Problem Assignment #3

Statistical Machine Learning

1 Problem 1. 2-D Gaussian Mixture

1.1 Model

生成模型如下

$$Z_i \sim \text{Discrete}(\pi) \tag{1}$$
$$X_i|Z_i \sim \mathcal{N}(\mu_{Z_i}, \sigma^2_{Z_i}) \tag{2}$$

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 $\{1, 2, \cdots, K\}$ 的随机变量, 并且满足 $\mathbb{P}(Z_i) = \pi_i$。$\pi \in \mathbb{R}^K, \mu_i \in \mathbb{R}^D, \sigma_i^2 \in \mathbb{R}^D$ 是需要学习的参数, 其中 $i \in \{1, 2, \cdots, K\}$。$N$、$K$ 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 (model-inference-learning), 用 [1] 步生成的数据进行训练 (推荐使用 zs.rws)。
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 model-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。

2 Problem 2. Gaussian Mixture VAE

我们考虑将 Gaussian Mixture 和 VAE 结合起来，生成模型如下

$$
\begin{aligned}
Z_i &\sim \text{Discrete}(\pi) && (3)\\
H_i|Z_i &\sim \mathcal{N}(\mu_{Z_i}, \sigma^2_{Z_i}) && (4)\\
X_i|H_i &\sim \text{Bernoulli}(\sigma_{NN}(H_i)) && (5)
\end{aligned}
$$

其中 Z_i, H_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 $\{1, 2, \cdots, K\}$ 的随机变量，并且满足 $\mathbb{P}(Z_i) = \pi_i$。$Z_i \in \mathbb{R}^K$, $H_i \in \mathbb{R}^D$, $X_i \in \{0, 1\}^{784}$。$\sigma_{NN}(H_i)$ 为神经网络，构成一个从 $\mathbb{R}^D$ 到 $[0, 1]^{784}$ 的映射。以 $\pi \in \mathbb{R}^K, \mu_i \in \mathbb{R}^D, \sigma_i^2 \in \mathbb{R}^D$ 和神经网络 $\sigma_{NN}(\cdot)$ 中的参数是需要学习的参数，其中 $i \in \{1, 2, \cdots, K\}$。$N$、$K$ 和 D 是固定的常数，都是正整数。

在此问题中我们固定 $D = 40$，$K = 10$，使用 MNIST 数据集 (具体使用可以看 examples.utils 中的 dataset.py 和 examples 中的 vae.py) 需要使用 variational inference 的方法求出参数 $\pi \in \mathbb{R}^K, \mu_i \in \mathbb{R}^D, \sigma_i^2 \in \mathbb{R}^D$ 和神经网络 $\sigma_{NN}(\cdot)$ 中的参数的极大似然估计，建议使用 Reweighted Wake Sleep 算法。

2.1 Requirements

代码需要包括：

1. 在 ZhuSuan 中实现生成模型 (model)。
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 (model-inference-learning)，用 MNIST 数据集进行训练 (推荐使用 zs.rws)。
4. 对于结果的可视化 (画出 samples，固定取 $Z_i = \{1, 2, \cdots, K\}$ 不变，对于每一种用生成模型生成多个 X_i，观察其聚类效果)。

报告需要包括：

1. 对于整个问题的形式化描述 (按照 model-infernece-learning 的框架), 尤其是设计的 variational posterior distribution 的描述与分析。
2. 对于数值结果的分析和 samples 的可视化展示与分析。

*3. 如果设计了多个 variational posterior，比较不同的 variational posterior 的结果。