Normal Mixtures". In: *The Annals of Statistics*. Vol. 33. Institute of Mathematical Statistics, 2005, pp. 2042–2065. URL: http://www.jstor.org/stable/3448634.

Proof

We suppose f_i satisfy the Gaussian distribution as below:

$$f_i(\theta; \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2}} \exp\{-\frac{1}{2}(\theta - \mu_i)^T \Sigma_i^{-1}(\theta - \mu_i)\}$$

Our goal is to find:

$$\mu^* = \arg\max_{\theta} \sum_{i=1}^{k} \alpha_i \log f_i = -\frac{1}{2} \sum_{i=1}^{k} \alpha_i (\theta - \mu_i)^T \Sigma_i^{-1}(\theta - \mu_i) + C$$

Let the derivation to be 0:

$$\begin{aligned} \frac{\partial \sum_{i=1}^{k} \alpha_i \log f_i}{\partial \theta} &= -\frac{1}{2} \sum_{i=1}^{k} \alpha_i (\Sigma_i^{-1} + (\Sigma_i^{-1})^T)(\theta - \mu_i) \\ &= -\sum_{i=1}^{k} \alpha_i \Sigma_i^{-1}(\theta - \mu_i) = 0 \end{aligned}$$

we can get:

$$\mu^* = \theta^* = (\sum_{i=1}^{k} \alpha_i \Sigma_i^{-1})^{-1} (\sum_{i=1}^{k} \alpha_i \Sigma_i^{-1} \mu_i)$$

Proof

Our Gaussian mixture model g need to satisfy the distribution of $N(\mu^*, \Sigma^*)$, we can take the second derivation to approximate the covariance matrix Σ^*, that is:

$$\Sigma^{*-1} = -\frac{\partial^2 \sum_{i=1}^{k} \alpha_i \log f_i}{\partial \theta^2} = \frac{\partial \sum_{i=1}^{k} \alpha_i \Sigma_i^{-1}(\theta - \mu_i)}{\partial \theta}$$

Find the inverse of the matrix above we can get:

$$\Sigma^* = (\sum_{i=1}^{k} \alpha_i \Sigma_i^{-1})^{-1} \qquad \square$$