概率论与机器学习

本文件旨在从最基础的概率论公式开始，逐步推导和讲解相关公式，以帮助读者深入理解概率论的核心概念和应用。

1. 概率基本规则

离散变量加和公式

$$p(X) = \sum_{Y} p(X, Y) \tag{1}$$

离散变量乘积公式

$$p(X, Y) = p(Y|X)p(X) = p(X|Y)p(Y) \tag{2}$$

连续变量加和公式

$$p(x) = \int p(x, y)dy \tag{3}$$

连续变量乘积公式

$$p(x, y) = p(y|x)p(x) = p(x|y)p(y) \tag{4}$$

2. 期望和方差

离散变量期望

$$E[f] = \sum_{x} p(x)f(x) \tag{5}$$

连续变量期望

$$E[f] = \int p(x)f(x)dx \tag{6}$$

二元离散变量期望

$$E_x[f|y] = \sum_{x} p(x|y)f(x) \tag{7}$$

方差公式

$$\begin{aligned} var[f] &= E[f(x)^2] - E[f(x)]^2 \\ var[x] &= E[x^2] = E[x]^2 \end{aligned} \tag{8}$$

协方差公式

$$cov[x, y] = E_{x,y}[(x - E[x])(y - E[y])] = E_{x,y}[xy] - E[x]E[y] \tag{9}$$

3. 贝叶斯定理

$$p(w|D) = \frac{p(D|w)p(w)}{p(D)} \tag{10}$$

4. 高斯分布

基本公式

$$\mathcal{N}(x|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{11}$$

期望和方差

$$\begin{aligned} E[x]&=\mu \\ var[x]&=\sigma^2 \end{aligned} \tag{12}$$

高维高斯分布

$$\mathcal{N}(x|\mu,\Sigma)=\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\Sigma|^{\frac{1}{2}}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \tag{13}$$

其中：

- D 是维度
- Σ 是一个 $D\times D$ 的协方差矩阵，代表方差
- μ 是一个维度为 D 的向量，代表均值
- $|\Sigma|$ 是 Σ 的行列式

似然函数

假设 $\mathbf{x}$ 是一个独立且遵循高斯分布 $\mathcal{N}(\mu,\sigma^2)$ 的样本集合，高斯分布的似然函数可以表示为：

$$L(\mu,\sigma^2)=\prod_{i=1}^{n}\mathcal{N}(x_i|\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} \tag{14}$$

对数似然函数

在实践中，最大化似然函数的对数更方便。因为对数是其参数的单调递增函数，所以函数对数的最大化等同于函数本身的最大化。取对数不仅简化了后续的数学分析，而且在数值上也有帮助，因为大量小概率的乘积很容易使计算机的数值精度下溢，而这可以通过计算对数概率的总和来解决。

$$\log L(\mu,\sigma^2)=\log\left(\prod_{i=1}^{n}\mathcal{N}(x_i|\mu,\sigma^2)\right)=\sum_{i=1}^{n}\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}\right) \tag{15}$$

进一步简化得到：

$$\log L(\mu,\sigma^2)=-\frac{N}{2}\log(2\pi)-N\log(\sigma)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{16}$$

将该函数视为 μ 的函数并求最大值，可得此时 μ 的取值为：

$$\mu_{ML}=\frac{1}{N}\sum_{n=1}^{N}x_n \tag{17}$$

- 这是 $\mathbf{x}$ (作为样本集合)的均值

将该函数视为 σ 的函数并求最大值，可得此时 σ 的取值为：

$$\sigma^2_{ML}=\frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{ML})^2 \tag{18}$$

- 这是 $\mathbf{x}$ (作为样本集合)的方差

考虑 μ_{ML} 和 σ^2_{ML} 的期望：

$$
\begin{aligned}
E[\mu_{ML}] &= \mu \\
E[\sigma^2_{ML}] &= \left(\frac{N-1}{N}\right)\sigma^2
\end{aligned} \tag{19}
$$

> **证明：**
>
> $$E[\sigma^2_{ML}] = E\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu_{ML})^2\right] \tag{20}$$
>
> 由于 μ_{ML} 是样本均值，可以写成：
>
> $$E[\sigma^2_{ML}] = E\left[\frac{1}{N}\sum_{n=1}^{N}((x_n - \mu) - (\mu_{ML} - \mu))^2\right] \tag{21}$$
>
> 展开平方项：
>
> $$E[\sigma^2_{ML}] = E\left[\frac{1}{N}\sum_{n=1}^{N}\left((x_n - \mu)^2 - 2(x_n - \mu)(\mu_{ML} - \mu) + (\mu_{ML} - \mu)^2\right)\right] \tag{22}$$
>
> 由于 $E[x_n - \mu] = 0$ 和 $E[(x_n - \mu)(\mu_{ML} - \mu)] = 0$，可以简化为：
>
> $$E[\sigma^2_{ML}] = E\left[\frac{1}{N}\sum_{n=1}^{N}(x_n - \mu)^2\right] - E\left[(\mu_{ML} - \mu)^2\right] \tag{23}$$
>
> 注意到 $E[(\mu_{ML} - \mu)^2] = \frac{\sigma^2}{N}$，所以：
>
> $$E[\sigma^2_{ML}] = \sigma^2 - \frac{\sigma^2}{N} = \left(\frac{N-1}{N}\right)\sigma^2 \tag{24}$$

因此将方差乘以一个系数

$$\tilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{ML} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{ML})^2 \tag{25}$$

5. 从概率角度看曲线拟合

曲线拟合问题的目标是能够根据一组训练数据，给定输入变量 x 的某个新值，对目标变量 $$t$ 进行预测，这些数据包含 N 个输入值 $x = （x_{1},⋯,x_{N}) ^{T}$ 及其相应的目标值 $t = （t_{1},⋯,t_{N})^{T}$。我们可以使用概率分布来表示我们对目标变量值的不确定性。为此，我们将假设，给定 x 的值，相应的 t 值具有高斯分布，其平均值等于 $y(x,w) = \sum_{j=0}^{N}w_{j}x^{j}$ 给出的多项式曲线的 $y(x,w)$ 的值。因此我们有：

$$p(t|x, w, \beta) = \mathcal{N}(t|y(x, w), \beta^{-1}) \tag{26}$$

其中 $\beta = \frac{1}{\sigma^2}$ 被称为精度

则我们可以得到其似然函数：

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N}\mathcal{N}(t_n|y(x_n\mathbf{w}), \beta^{-1}) \tag{27}$$

将其变为对数似然函数：

$$\log p(\mathbf{t}|\mathbf{x},\mathbf{w},\beta)=\sum_{n=1}^{N}\log\mathcal{N}(t_n|y(x_n,\mathbf{w}),\beta^{-1})=-\frac{N}{2}\log(2\pi)+\frac{N}{2}\log\beta-\frac{\beta}{2}\sum_{n=1}^{N}(t_n-y(x_n,\mathbf{w}))^2 \quad (28)$$

首先考虑确定多项式系数的最大似然解，该系数将用 w_{ML} 表示。这些是通过以 w 为自变量而将对数似然函数最大化来确定的。为此，我们可以省略上式右侧的前两项，因为它们与 w 无关。此外，我们还注意到，用正常数系数缩放对数似然不会改变 w 在函数最大时的取值，因此我们可以将系数 $\frac{\beta}{2}$ 替换为 $\frac{1}{2}$ 。最后，我们可以等效地最小化负对数似然，而不是最大化对数似然。因此，我们看到，就确定w而言，最大化似然等同于最小化平方和误差函数。因此，平方和误差函数是在高斯噪声分布的假设下最大化似然的结果。

对于 β 而言，我们同样可以通过最小化下式来得到其最大值：

$$\frac{1}{\beta_{ML}}=\frac{1}{N}\sum_{n=1}^{N}y((x_n,\mathbf{w}_{ML})-t_n)^2 \quad (29)$$

确定了参数 $\mathbf{w}$ 和 β 之后，我们现在就可以预测 x 的新值了。因为我们现在有一个概率模型，这个模型是用预测分布来表示的，该分布给出了 t 上的概率分布，而不是简单的点估计。

将前面求得的参数代回似然函数可得：

$$p(\mathbf{t}|\mathbf{x},\mathbf{w_{ML}},\beta_{ML})=\prod_{n=1}^{N}\mathcal{N}(t_n|y(x_n\mathbf{w_{ML}}),\beta_{ML}^{-1}) \quad (30)$$

在确定了模型参数后，我们可以对新的输入数据进行预测。预测分布可以表示为：

$$p(t|x,\mathbf{w_{ML}},\beta_{ML})=\mathcal{N}(t|y(x,\mathbf{w_{ML}}),\beta_{ML}^{-1}) \quad (31)$$

其中， $y(x,\mathbf{w_{ML}})$ 是基于最大似然估计的多项式曲线， β_{ML}^{-1} 是噪声的方差。

通过这个预测分布，我们不仅可以得到预测值的期望，还可以得到预测值的不确定性。

现在让我们朝着更贝叶斯的方法迈出一步，并引入多项式系数 w 的先验分布。为简单起见，我们考虑以下形式的高斯分布：

$$p(\mathbf{w}|\alpha)=\mathcal{N}(\mathbf{w}|0,\alpha^{-1}\mathbf{I})=\left(\frac{\alpha}{2\pi}\right)^{\frac{M+1}{2}}e^{-\frac{\alpha}{2}\mathbf{w}^T\mathbf{w}} \quad (32)$$

其中：

- α 是分布的精度，是超参数。
- $M+1$ 是 M 阶多项式的向量 $\mathbf{w}$ 中的元素总数。

贝叶斯定理给出了后验分布的表达式：

$$p(\mathbf{w}\mid\mathbf{x},\mathbf{t},\alpha,\beta)=\frac{p(\mathbf{t}\mid\mathbf{x},\mathbf{w},\beta)p(\mathbf{w}\mid\alpha)}{p(\mathbf{t}\mid\mathbf{x},\alpha,\beta)} \quad (33)$$

其中：

- $p(\mathbf{w}\mid\mathbf{x},\mathbf{t},\alpha,\beta)$ 是参数 $\mathbf{w}$ 的后验分布，条件是数据 $\mathbf{x}$ 和 $\mathbf{t}$ ，以及超参数 α （先验的精度）和 β （似然的精度）。
- $p(\mathbf{t}\mid\mathbf{x},\mathbf{w},\beta)$ 是似然函数，表示给定参数 $\mathbf{w}$ 和数据 $\mathbf{x}$ ，目标变量 $\mathbf{t}$ 的概率分布。
- $p(\mathbf{w}\mid\alpha)$ 是参数 $\mathbf{w}$ 的先验分布，通常由超参数 α 控制。
- $p(\mathbf{t}\mid\mathbf{x},\alpha,\beta)$ 是证据（evidence）或边缘似然（marginal likelihood），它是 $\mathbf{w}$ 上的积分：

$$p(\mathbf{t}\mid\mathbf{x},\alpha,\beta)=\int p(\mathbf{t}\mid\mathbf{x},\mathbf{w},\beta)p(\mathbf{w}\mid\alpha)\,d\mathbf{w} \quad (34)$$

由于证据 $p(\mathbf{t}\mid\mathbf{x},\alpha,\beta)$ 是一个归一化常数（与 $\mathbf{w}$ 无关），在计算后验分布的形状时可以省略。因此，后验分布与以下表达式的乘积成正比：

$$p(\mathbf{w}\mid\mathbf{x},\mathbf{t},\alpha,\beta)\propto p(\mathbf{t}\mid\mathbf{x},\mathbf{w},\beta)p(\mathbf{w}\mid\alpha) \quad (35)$$

我们现在可以通过找到给定数据的最可能 $\mathbf{w}$ 值来确定 $\mathbf{w}$ ，换句话说，通过最大化后验分布来确定 $\mathbf{w}$ 。这种技术称为最大后验，或简称为 MAP。我们将似然函数都取负对数后把与 $\mathbf{w}$ 无关的项全部丢掉，最后得到优化目标，即最小化下式：

$$\frac{\beta}{2}\sum_{n=1}^{N}\{y(x_n,\mathbf{w})-t_n\}^2+\frac{\alpha}{2}\mathbf{w}^T\mathbf{w} \tag{36}$$

在曲线拟合问题中，我们得到训练数据 $\mathbf{x}$ 和 $\mathbf{t}$ ，以及一个新的测试点 x ，我们的目标是预测 t 的值。因此，我们希望评估预测分布 $p(t|x,\mathbf{x},\mathbf{t})$ 。在这里，我们将假设参数 α 和 β 是固定的，并且事先已知。

应用乘积公式可得：

$$p(t|x,\mathbf{x},\mathbf{w})=\int p(t|x,\mathbf{w})p(\mathbf{w}|\mathbf{x},\mathbf{t})d\mathbf{w} \tag{37}$$

为了符号简便，此处将所有的 α 和 β 都删去了。最终结果是预测分布可以写成如下形式：

$$p(t\mid x,\mathbf{x},\mathbf{t})=\mathcal{N}(t\mid m(x),s^2(x)) \tag{38}$$

- 其中均值和方差分别由以下公式给出：

$$m(x)=\beta\phi(x)^T\mathbf{S}\sum_{n=1}^{N}\phi(x_n)t_n \tag{39}$$

$$s^2(x)=\beta^{-1}+\phi(x)^T\mathbf{S}\phi(x) \tag{40}$$

- 矩阵 $\mathbf{S}$ 由以下表达式给出：

$$\mathbf{S}^{-1}=\alpha\mathbf{I}+\beta\sum_{n=1}^{N}\phi(x_n)\phi(x_n)^T \tag{41}$$

- 其中 $\mathbf{I}$ 是单位矩阵
- 向量 $\phi(x)$ 的元素为 $\phi_i(x)=x^i$，其中 $i=0,\ldots,M$。

6. 信息论

离散变量的熵

我们首先考虑一个离散随机变量 x ，然后询问当我们观察到该变量的特定值时，我们收到了多少信息。信息量可以看作是了解到 x 的值时的“惊讶程度”。如果我们被告知一个极不可能的事件刚刚发生，我们将收到比我们被告知某个非常可能的事件刚刚发生的更多信息，如果我们知道该事件肯定会发生，我们将不会收到任何信息。因此，我们对信息内容的度量将取决于概率分布 $p(x)$ ，因此我们寻找一个量 $h(x)$ ，它是概率 $p(x)$ 的单调递减函数，用于表示信息内容。

如果我们有两个不相关的事件 x 和 y ，那么观察它们两个事件所获得的信息应该是分别从它们中每个事件获得的信息之和，因此 $h(x,y)=h(x)+h(y)$ 。两个不相关的事件在统计上是独立的，因此 $p(x,y)=p(x)p(y)$ 。从以上两个关系中，很容易证明 $h(x)$ 必须由 $p(x)$ 的负对数给出，因此我们有：

$$h(x)=-\log_2 p(x) \tag{42}$$

其中负号确保信息量 $h(x)\geq 0$ 。对数的底的选择是任意的，目前我们将采用信息论中流行的惯例，即使用以 2 为底的对数。在这种情况下，正如我们稍后将看到的，h（x） 的单位是 bit

现在假设发送者希望将随机变量的值传输给接收者。它们在此过程中传输的平均信息量是通过对分布 $p(x)$ 的期望值获得的，由下式给出：

$$H[x]=-\sum_{x}p(x)\log_2 p(x) \tag{43}$$

这个重要的量称为随机变量 x 的熵。请注意， $\lim_{p\to 0}p\ln p=0$ ，因此每当我们遇到 $p(x)=0$ 时，我们应该取 $p(x)\ln p(x)=0$ 。

接下来为了与后文衔接流畅，将对数的底数改为 e ，此时熵的单位为 $nats$ ，其与 bit 相差 $\ln 2$ 倍。

即对于一个离散变量 X ， $p(X=x_i)=p_i$ ，我们能够得知 X 的熵为：

$$H[p]=-\sum_i p(x_i)\ln(p(x_i)) \tag{44}$$

最大熵可以通过使用拉格朗日乘子最大化 H 来增强对概率的归一化约束来找到。之所以这么做，是因为：

- 约束优化问题：熵的最大化需要满足概率和为1的约束条件。直接对熵 H 求极值可能违反这一约束，因此必须将约束纳入优化过程。
- 通过引入拉格朗日乘子将约束条件整合到目标函数中，构造新的函数，从而将问题转化为无约束优化。此时，极值点需同时满足：
  - 熵的极值条件；
  - 概率归一化约束。

因此，我们最大化

$$\tilde{H}=H[p]+\lambda\left(\sum_i p(x_i)-1\right) \tag{45}$$

显然当变量均匀分布的时候，整个熵最大，此时假设离散变量 X 被分为 M 个相同的种类（盒子），那么有：

$$p(x_i)=\frac{1}{M}\quad ,\quad H=\ln M \tag{46}$$

在验证其为最大值的时候，我们对这个熵求二阶导，如果二阶导为负数，则为极大值， $\tilde{H}$ 的二阶导如下：

$$\frac{\partial^2\tilde{H}}{\partial p(x_i)\partial p(x_j)}=-I_{ij}\frac{1}{p(x_i)} \tag{47}$$

其中

- I_{ij} 是单位矩阵中的元素

该二阶导的证明过程较为简单，先对 x_i 求导后对 x_j 求导，分析 i 和 j 是否相等即可。

连续变量的熵

要将熵扩展到连续变量 x，需离散化处理：

步骤1：离散化连续变量

将 x 分为宽度为 Δ 的区间，根据积分中值定理，每个区间存在 x_i 使得：

$$\int_{i\Delta}^{(i+1)\Delta}p(x)\,dx=p(x_i)\Delta. \tag{48}$$

此时，连续变量被量化为离散分布，概率为 $p(x_i)\Delta$。

步骤2：计算离散化后的熵

离散分布的熵为：

$$H_\Delta=-\sum_i \underbrace{p(x_i)\Delta}_{\text{离散概率}}\ln\left(p(x_i)\Delta\right). \tag{49}$$

展开后：

$$H_\Delta = -\sum_i p(x_i)\Delta \ln p(x_i) - \sum_i p(x_i)\Delta \ln \Delta. \tag{50}$$

由于 $\sum_i p(x_i)\Delta = 1$（概率归一化），第二项简化为：

$$-\ln \Delta \sum_i p(x_i)\Delta = -\ln \Delta. \tag{51}$$

因此，熵表达式为：

$$H_\Delta = -\sum_i p(x_i)\Delta \ln p(x_i) - \ln \Delta. \tag{52}$$

步骤3： 定义连续熵（微分熵）

为得到连续变量的熵，需取极限 $\Delta \to 0$，但注意到：

- 第一项 $-\sum_i p(x_i)\Delta \ln p(x_i)$ 在极限下趋向于积分：

$$-\int p(x) \ln p(x)\, dx. \tag{53}$$

- 第二项 $-\ln \Delta$ 随 $\Delta \to 0$ 趋向于正无穷，但此项与分布的“分辨率”相关，无实际物理意义。因此，**省略此项**，定义连续熵（微分熵）为：

$$H = -\int p(x) \ln p(x)\, dx. \tag{54}$$

---

4. 关键解释

- **为何省略** $-\ln \Delta$：该项表示离散化引入的信息量，与分布本身无关。在连续极限下，其趋向于无穷，但实际应用中关注的是熵的相对差异，而非绝对量。
- **微分熵的性质**：
  微分熵不具有离散熵的“非负性”，可能为负值（例如，当 $p(x)$ 在某些区域高度集中导致该处概率密度函数大于1时）。

结论

通过离散化连续变量并取极限，最终定义了微分熵：

$$H = -\int p(x) \ln p(x)\, dx. \tag{55}$$

此过程结合了离散熵的优化方法和连续分析的极限思想，揭示了熵从离散到连续的扩展逻辑。

通过引入拉格朗日乘子，我们能够得到如下结论：

- **使微分熵最大化的分布是高斯分布，其微分熵如下：**

$$H[x] = \frac{1}{2}(1 + \ln(2\pi\sigma^2)) \tag{56}$$

相对熵（KL散度）

1. 相对熵的定义

相对熵用于衡量两个概率分布 $p(x)$ 和 $q(x)$ 之间的差异。假设我们用 $q(x)$ 近似真实分布 $p(x)$，则在使用 $q(x)$ 编码 x 时，所需的额外信息量（以 nats 为单位）为：

$$\mathrm{KL}(p\|q) = -\int p(x)\ln q(x)\,dx - \left(-\int p(x)\ln p(x)\,dx\right). \tag{57}$$

展开后得到：

$$\mathrm{KL}(p\|q) = -\int p(x)\ln\left\{\frac{q(x)}{p(x)}\right\}dx. \tag{58}$$

这就是 **Kullback-Leibler 散度** 的定义。

---

2. KL 散度的非负性

要证明 $\mathrm{KL}(p\|q) \geq 0$，且等号成立当且仅当 $p(x) = q(x)$，需要用到 **凸函数（Convex Function）** 的性质。

(1) 凸函数的定义

函数 $f(x)$ 是凸函数，如果对于任意 $x = a$ 和 $x = b$，以及 $0 \leq \lambda \leq 1$，满足：

$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda)f(b). \tag{59}$$

即函数图像上的任意弦位于函数图像之上。

(2) 应用到 KL 散度

选择凸函数 $f(x) = -\ln(x)$，并利用 **Jensen 不等式**（凸函数的性质）：

$$f\left(\mathbb{E}[x]\right) \leq \mathbb{E}[f(x)]. \tag{60}$$

将 $x = \frac{q(x)}{p(x)}$ 代入，得到：

$$-\ln\left(\int p(x)\cdot\frac{q(x)}{p(x)}\,dx\right) \leq \int p(x)\cdot\left(-\ln\frac{q(x)}{p(x)}\right)dx. \tag{61}$$

化简左边：

$$-\ln\left(\int q(x)\,dx\right) = -\ln(1) = 0. \tag{62}$$

右边即为 KL 散度的定义：

$$\int p(x)\cdot\left(-\ln\frac{q(x)}{p(x)}\right)dx = \mathrm{KL}(p\|q). \tag{63}$$

因此，得到：

$$0 \leq \mathrm{KL}(p\|q). \tag{64}$$

等号成立当且仅当 $\frac{q(x)}{p(x)} = 1$，即 $p(x) = q(x)$。

---

3. KL 散度的不对称性

KL 散度不满足对称性，即：

$$\mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p). \tag{65}$$

这是因为 KL 散度的定义中，$p(x)$ 和 $q(x)$ 的作用不同，导致结果不对称。

---

4. 关键公式总结

1. **KL 散度定义**：

$$\mathrm{KL}(p\|q) = -\int p(x)\ln\left\{\frac{q(x)}{p(x)}\right\}dx. \tag{66}$$

2. **非负性证明**：

$$\mathrm{KL}(p\|q) \geq 0, \quad \text{等号成立当且仅当 } p(x) = q(x). \tag{67}$$

3. **不对称性**：

$$\mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p). \tag{68}$$

结论

- KL 散度衡量了两个分布 $p(x)$ 和 $q(x)$ 之间的差异，具有非负性。
- 其不对称性表明，KL 散度不是距离度量，但在信息论和模式识别中有重要应用。
- 通过凸函数和 Jensen 不等式，可以严格证明 KL 散度的非负性。

密度估计与互信息

1. KL 散度与密度估计

- **核心思想**：KL 散度衡量了两个概率分布 $p(\mathbf{x})$ 和 $q(\mathbf{x})$ 之间的差异。在密度估计问题中，我们希望通过一个参数化分布 $q(\mathbf{x}|\boldsymbol{\theta})$ 来近似未知的真实分布 $p(\mathbf{x})$。
- **优化目标**：通过最小化 KL 散度 $\mathrm{KL}(p\|q)$ 来确定参数 $\boldsymbol{\theta}$。由于 $p(\mathbf{x})$ 未知，我们使用训练数据 $\{\mathbf{x}_n\}_{n=1}^N$，这些数据是从 $p(x)$ 中独立同分布采样得到的。

$$\mathrm{KL}(p\|q) = \int p(\mathbf{x})\ln\left(\frac{p(\mathbf{x})}{q(\mathbf{x}|\boldsymbol{\theta})}\right)d\mathbf{x}. \tag{69}$$

展开后：

$$\mathrm{KL}(p\|q) = \int p(\mathbf{x})\ln p(\mathbf{x})\,d\mathbf{x} - \int p(\mathbf{x})\ln q(\mathbf{x}|\boldsymbol{\theta})\,d\mathbf{x}. \tag{70}$$

其中，第一项是 $p(\mathbf{x})$ 的熵，与 $\boldsymbol{\theta}$ 无关；第二项是交叉熵，依赖于 $\boldsymbol{\theta}$。因此我们将第一项删去之后得到

$$\mathrm{KL}(p\|q) \simeq -\sum_{n=1}^{N}\{\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\}. \tag{71}$$

其中，第二项与 $\boldsymbol{\theta}$ 无关，因此最小化 KL 散度等价于最大化似然函数：

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{n=1}^{N}\ln q(\mathbf{x}_n|\boldsymbol{\theta}) \tag{72}$$

---

2. 互信息（Mutual Information）

- **定义**：互信息衡量两个随机变量 $\mathbf{x}$ 和 $\mathbf{y}$ 之间的依赖关系，定义为：

$$I[\mathbf{x},\mathbf{y}] = \mathrm{KL}(p(\mathbf{x},\mathbf{y})\|p(\mathbf{x})p(\mathbf{y})). \tag{73}$$

  展开后：

$$I[\mathbf{x},\mathbf{y}] = -\iint p(\mathbf{x},\mathbf{y})\ln\left(\frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x},\mathbf{y})}\right)d\mathbf{x}d\mathbf{y}. \tag{74}$$

- **性质**：
  - 互信息 $I[\mathbf{x}, \mathbf{y}] \geq 0$，等号成立当且仅当 $\mathbf{x}$ 和 $\mathbf{y}$ 独立。
  - 互信息与条件熵的关系：

$$I[\mathbf{x}, \mathbf{y}] = H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}] = H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}]. \tag{75}$$

    表示通过观测 $\mathbf{y}$ 减少的关于 $\mathbf{x}$ 的不确定性（或反之）。
- **贝叶斯视角**：将 $p(\mathbf{x})$ 视为先验分布，$p(\mathbf{x}|\mathbf{y})$ 视为后验分布，互信息表示观测 $\mathbf{y}$ 后对 $\mathbf{x}$ 不确定性的减少量。

---

3. 关键公式总结

1. **KL 散度与似然函数**：

$$\mathrm{KL}(p\|q) \simeq \sum_{n=1}^{N} \{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\}. \tag{76}$$

   最小化 KL 散度等价于最大化似然函数。
2. **互信息定义**：

$$I[\mathbf{x}, \mathbf{y}] = \mathrm{KL}(p(\mathbf{x}, \mathbf{y})\|p(\mathbf{x})p(\mathbf{y})). \tag{77}$$

3. **互信息与条件熵的关系**：

$$I[\mathbf{x}, \mathbf{y}] = H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}] = H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}]. \tag{78}$$

---

4. 结论

- KL 散度在密度估计中用于衡量模型分布与真实分布之间的差异，最小化 KL 散度等价于最大化似然函数。
- 互信息是 KL 散度的一种特殊形式，用于衡量两个随机变量之间的依赖关系，具有非负性，且在变量独立时为零。
- 通过互信息，可以量化观测一个变量对另一个变量不确定性的减少量。