Deep|Bayes summer school

Practical Session: Approximate Bayesian inference

August 20, 2019

Consider a clustering problem: we have a dataset $X = \{x_i\}_{i=1}^N$, where $x_i \in \mathbb{R}$ — scalar features, and want to group these objects into K clusters. To do so we assume that points from the dataset were generated from a Gaussian Mixture Model (GMM) with K components. For each object x_i we establish additional latent variable z_i which denotes the index of the gaussian from which i-th object was generated. For convenience we use binary vector notation for z_i: $z_i \in \{0,1\}^K, \quad \sum_{k=1}^K z_{ik} = 1$.

Below we consider several variations of Gaussian Mixture Model for clustering problem with different prior assumptions. The task is to choose a suitable method of training for each of them and write down formulas for training: find optimal values of parameters and find a posterior distribution of latent variables.

1. [Basic GMM – no additional priors] A probabilistic model of standard GMM looks as follows:

$$p(X, Z|\pi, \mu, \lambda) = \prod_{i=1}^N p(z_i|\pi)p(x_i|z_i, \mu, \lambda) = \prod_{i=1}^N \prod_{k=1}^K \left[\pi_k \mathcal{N}(x_i|\mu_k, \lambda_k^{-1})\right]^{z_{ik}}$$

Parameter $\pi = (\pi_1, \dots, \pi_K)$ denotes probabilities of Gaussian components in the mixture and is restricted to a simplex: $\sum_{k=1}^K \pi_k = 1$ and $\pi_k \geq 0$ for $k = 1, \dots, K$. μ and λ contain parameters of Gaussian components (λ contains inverse variances).

Before reading further, try to choose a suitable method to train this model by yourself.

We can train this model with EM-algorithm because (i) there are both latent variables and parameters in the model and (ii) the prior on Z and the likelihood are conjugate.

   (a) Check that the prior distribution $p(Z|\pi)$ and the likelihood $p(X|Z, \pi, \mu, \lambda)$ are conjugate.

   (b) E-step: derive the posterior distribution $p(Z|X, \pi, \mu, \lambda)$. Values of π, μ, λ are fixed on this step.

   (c) M-step: compute optimal values of π, μ, λ by maximizing $\mathbb{E}_{p(Z|X,\pi,\mu,\lambda)} \log p(X, Z|\pi, \mu, \lambda)$. Posterior distribution $p(Z|X, \pi, \mu, \lambda)$ is fixed on this step.

2. [GMM with prior on π] We can add Dirichlet prior on probabilities of Gaussian components π to obtain either sparse or more even clusters (depending on values of parameters α of the prior). A probabilistic model, in this case, looks as follows:

$$p(X, Z, \pi|\mu, \lambda) = p(\pi) \prod_{i=1}^N p(z_i|\pi)p(x_i|z_i, \mu, \lambda) = \mathrm{Dir}(\pi|\alpha) \prod_{i=1}^N \prod_{k=1}^K \left[\pi_k \mathcal{N}(x_i|\mu_k, \lambda_k^{-1})\right]^{z_{ik}}$$

Here parameter α is fixed (we do not train it). $\alpha < 1$ gives sparse solutions, $\alpha > 1$ results in more even clusters.

Before reading further, try to choose a suitable method to train this model by yourself.

To train this model we need variational EM-algorithm because the prior on Z, π and the likelihood are not conjugate.

   (a) Check that the prior distribution $p(Z, \pi)$ and the likelihood $p(X|Z, \pi, \mu, \lambda)$ are not conjugate.

   (b) Check that there is a conditional conjugacy between the prior and the likelihood if we use the following factorization: $p(Z, \pi|X, \mu, \lambda) \approx q(Z, \pi) = q(Z)q(\pi)$.

   (c) E-step: write down update rules for $q(Z)$ and $q(\pi)$ as in variational inference. Values of μ, λ are fixed on this step.

   (d) M-step: compute optimal values of μ, λ by maximizing $\mathbb{E}_{q(Z,\pi)} \log p(X, Z, \pi|\mu, \lambda)$. Posterior approximation $q(Z)q(\pi)$ is fixed on this step.

* [GMM with priors on π, μ, λ] We can add Normal-Gamma priors on parameters of Gaussian components μ_k, λ_k to make the model prefer clusters of specific form or location. A probabilistic model, in this case, looks as follows:

$$\begin{aligned} p(X, Z, \pi, \mu, \lambda) =& p(\pi) \left[\prod_{k=1}^K p(\mu_k, \lambda_k) \right] \prod_{i=1}^N p(z_i|\pi) p(x_i|z_i, \mu, \lambda) \\ &= \mathrm{Dir}(\pi|\alpha) \left[\prod_{k=1}^K \mathcal{N}(\mu_k|m, (\beta\lambda_k)^{-1}) \mathcal{G}(\lambda_k|a, b) \right] \prod_{i=1}^N \prod_{k=1}^K \left[\pi_k \mathcal{N}(x_i|\mu_k, \lambda_k^{-1}) \right]^{z_{ik}} \end{aligned}$$

Here α, m, β, a, b are fixed (we do not train them).

We can train this model with mean-field variational inference because (i) the model contains only latent variables and not parameters and (ii) the prior on Z, π, μ, λ and the likelihood are not conjugate. To approximate the posterior the following factorization should be used:

$$p(Z, \pi, \mu, \lambda|X) \approx q(Z, \pi, \mu, \lambda) = q(Z) q(\pi, \mu, \lambda).$$

Useful formulas

Normal distribution:

$$\mathcal{N}(x|\mu, \lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} \exp\left[-\frac{1}{2}(x-\mu)^2 \lambda \right]$$

Dirichlet distribution:

$$\mathrm{Dir}(\pi|\alpha) = \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{k=1}^K \pi_k^{\alpha_k - 1}, \quad \pi \in S_K$$