

Problem Assignment #3

Statistical Machine Learning

1 Problem 1. 2-D Mixture Gaussian

1.1 Model

生成模型如下

$$Z_i \sim \text{Discrete}(\pi) \quad (1)$$

$$X_i|Z_i \sim \mathcal{N}(\mu_{Z_i}, \sigma_{Z_i}^2) \quad (2)$$

其中 Z_i 和 X_i 是随机变量, $\{x_i\}_{i=1}^N$ 是 X_i 的具体的观测。Discrete(x) 为取值为 $\{1, 2, \dots, K\}$ 的随机变量, 并且满足 $\mathbb{P}(Z_i) = \pi$ 。 $\pi \in \mathbb{R}^K, \mu_i \in \mathbb{R}^D, \sigma_i^2 \in \mathbb{R}^D$ 是需要学习的参数, 其中 $i \in \{1, 2, \dots, K\}$ 。 N, K 和 D 是固定的常数, 都是正整数。

在此问题中我们固定 $D = 2$, 需要使用 variational inference 的方法求出参数 $\pi \in \mathbb{R}^K, \mu_i \in \mathbb{R}^D, \sigma_i^2 \in \mathbb{R}^D$ 的极大似然估计。

1.2 Requirements

代码需要包括：

1. 在 ZhuSuan 中实现生成模型 (model), 并使用你写的模型生成数据 (这一部分可以参考 vae 的教程的生成图像的部分)。
2. 在 ZhuSuan 中设计合理的 variational posterior distribution(q_net)。
3. 基于 ZhuSuan 的框架实现整个算法 (generative-inference-learning), 用 [1] 步生成的数据进行训练。
4. 对于结果的可视化 (可以使用 matplotlib)。

报告需要包括：

1. 对于整个问题的形式化描述 (按照 generative-infernece-learning 的框架), 尤其是设计的 variational posterior distribution 的描述与分析。
2. 固定 $N = 100, K = 3$, 生成数据并完成训练, 从数值角度和可视化的角度展示你的结果 (需要有具体的图)。
- *3. 探讨 N, K 以及参数的真值 (比如 μ_i 之间是否比较近) 的变化对你的方法产生结果的影响。
- *4. 如果设计了多个 variational posterior, 比较不同的 variational posterior 的结果。

关于数据生成：先固定 N, D, K , 随机生成或者手动选取参数的真值 $\pi \in \mathbb{R}^K, \mu_i \in \mathbb{R}^D, \sigma_i^2 \in \mathbb{R}^D$, 使用生成模型生成 N 组数据 $\{x_i\}_{i=1}^N$, 在求解极大似然估计的时候应该只用生成的数据 $\{x_i\}_{i=1}^N$ 而“忘记”参数的真值, 最后比较极大似然估计的估计值和参数的真值。