作业 3

截止日期 2024/12/15 23:59

1 KL 散度和最大似然

Kullback-Leibler（KL）散度是衡量两个概率分布差异大小的指标。它是一个起源于信息论的概念，但已经进入了其他几个领域，包括统计学、机器学习、信息几何等等。在机器学习中，KL 散度起着至关重要的作用，将各种看似无关的概念联系起来。

在这个问题中，我们将介绍离散分布上的 KL 散度，练习一些简单的操作，并了解它与最大似然估计的关系。

在空间 $\mathcal{X}$ 上，两个离散值分布的 KL 散度 $P(X), Q(X)$ 定义如下：

$$D_{KL}(P\|Q) = \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)}$$

为了标记方便，我们假定 $P(x) > 0, \forall x.$ (另外默认"$0\log 0 = 0$.") 有时，我们也把 KL 散度写的更清楚一些，如：$D_{KL}(P||Q) = D_{KL}(P(X)||Q(X))$.

信息论的背景

在我们深入探讨之前，我们简要（可选）介绍了 KL 散度的信息论背景。虽然这篇介绍不是回答作业问题的必要条件，但它可以帮助你更好地理解和欣赏我们为什么研究 KL 散度，以及信息理论如何与机器学习相关。

我们从一个概率分布 $P(X)$ 的熵 $H(P)$ 开始，它被定义为

$$H(P) = -\sum_{x\in\mathcal{X}} P(x)\log P(x).$$

直观地说，熵衡量概率分布的分散程度。例如，均匀分布被认为具有非常高的熵（即大量的不确定性），而将所有质量分配给单个点的分布被认为熵为零（即没有不确定性）。值得注意的是，可以证明，在 $\mathbb{R}$ 上的连续分布中，高斯分布 $N(\mu, \sigma^2)$ 在所有具有给定均值 μ 和方差 σ^2 的可能分布中具有最高的熵（最高的不确定性）。

为了进一步巩固我们的直觉，我们从传播理论中提出了动机。假设我们想从源通信到目的地，我们的消息总是（一系列）空间 X 上的离散符号（例如，X 可以是字母 $\{a, b, \ldots, z\}$）。我们希望以通过信道传输的二进制比特序列的形式为我们的符号构建一个编码方案。此外，假设从长远来看，符号的出现频率遵循概率分布 $P(X)$。这意味着，从长远来看，符号 x 被传输的次数分数是 $P(x)$。

一个通用的思路是构建一种编码方案，使每个传输符号的平均比特数保持尽可能小。直观地说，这意味着我们希望将较常出现符号分配少量的比特。同样，由于我们希望从整体上减少每个符号的平均比特数，因此可以容忍将不常出现的符号分配大量的比特，因为这样对平均值的影响很小。编码方案可以根据我们的想法而复杂，例如，单个比特可能表示多个符号的长序列（如果特定的符号模式非常常见）。概率分布 $P(X)$ 的熵是其最佳比特率，即如果符号 $X \in \mathcal{X}$ 根据 $P(X)$ 出现，则

每条消息可能达到的最低平均比特数。它没有具体告诉我们如何构建最佳编码方案。它只是告诉我们，任何编码都不可能给我们每个消息提供比 $H(P)$ 更低的长期比特。

一个具体的例子如下：假设我们的消息有一个 $K = 32$ 符号的词汇表，并且每个符号在长期内具有相等的传输概率（即均匀概率分布）。一种适用于这种情况的编码方案是每个符号有 $\log_2 K$ 比特，并为每个符号分配 $\log_2 K$ 比特的某种唯一组合。事实上，事实证明，这是均匀分布场景下能想到的最有效的编码。

您现在可能已经想到，每条消息的长期平均比特数仅取决于符号的出现频率。理论上，只要场景 B 的符号遵循与场景 A 的符号相同的概率分布，场景 A 的编码方案就可以在具有不同符号集的场景 B 中重复使用（为简单起见，假设词汇量相等），并具有相同的长期效率。您可能也会想到，对于场景 B 中具有不同符号概率的消息，重复使用为场景 A 设计的最佳编码方案对于场景 B 来说总是次优的。需要明确的是，我们不需要知道在这两种场景中具体的最佳方案是什么。只要我们知道它们符号的分布，我们就可以说，如果分布不同，为场景 A 设计的最优方案对场景 B 来说将是次优的。

具体来说，如果我们将针对符号分布 $Q(X)$ 场景设计的最优方案重新用到符号分布 $P(X)$ 场景中，则所实现的长期平均每符号位数称为交叉熵 (Cross Entropy)，记为 $H(P, Q)$：

$$H(P, Q) = -\sum_{x \in \mathcal{X}} P(x) \log Q(x).$$

概括来说，熵 $H(P)$ 是在符号分布 $P(X)$ 下通过使用专门为 $P(X)$ 设计的编码方案（可能未知）可以实现的最佳长期平均每条消息位数（最优）。交叉熵 $H(P, Q)$ 是在符号分布 $P(X)$ 下通过重复使用针对符号分布 $Q(X)$ 场景进行优化的编码方案（可能未知）得到的长期平均每条消息位数（次优）。

KL 散度是我们为使用 $Q(X)$ 的最优方案而付出的代价，以平均位数来衡量，在符号实际分布为 $P(X)$ 的情况下。很容易看出这一点

$$\begin{aligned} D_{KL}(P \| Q) &= \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \\ &= \sum_{x \in \mathcal{X}} P(x) \log P(x) - \sum_{x \in \mathcal{X}} P(x) \log Q(x) \\ &= H(P, Q) - H(P). \quad \text{(平均比特数的差值.)} \end{aligned}$$

如果 P 和 Q 之间的交叉熵是 $H(P)$（因此 $D_{KL}(P||Q) = 0$），则必然意味着 $P = Q$。在机器学习中，找到一个“接近”另一个分布 P 的分布 Q 是一项常见任务。为了实现这一点，通常使用 $D_{KL}(Q||P)$ 作为要优化的损失函数。正如我们将在下面的这个问题中看到的那样，最大似然估计是一种常用的优化目标，它等同于最小化训练数据（即数据的经验分布）和模型之间的 KL 散度。

现在，我们回来证明一些 KL 散度的简单性质。

(a) 非负性.

证明：

$$\forall P, Q. \quad D_{KL}(P \| Q) \geq 0$$

和

$$D_{KL}(P\|Q) = 0 \quad \text{当且仅当} \quad P = Q.$$

[提示: 你可以使用 Jensen 不等式的以下结论：如果 f 是一个凸函数，且 X 是一个随机变量，那么 $E[f(X)] \geq f(E[X])$。此外，如果 f 严格凸的 (如果 f 的 Hessian 满足 $H \geq 0$ 则它是凸的，如果满足 $H > 0$ 则它是严格凸的；例如 $f(x) = -\log x$ 是严格凸的), 那么 $E[f(X)] = f(E[X])$ 表明 $X = E[X]$ 概率为 1，即 X 实际上是一个常数。]

(b) KL 散度的链式法则.

两个条件分布的 DL 散度 $P(X|Y), Q(X|Y)$ 定义如下：

$$D_{KL}(P(X|Y)\|Q(X|Y)) = \sum_y P(y) \left(\sum_x P(x|y) \log \frac{P(x|y)}{Q(x|y)} \right)$$

这可以被认为是 x 上相应条件分布（即 $P(X|Y = y)$ 和 $Q(X|Y = y)$ 之间）的预期 KL 散度，其中期望取自随机 y。

证明 KL 散度的如下链式法则：

$$D_{KL}(P(X,Y)\|Q(X,Y)) = D_{KL}(P(X)\|Q(X)) + D_{KL}(P(Y|X)\|Q(Y|X)).$$

(c) KL 和最大似然

考虑一个密度估计问题，假设我们给定一个训练集 $\{x^{(i)}; i = 1, \ldots, n\}$。设经验分布为 $\hat{P}(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{x^{(i)} = x\}$。（$\hat{P}(x)$ 只是训练集上的均匀分布；即，从经验分布中抽样与从训练集中随机挑选一个示例相同。）

假设我们有一些分布族 P_θ，其参数为 θ。（如果您愿意，可以将 $P_\theta(x)$ 视为 $P(x;\theta)$ 的另一种表示法。）证明找到参数 θ 的最大似然估计等同于找到与 $\hat{P}(x)$ 具有最小 KL 散度的 P_θ。即证明：

$$\arg\min_\theta D_{KL}(\hat{P}\|P_\theta) = \arg\max_\theta \sum_{i=1}^{n} \log P_\theta(x^{(i)})$$

注. 考虑部分 (b-c) 与多变量伯努利朴素贝叶斯参数估计之间的关系。在朴素贝叶斯模型中，我们假设 P_θ 具有以下形式：$P_\theta(x, y) = p(y) \prod_{i=1}^{d} p(x_i|y)$。根据 KL 散度的链式法则，我们因此得到：

$$D_{KL}(\hat{P}\|P_\theta) = D_{KL}(\hat{P}(y)\|p(y)) + \sum_{i=1}^{d} D_{KL}(\hat{P}(x_i|y)\|p(x_i|y))$$

这表明，寻找参数的最大似然/最小 KL 散度估计可分解为 $2n + 1$ 个独立优化问题：一个针对类先验 $p(y)$，另一个针对每个特征 x_i 的条件分布 $p(x_i|y)$（给定 y 的两个可能标签）。具体而言，单独寻找这些问题中每个问题的最大似然估计也会最大化联合分布的似然。（如果您知道什么是贝叶斯网络，类似的评论也适用于它们的参数估计。）

2 半监督的 EM

期望最大化 (EM) 是一种经典的无监督学习算法（即使用隐变量进行学习）。在这个问题中，我们将探索 EM 算法如何适应半监督设置。数据中有一些有标签，有一些无标签。

在标准无监督设置中，我们有 $n \in \mathbb{N}$ 个无标签数据 $\{x^{(1)}, \ldots, x^{(n)}\}$。我们希望从数据中学习 $p(x, z; \theta)$ 的参数，但未观察到 $z^{(i)}$。经典 EM 算法就是为此目的而设计的，我们通过迭代执行 E-step 和 M-step 来间接最大化难以处理的 $p(x; \theta)$，每次最大化 $p(x; \theta)$ 的可处理下限。我们的目标可以具体写成：

$$\begin{aligned}\ell_{\text{unsup}}(\theta) &= \sum_{i=1}^{n} \log p(x^{(i)}; \theta) \\ &= \sum_{i=1}^{n} \log \sum_{z^{(i)}} p(x^{(i)}, z^{(i)}; \theta)\end{aligned}$$

现在，我们将尝试将 EM 到拓展到半监督设置。假设我们有额外的 $\tilde{n} \in \mathbb{N}$ 个有标签数据 $\{(\tilde{x}^{(1)}, \tilde{z}^{(1)}), \ldots, (\tilde{x}^{(\tilde{n})}, \tilde{z}^{(\tilde{n})})\}$，其中 x 和 z 都是可观察到的。我们希望通过优化它们的加权和（使用超参数 α），同时最大化二者的总似然。具体来说，我们的半监督目标 $\ell_{\text{semi-sup}}(\theta)$ 可以写成：

$$\ell_{\text{sup}}(\theta) = \sum_{i=1}^{\tilde{n}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta)$$

$$\ell_{\text{semi-sup}}(\theta) = \ell_{\text{unsup}}(\theta) + \alpha \ell_{\text{sup}}(\theta)$$

我们可以使用与之前相同的方法和步骤推导出半监督设置的 EM 步骤。强烈建议您自己推导（无需提交）以下结果：

E-step (半监督)

对于所有 $i \in \{1, \ldots, n\}$，设置

$$Q_i^{(t)}(z^{(i)}) := p(z^{(i)} | x^{(i)}; \theta^{(t)})$$

M-step (半监督)

$$\theta^{(t+1)} := \arg\max_{\theta} \left[\sum_{i=1}^{n} \left(\sum_{z^{(i)}} Q_i^{(t)}(z^{(i)}) \log \frac{p(x^{(i)}, z^{(i)}; \theta)}{Q_i^{(t)}(z^{(i)})} \right) + \alpha \left(\sum_{i=1}^{\tilde{n}} \log p(\tilde{x}^{(i)}, \tilde{z}^{(i)}; \theta) \right) \right]$$

(a) 收敛性.

首先，我们将证明该算法最终会收敛。为了证明这一点，只需证明我们的半监督目标 $\ell_{\text{semi-sup}}(\theta)$ 随着 E-step 和 M-step 的每次迭代单调增加即可。具体来说，让 $\theta(t)$ 成为 t 个 EM-step 结束时获得的参数。证明 $\ell_{\text{semi-sup}}(\theta^{(t+1)}) \geq \ell_{\text{semi-sup}}(\theta^{(t)})$。

半监督的 GMM

现在我们将重新讨论高斯混合模型 (GMM)，以应用我们的半监督 EM 算法。让我们考虑这样一种场景，其中数据是从 $k \in \mathbb{N}$ 个高斯分布生成的，具有未知均值 $\mu_j \in \mathbb{R}^d$ 和协方差 $\Sigma_j \in \mathbb{S}_+^d$，其中 $j \in \{1, \ldots, k\}$。我们有 n 个数据点 $x^{(i)} \in \mathbb{R}^d, i \in \{1, \ldots, n\}$，每个数据点都有一个对应的隐变量 $z^{(i)} \in \{1, \ldots, k\}$，指示 $x^{(i)}$ 属于哪个分布。具体来说，$z^{(i)} \sim \text{Multinomial}(\phi)$，使得 $\sum_{j=1}^{k} \phi_j = 1$ 且对于所有 j，$\phi_j \geq 0$，并且 $x^{(i)} | z^{(i)} \sim \mathcal{N}(\mu_{z^{(i)}}, \Sigma_{z^{(i)}})$ i.i.d。因此，μ、Σ 和 ϕ 是模型参数。

我们还有额外的 $\tilde{n}$ 个数据点 $\tilde{x}^{(i)} \in \mathbb{R}^d, i \in \{1, \ldots, \tilde{n}\}$，以及 n 个相关观测变量 $\tilde{z}^{(i)} \in \{1, \ldots, k\}$，表示 $\tilde{x}^{(i)}$ 所属的分布。请注意，$\tilde{z}^{(i)}$ 是已知常数（与 $z^{(i)}$ 是未知随机变量相反）。和以前一样，我们假设 $\tilde{x}^{(i)} | \tilde{z}^{(i)} \sim \mathcal{N}(\mu_{\tilde{z}^{(i)}}, \Sigma_{\tilde{z}^{(i)}})$ i.i.d.

总之，我们有 $n+\tilde{n}$ 个数据，其中 n 是无标签的的数据点 x 和未观察到的 z，而 $\tilde{n}$ 是有标签的数据点 $\tilde{x}^{(i)}$ 和相应的观察到的标签 $\tilde{z}^{(i)}$。传统的 EM 算法被设计为只将 n 个未标记的例子作为输入，并学习模型参数 μ、Σ 和 ϕ。

我们现在的任务是将半监督 EM 算法应用于 GMM，以便利用额外的 $\tilde{n}$ 个标记示例，并提出特定于 GMM 的半监督 E-step 和 M-step 更新规则。如果需要，您可以引用 Lecture Notes 的推导和步骤。

(b) 半监督的 E-Step

明确指出哪些隐变量需要在 E-step 中重新估计。推导 E-step 的公式以重新估计所有隐变量。最终表达式必须仅包含 x, z, μ, Σ, ϕ 和通用常数。（助教注：您需要计算隐变量数据点 (i) 关于类别 j 的系数 $w_j^{(i)}$。）

(c) 半监督的 M-Step

明确指出哪些参数需要在 M-step 中重新估计。推导 M-step 的公式以重新估计所有参数。具体来说，基于半监督目标推导 $\mu^{(t+1)}$、$\Sigma^{(t+1)}$ 和 $\phi^{(t+1)}$ 的参数更新规则的闭式表达式。

3 商店聚类

助教提示：对于 (a) (b) (c) 只需要给出答案，不写过程；对于 (d) (e) (f) 过程控制在 10 行以内。

Amazing 公司将在 Megacity 地区开设第一家 k 家商店。他们知道第 i 位顾客（总共 n 位顾客）住在 $x^{(i)}$ 位置。他们将在 $\mu^{(j)}$ 位置开设第 j 家商店（其中 $j = 1, \ldots, k$）。Amazing 公司希望最小化顾客位置和商店之间的平方距离，因此他们决定使用 k-means 来选择商店位置。

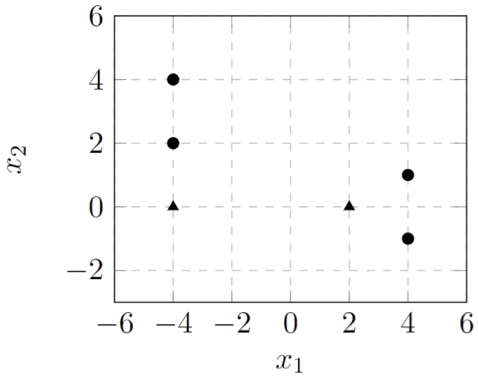
Amazing 公司聚焦商店的选址，决定将问题表述为寻找 μ 使损失函数最小化：

$$L(\mu) = \sum_{i=1}^{n} \min_{j \in 1,\ldots,k} \left\| x^{(i)} - \mu^{(j)} \right\|^2$$

对于给定的一组聚类均值 μ，在选择了数据点的最佳分配（即 $\min_j$）后，$L(\mu)$ 即为 k-means 的损失函数。

(a)

Amazing 公司在 Megacity 有四个客户，位于下图中圆圈所示的位置 (x_1, x_2)，并计划开设两家商店。他们初步规划了两个商店的位置，如下图三角形所示。初始的损失函数 $L(\mu)$ 是多少？（助教注：假设边缘上的两个点具有整数坐标。）



(b)

从以上初始值开始，Amazing 公司运行了 k-means 聚类算法直到收敛 ($L(\mu)$ 不再降低)。计算两个商店最终的坐标和对应的损失函数 $L(\mu)$。

(c)

近年来 Amazing 公司发展迅速，拥有一个由 r 条记录组成的数据库：第 i 条记录，$i = 1, \ldots, r$，含有 $c^{(i)}$ 和 $x^{(i)}$，其中 $c^{(i)}$ 是位于位置 $x^{(i)}$ 的客户数量。他们仍然希望将每个商店（索引为 $j = 1, \ldots, k$）放置在位置 $\mu^{(j)}$，以最小化客户和商店位置之间的平方距离，求出所有客户（和商店位置平方距离）之和。注意每个位置不再代表单个客户，而是拥有 $c^{(i)}$ 个客户。

使用数据库给出信息，重新定义一个损失函数 $L_C(\mu)$ 。你可以在公式中任意定义新表达式。（即定义一些新的符号，尽管在本题中可能用不上）

(d)

Amazing 公司在 One-di City 测试了他们的方法，该城市客户数量较少。他们的数据集 $\mathcal{D}$ 由一维位置 x 和客户数量 c 组成，成对出现，$(x, c) : \mathcal{D} = ((-1, 10), (1, 4))$。Amazing 公司只打算建一家商店：它应该设在哪里？

(e)

Amazing 公司在 Megacity 也有一个配送中心 (DC)，位于 x_{DC} ，为 Megacity 的所有商店供货。将货物从 DC 运输到每个商店会产生成本，该成本随着距离的平方和商店服务的客户数量的增加而增加。具体来说，对于每个商店，该成本等于分配给集群 j 的客户数量乘以从 $\mu^{(j)}$ 到 x_{DC} 的距离的平方。Amazing 公司希望最大限度地降低自己的运输成本和客户访问其商店的运输成本，后者是 (c) 中的损失。为此，他们决定最小化这两个成本的总和。

定义损失函数 $L_S(\mu)$，该函数表示 Amazing 公司力求最小化的总体损失：Amazing 公司的运输成本和客户到商店成本之和。根据问题中数量关系的需要，你可以在公式中定义新的表达式，需以方程的形式呈现（而不仅仅是文字定义）。以下是一个新的表达式定义，可能对你有帮助：$y^{(i)} = \arg\min_j \left\| x^{(i)} - \mu^{(j)} \right\|^2$ 。（即你可以在公式中使用 $y^{(i)}$，也可以自己定义一些符号包含在公式中。）

(f)

Amazing 公司回到 One-di City，看看应该把唯一的商店开在哪里。客户及数量的数据与 (d) 中相同，但现在我们还知道他们的配送中心位于 $x_{DC} = 10$。如果我们的目标是最小化 (e) 中所需的目标函数，那么他们的商店应该设在什么位置 μ?

4 扩散模型：一个变分推断的例子

给定来自感兴趣分布的观测样本 $\boldsymbol{x}$，生成模型的目标是学习建模其真实的数据分布 $p(\boldsymbol{x})$。一旦模型学会了这一点，我们就可以根据我们近似的模型随意生成新的样本。此外，在某些情况下，我们还可以使用学习到的模型来评估观测数据或采样数据的似然性。

然而，在通常情况下直接学习真实的数据分布 $p(\boldsymbol{x})$ 是一个非常困难的任务。扩散模型提供了一种方式，可以把任意分布 $p(\boldsymbol{x})$ 映射到一个已知分布 (例如正态分布) 上，然后学习这个映射的逆，间接地完成生成任务。

(a) 扩散过程

给定一个数据分布 $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0)$, 我们向这个分布中加入逐渐增大的噪声，构造如下的 Markov 过程：

$$\boldsymbol{x}_t = \sqrt{1-\beta_t}\boldsymbol{x}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \boldsymbol{0}, \boldsymbol{I})$$

其中 $t \in \{1, 2, \cdots, T\}, \beta_t \in (0, 1)$ 逐渐增大，记 $\alpha_t := 1 - \beta_t, \bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$。证明：

$$q(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I})$$

反向过程

当 $T \to \infty, \beta_t \to 1$ 时，扩散过程的最终结果 $q(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x}_T; \boldsymbol{0}, \boldsymbol{I})$, 我们把扩散过程的 Markov 链记为：

$$q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0) = \prod_{t=1}^{T} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$$

反转上面的扩散过程，我们可以从一个高斯分布中逐步去噪，" 还原" 出原本的数据：

$$q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_q(\boldsymbol{x}_t, \boldsymbol{x}_0), \boldsymbol{\Sigma}_q(t)\boldsymbol{I})$$

其中：（此为给出结论，无需证明）

$$\boldsymbol{\mu}_q(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}$$

$$\boldsymbol{\Sigma_q(t)} = \sigma_q^2\boldsymbol{I} = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{I}$$

(b) 变分下界

我们并不知道反向过程的转移概率 $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$，在这里我们学习一个 $p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ 为来建模它，在这种表示下，反向过程的联合分布：

$$p(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$$

使用最大似然估计，我们希望最大化：$\log p(\boldsymbol{x}_0)$。使用变分推断证明上面这个对数似然的 ELBO 是：

$$\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}]$$

注意，你需要把变分推断的结果表示为 ELBO+ 一个 KL 散度的形式。
(提示：对于一个未知分布 $p(\boldsymbol{x})$ 我们可以把它表示为 $\frac{p(\boldsymbol{x},\boldsymbol{z})}{p(\boldsymbol{z}|\boldsymbol{x})}$ 引入隐变量 $\boldsymbol{z}$)

(c) 优化目标

结合上面的推导，我们可以进一步把 ELBO 拆成下面三项：

$$\begin{aligned} ELBO = &\underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_0 \mid \boldsymbol{x}_1\right)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}\left(q\left(\boldsymbol{x}_T \mid \boldsymbol{x}_0\right) \| p\left(\boldsymbol{x}_T\right)\right)}_{\text{prior matching term}} \\ &- \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}\left(q\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0\right) \| p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t\right)\right)\right]}_{\text{denoising matching term}} \end{aligned}$$

其中第一项我们使用第三项在 $t=1$ 时的近似代替，第二项与要学习的参数 $\boldsymbol{\theta}$ 无关，我们忽略掉它。

考虑 (b) 中我们导出的 $\boldsymbol{\mu}_q(\boldsymbol{x}_t, \boldsymbol{x}_0)$ 与 $\boldsymbol{x}_0$ 有关，我们学习一个模型来预测它 $\hat{\boldsymbol{x}}_0 = \hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t)$

$$\boldsymbol{\mu}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{\boldsymbol{x}}_0(\boldsymbol{x}_t, t)}{1-\bar{\alpha}_t}$$

$$p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_q(t))$$

证明我们最终的优化目标是：

$$\arg_{\boldsymbol{\theta}}\max ELBO = \arg_{\boldsymbol{\theta}}\min \mathbb{E}_{t\sim\mathbb{U}[1,T], q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{2\sigma_q^2(1-\bar{\alpha}_t)^2}[\|\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t) - \boldsymbol{x}_0\|_2^2]$$

(提示:

$$\begin{aligned} &D_{\mathrm{KL}}\left(\mathcal{N}\left(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x\right) \| \mathcal{N}\left(\boldsymbol{y}; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y\right)\right) \\ =&\frac{1}{2}\left[\log \frac{|\boldsymbol{\Sigma}_y|}{|\boldsymbol{\Sigma}_x|} - d + \operatorname{tr}\left(\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_x\right) + \left(\boldsymbol{\mu}_y - \boldsymbol{\mu}_x\right)^T \boldsymbol{\Sigma}_y^{-1}\left(\boldsymbol{\mu}_y - \boldsymbol{\mu}_x\right)\right] \end{aligned}$$

，其中 d 表示维数)