指数族分布 (Exponential Family Distribution)

1. 背景 (Background)

指数族分布 (Exponential Family Distribution) 是一类在概率统计和机器学习中非常重要的分布族。许多常见的概率分布，如高斯分布 (Gaussian)、伯努利分布 (Bernoulli)、二项分布 (Binomial)、泊松分布 (Poisson)、Beta 分布、Dirichlet 分布、Gamma 分布等，都属于指数族分布。

1.1 定义 (Definition)

一个分布如果可以写成如下形式，则称为指数族分布：

$$P(x|\eta) = h(x)\exp(\eta^T\phi(x) - A(\eta)) \tag{1}$$

其中各个量的含义如下：

- η **(Eta)**: 自然参数 (Natural Parameter) 或 典范参数 (Canonical Parameter)。
- x: 随机变量 (Random Variable)。
- $\phi(x)$: 充分统计量 (Sufficient Statistic)。对于许多常见分布，$\phi(x) = x$。
- $A(\eta)$: 对数配分函数 (Log Partition Function)。这是指数族分布中极其重要的一项。
    - **配分函数 (Partition Function)**: $Z(\eta)$。它的作用是归一化，确保概率密度函数的积分等于 1。

      $$P(x|\theta) = \frac{1}{Z}\hat{P}(x|\theta) \tag{2}$$

      其中 $Z(\eta) = \int h(x)\exp(\eta^T\phi(x))dx$。
    - **对数配分函数**: $A(\eta) = \log Z(\eta)$，即 $\exp(A(\eta)) = Z(\eta)$。
    - **推导关系**:

      $$\begin{aligned} P(x|\eta) &= h(x)\exp(\eta^T\phi(x) - A(\eta)) \\ &= h(x)\exp(\eta^T\phi(x))\exp(-A(\eta)) \\ &= \frac{1}{\exp(A(\eta))}h(x)\exp(\eta^T\phi(x)) \\ &= \frac{1}{Z(\eta)}\hat{P}(x|\eta) \end{aligned} \tag{3}$$

      其中 $\hat{P}(x|\eta) = h(x)\exp(\eta^T\phi(x))$ 是未归一化的概率密度。
- $h(x)$: 不依赖于参数 η 的函数 (Base Measure)。

1.2 为什么要研究指数族分布？ (Why Exponential Family?)

指数族分布在机器学习的多个领域都有核心地位：

1. **充分统计量 (Sufficient Statistic)**: $\phi(x)$ 是数据的充分统计量，这意味着它包含了估计参数 η 所需的所有信息。
    - **在线学习 (Online Learning)**: 充分统计量的性质使得指数族分布非常适合在线学习。我们只需要维护一个充分统计量的累加值，而不需要存储所有历史数据。

2. **广义线性模型 (Generalized Linear Models, GLM)**: 指数族分布是 GLM 的构建基石。
    - **结构**: 线性组合 $w^T x \rightarrow$ Link Function (激活函数的反函数) $\rightarrow$ 指数族分布。
    - **模型实例**:
        - 线性回归 (Linear Regression): $y|x \sim \mathcal{N}(\mu, \Sigma)$ (Gaussian)。
        - 分类 (Classification): $y|x \sim \text{Bernoulli}$ (Logistic Regression)。
        - 计数模型 (Count Data): $y|x \sim \text{Poisson}$ (Poisson Regression)。
3. **概率图模型 (Probabilistic Graphical Models)**:
    - 在无向图模型 (如马尔可夫随机场) 中，势函数常采用指数形式。
    - **受限玻尔兹曼机 (Restricted Boltzmann Machine, RBM)** 是其典型代表。
4. **变分推断 (Variational Inference)**: 在变分推断中，平均场近似 (Mean Field Approximation) 通常假设变分后验分布属于指数族分布，从而大大简化推导。
5. **共轭先验 (Conjugate Priors)**: 指数族分布与其共轭先验具有良好的代数性质，使得贝叶斯推断中的后验分布计算变得封闭 (Closed-form)。
    - 似然 (Likelihood) $\times$ 先验 (Prior) $\propto$ 后验 (Posterior)。
6. **最大熵原理 (Maximum Entropy)**: 在给定某些约束条件下 (如均值、方差已知)，熵最大的分布属于指数族分布。
    - 这对应于**无信息先验 (Non-informative Prior)** 的概念，即在没有更多先验知识的情况下，选择熵最大的分布是最“客观”的假设。

1.3 贝叶斯推断与先验选择 (Bayesian Inference & Prior Selection)

在贝叶斯推断中，核心是通过观测数据 x 来推断参数 z 的后验分布：

$$P(z|x) = \frac{P(x|z)P(z)}{P(x)} \propto P(x|z) \cdot P(z) \tag{4}$$

即：**后验 (Posterior)** $\propto$ **似然 (Likelihood)** $\times$ **先验 (Prior)**。

实例 (Example):

- 如果我们选择 **Beta 分布** 作为先验 $P(z)$。
- 选择 **二项分布 (Binomial)** 作为似然 $P(x|z)$。
- 那么推导出的后验 $P(z|x)$ 依然服从 **Beta 分布**。
- 这就是共轭先验的一个典型例子。

先验分布的选择 (Choice of Priors)

根据板书，常见的先验选择主要有以下几种角度：

1. **共轭先验 (Conjugate Prior)**:
    - **特点**: 后验分布与先验分布属于同一个分布族。
    - **优势**: **计算上的便利 (Computational Convenience)**。它可以避免复杂的积分运算，直接通过代数更新参数即可得到后验分布。
2. **最大熵先验 (Maximum Entropy Prior)**:
    - **特点**: 在满足已知约束条件的情况下，选择熵最大的分布。

- **意义**: **无信息先验 (Non-informative Prior)**。它代表了在该约束下最“随机”、最不预设偏见的分布选择。

3. **Jeffreys Prior**:
   - **特点**: 它是无信息先验的一种特殊形式，基于 Fisher 信息矩阵构造。
   - **优势**: **平移不变性 (Translation Invariance)**。即无论参数如何变换（例如从 σ 变换为 σ^2），推导出的先验性质保持一致。

2. 高斯分布的指数族形式 (Gaussian Distribution as Exponential Family)

我们将一元高斯分布 (Univariate Gaussian) 转换为指数族分布的标准形式，从而识别出其自然参数 η、充分统计量 $\phi(x)$ 和 对数配分函数 $A(\eta)$。

2.1 推导过程 (Derivation)

已知一元高斯分布的概率密度函数为：

$$P(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{5}$$

我们将指数项展开：

$$\begin{aligned} P(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2\sigma^2}(x^2-2\mu x+\mu^2)\right\} \\ &= \exp\left\{\log(2\pi\sigma^2)^{-1/2}\right\}\exp\left\{-\frac{1}{2\sigma^2}(x^2-2\mu x)-\frac{\mu^2}{2\sigma^2}\right\} \\ &= \exp\left\{-\frac{1}{2\sigma^2}\begin{pmatrix}-2\mu & 1\end{pmatrix}\begin{pmatrix}x\\x^2\end{pmatrix}-\frac{\mu^2}{2\sigma^2}-\frac{1}{2}\log(2\pi\sigma^2)\right\} \\ &= \exp\left\{\begin{pmatrix}\frac{\mu}{\sigma^2} & -\frac{1}{2\sigma^2}\end{pmatrix}\begin{pmatrix}x\\x^2\end{pmatrix}-\left(\frac{\mu^2}{2\sigma^2}+\frac{1}{2}\log(2\pi\sigma^2)\right)\right\} \end{aligned} \tag{6}$$

将其对应到指数族分布的标准形式：

$$P(x|\eta) = h(x)\exp\left\{\eta^T\phi(x)-A(\eta)\right\} \tag{7}$$

2.2 参数对应 (Parameter Mapping)

1. 自然参数 η (Natural Parameter) 与 充分统计量 $\phi(x)$ (Sufficient Statistic)

我们将 x 和 x^2 的系数提取出来，可以构造向量：

$$\eta = \begin{pmatrix}\eta_1\\\eta_2\end{pmatrix} = \begin{pmatrix}\frac{\mu}{\sigma^2}\\-\frac{1}{2\sigma^2}\end{pmatrix} \tag{8}$$

$$\phi(x) = \begin{pmatrix}x\\x^2\end{pmatrix} \tag{9}$$

这样，$\eta^T\phi(x) = \eta_1 x + \eta_2 x^2 = \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2$，与推导中的第一部分吻合。

2. 从 η 反解原始参数 $\theta = (\mu, \sigma^2)$

由 $\eta_2 = -\frac{1}{2\sigma^2}$ 可得：

$$\sigma^2 = -\frac{1}{2\eta_2} \tag{10}$$

将 σ^2 代入 $\eta_1 = \frac{\mu}{\sigma^2}$：

$$\mu = \eta_1\sigma^2 = \eta_1\left(-\frac{1}{2\eta_2}\right) = -\frac{\eta_1}{2\eta_2} \tag{11}$$

总结：

$$\begin{cases}\mu = -\frac{\eta_1}{2\eta_2} \\ \sigma^2 = -\frac{1}{2\eta_2}\end{cases} \tag{12}$$

(注意：为了保证方差 $\sigma^2 > 0$，必须要求 $\eta_2 < 0$)

3. 对数配分函数 $A(\eta)$ (Log Partition Function)

剩余的常数项即为 $A(\eta)$：

$$A(\theta) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2) \tag{13}$$

我们将 μ, σ^2 用 η 替换：

$$\begin{aligned} A(\eta) &= \frac{(-\frac{\eta_1}{2\eta_2})^2}{2(-\frac{1}{2\eta_2})} + \frac{1}{2}\log\left(2\pi\cdot(-\frac{1}{2\eta_2})\right) \\ &= \frac{\frac{\eta_1^2}{4\eta_2^2}}{-\frac{1}{\eta_2}} + \frac{1}{2}\log\left(-\frac{\pi}{\eta_2}\right) \\ &= -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log\left(-\frac{\pi}{\eta_2}\right) \end{aligned} \tag{14}$$

4. Base Measure $h(x)$

在前面的推导中，我们把所有项都放进了指数里，所以：

$$h(x) = 1 \tag{15}$$

2.3 总结：高斯分布的指数族形式 (Summary)

根据上述推导，我们将高斯分布的各个指数族组件汇总如下（如板书所示）：

1. **自然参数 (Natural Parameter)**:

$$\eta = \begin{pmatrix}\eta_1 \\ \eta_2\end{pmatrix} = \begin{pmatrix}\frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2}\end{pmatrix} \tag{16}$$

2. **充分统计量 (Sufficient Statistic)**:

$$\phi(x) = \begin{pmatrix}x \\ x^2\end{pmatrix} \tag{17}$$

3. **对数配分函数 (Log Partition Function)**:

$$A(\eta) = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log(-\frac{\pi}{\eta_2}) \tag{18}$$

4. **概率密度形式**:

$$P(x|\eta) = \exp\left\{\eta^T \phi(x) - A(\eta)\right\} \tag{19}$$

3. 对数配分函数的性质 (Properties of Log Partition Function)

对数配分函数 $A(\eta)$ 不仅起到了归一化的作用，它还蕴含了分布的重要矩信息（Moments）。

3.1 期望与一阶导数 (Expectation & First Derivative)

结论:

$$A'(\eta) = \mathbb{E}_{P(x|\eta)}[\phi(x)] \tag{20}$$

即：对数配分函数的一阶导数等于充分统计量的数学期望。

推导 (Derivation):
根据定义，我们要证明：

$$\frac{\partial A(\eta)}{\partial \eta} = \int \phi(x) P(x|\eta) dx \tag{21}$$

从配分函数的定义出发：

$$\exp(A(\eta)) = \int h(x) \exp(\eta^T \phi(x)) dx \tag{22}$$

两边同时对 η 求导：

$$\begin{aligned} \text{左边} &= \frac{\partial}{\partial \eta} \exp(A(\eta)) = \exp(A(\eta)) \cdot A'(\eta) \\ \text{右边} &= \frac{\partial}{\partial \eta} \int h(x) \exp(\eta^T \phi(x)) dx \\ &= \int h(x) \frac{\partial}{\partial \eta} \exp(\eta^T \phi(x)) dx \quad (\text{交换积分与微分}) \\ &= \int h(x) \exp(\eta^T \phi(x)) \cdot \phi(x) dx \end{aligned} \tag{23}$$

令左右两边相等：

$$\exp(A(\eta)) \cdot A'(\eta) = \int h(x) \exp(\eta^T \phi(x)) \cdot \phi(x) dx \tag{24}$$

将 $\exp(A(\eta))$ 除到右边：

$$\begin{aligned} A'(\eta) &= \frac{1}{\exp(A(\eta))} \int h(x) \exp(\eta^T \phi(x)) \cdot \phi(x) dx \\ &= \int \frac{h(x) \exp(\eta^T \phi(x))}{\exp(A(\eta))} \cdot \phi(x) dx \\ &= \int P(x|\eta) \cdot \phi(x) dx \\ &= \mathbb{E}_{P(x|\eta)}[\phi(x)] \end{aligned} \tag{25}$$

得证。

3.2 方差与二阶导数 (Variance & Second Derivative)

结论:

$$A''(\eta) = \mathrm{Var}[\phi(x)] \tag{26}$$

即：对数配分函数的二阶导数等于充分统计量的方差（或协方差矩阵）。

3.3 凸性 (Convexity)

由于方差（或协方差矩阵）总是半正定的 ($A''(\eta) \succeq 0$)，这意味着：

$A(\eta)$ **是凸函数 (Convex Function)。**

3.4 实例：验证高斯分布的均值 (Example: Verifying Gaussian Mean)

利用上述性质，我们可以直接通过 $A(\eta)$ 对 η 求导来推导高斯分布的期望（均值）。

已知:

1. **充分统计量**: $\phi(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$
   - 这意味着 $\mathbb{E}[\phi(x)] = \begin{pmatrix} \mathbb{E}[x] \\ \mathbb{E}[x^2] \end{pmatrix}$。
   - 我们的目标是求 $\mathbb{E}[x]$，即对应 η_1 的分量。
2. **对数配分函数**:

$$A(\eta) = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log(-\frac{\pi}{\eta_2}) \tag{27}$$

求导:
我们对 η_1 求偏导：

$$\frac{\partial A(\eta)}{\partial \eta_1} = \frac{\partial}{\partial \eta_1}\left(-\frac{\eta_1^2}{4\eta_2}\right) \tag{28}$$

(第二项 $\frac{1}{2}\log(-\frac{\pi}{\eta_2})$ 与 η_1 无关，导数为 0)

$$\frac{\partial A(\eta)}{\partial \eta_1} = -\frac{1}{4\eta_2} \cdot 2\eta_1 = -\frac{\eta_1}{2\eta_2} \tag{29}$$

代入参数:
回顾参数映射关系：

$$\eta_1 = \frac{\mu}{\sigma^2}, \quad \eta_2 = -\frac{1}{2\sigma^2} \tag{30}$$

代入求导结果：

$$A'(\eta_1) = -\frac{\frac{\mu}{\sigma^2}}{2 \cdot (-\frac{1}{2\sigma^2})} = -\frac{\frac{\mu}{\sigma^2}}{-\frac{1}{\sigma^2}} = \mu \tag{31}$$

结论:

$$\mathbb{E}[x] = \mu \tag{32}$$

这与我们熟知的高斯分布均值完全一致！通过这种方式，我们不需要进行复杂的积分运算 ($\int x \cdot \frac{1}{\sqrt{2\pi}\sigma} e^{-\cdots} dx$)，只需要简单的代数求导即可得到期望。同样地，通过对 η_2 求导或求二阶导数，我们可以推导出 $\mathbb{E}[x^2]$ 和方差。

4. 最大似然估计与充分统计量 (Maximum Likelihood Estimation & Sufficient Statistics)

本节我们将证明一个非常优美的结论：**对于指数族分布，最大似然估计 (MLE) 等价于让模型产生的充分统计量的期望等于经验分布的充分统计量的均值。**

4.1 问题设定

假设我们有 N 个独立同分布 (i.i.d.) 的样本 $D = \{x_1, x_2, \dots, x_N\}$。

我们的目标是找到最优参数 η_{MLE}，使得对数似然函数最大化：

$$\eta_{MLE} = \underset{\eta}{\operatorname{argmax}} \log P(D|\eta) \tag{33}$$

4.2 推导过程 (Derivation)

$$\begin{aligned}
\log P(D|\eta) &= \log \prod_{i=1}^{N} P(x_i|\eta) \\
&= \sum_{i=1}^{N} \log P(x_i|\eta) \\
&= \sum_{i=1}^{N} \log \left(h(x_i) \exp(\eta^T \phi(x_i) - A(\eta)) \right) \\
&= \sum_{i=1}^{N} \left(\log h(x_i) + \eta^T \phi(x_i) - A(\eta) \right) \\
&= \sum_{i=1}^{N} \log h(x_i) + \sum_{i=1}^{N} \eta^T \phi(x_i) - NA(\eta)
\end{aligned} \tag{34}$$

为了求最大值，我们需要对 η 求导并令其为 0。注意第一项 $\sum \log h(x_i)$ 与 η 无关，导数为 0。

$$\begin{aligned}
\frac{\partial}{\partial \eta} \log P(D|\eta) &= \frac{\partial}{\partial \eta} \left(\sum_{i=1}^{N} \eta^T \phi(x_i) - NA(\eta) \right) \\
&= \sum_{i=1}^{N} \phi(x_i) - N \frac{\partial A(\eta)}{\partial \eta} \\
&= \sum_{i=1}^{N} \phi(x_i) - NA'(\eta)
\end{aligned} \tag{35}$$

令导数为 0：

$$\sum_{i=1}^{N} \phi(x_i) - NA'(\eta_{MLE}) = 0 \tag{36}$$

整理得：

$$A'(\eta_{MLE}) = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i) \tag{37}$$

4.3 结论与几何解释 (Conclusion)

回顾第 3 节的性质 $A'(\eta) = \mathbb{E}_{P(x|\eta)}[\phi(x)]$，我们可以将上述结果重写为：

$$\mathbb{E}_{P(x|\eta_{MLE})}[\phi(x)] = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i) \tag{38}$$

- **左边**: 模型在参数 η_{MLE} 下预测的充分统计量的期望 (Theoretical Moment)。
- **右边**: 观测数据的充分统计量的样本均值 (Empirical Moment)。

这意味着：最大似然估计的过程，本质上就是调整参数 η，使得“模型眼中的世界”（理论期望）与“我们看到的世界”（经验均值）在充分统计量上完全一致。

5. 最大熵角度 (Maximum Entropy View)

最大熵原理认为，在满足所有已知约束条件的情况下，我们应该选择**熵 (Entropy)** 最大的分布作为我们的模型。这也等价于认为该分布是“最无偏的”或“最随机的”。

5.1 定义 (Definition)

- **信息量 (Information)**: $-\log p(x)$
- **熵 (Entropy)**: 信息量的期望。

$$H[P] = \mathbb{E}[-\log p(x)] = -\sum_{x} p(x)\log p(x) \quad \text{(离散情形)} \tag{39}$$

$$H[P] = -\int p(x)\log p(x)dx \quad \text{(连续情形)} \tag{40}$$

5.2 没有任何约束时的最大熵分布 (MaxEnt with No Constraints)

假设 x 是离散随机变量，取值范围为 $1, 2, \ldots, K$。
已知约束仅为概率归一化条件：

$$\sum_{i=1}^{K} p_i = 1 \tag{41}$$

我们要最大化熵：

$$\begin{aligned} \hat{P} &= \operatorname*{argmax}_{P} H[P] \\ &= \operatorname*{argmax}_{P} \left(-\sum_{i=1}^{K} p_i \log p_i \right) \end{aligned} \tag{42}$$

这等价于最小化负熵：

$$\min_{P} \sum_{i=1}^{K} p_i \log p_i \quad \text{s.t.} \quad \sum_{i=1}^{K} p_i = 1 \tag{43}$$

拉格朗日乘子法 (Lagrange Multipliers):

构造拉格朗日函数：

$$L(p,\lambda)=\sum_{i=1}^{K}p_i\log p_i+\lambda\left(1-\sum_{i=1}^{K}p_i\right) \tag{44}$$

对 p_i 求导并令其为 0：

$$\frac{\partial L}{\partial p_i}=\log p_i+p_i\cdot\frac{1}{p_i}-\lambda=\log p_i+1-\lambda=0 \tag{45}$$

求解 p_i：

$$\log p_i=\lambda-1\implies p_i=\exp(\lambda-1) \tag{46}$$

由于 λ 是常数，这意味着所有的 p_i 都相等（Constant）。
结合归一化条件 $\sum p_i=1$：

$$\sum_{i=1}^{K}p_i=\sum_{i=1}^{K}c=K\cdot c=1\implies c=\frac{1}{K} \tag{47}$$

结论:

$$p_1=p_2=\ldots=p_K=\frac{1}{K} \tag{48}$$

5.3 有约束时的最大熵分布 (MaxEnt with Constraints)

现在我们要处理更一般的情况：除了归一化条件外，我们还已知某些特征函数 $f(x)$ 的期望值。

问题设定:

- **已知数据**: $Data=\{x_1,x_2,\ldots,x_N\}$
- **经验分布**: $\hat{P}(x)=\frac{\text{count}(x)}{N}$
- **特征函数**: $f(x)$ 是任意关于 x 的函数向量。
- **约束条件**: 我们要求模型分布 $P(x)$ 对特征 $f(x)$ 的期望等于经验分布的期望（即已知事实）。

$$\mathbb{E}_P[f(x)]=\mathbb{E}_{\hat{P}}[f(x)]=\Delta\quad(\text{已知常数向量}) \tag{49}$$

优化目标:

$$\begin{cases}\min_P & \sum_x p(x)\log p(x)\\ \text{s.t.} & \sum_x p(x)=1\\ & \sum_x p(x)f(x)=\Delta\end{cases} \tag{50}$$

拉格朗日乘子法:
引入拉格朗日乘子 λ_0 (对应归一化约束) 和向量 λ (对应期望约束)。

$$L(p,\lambda_0,\lambda)=\sum_x p(x)\log p(x)+\lambda_0\left(1-\sum_x p(x)\right)+\lambda^T\left(\Delta-\sum_x p(x)f(x)\right) \tag{51}$$

求解:
对 $p(x)$ 求偏导并令其为 0：

$$\begin{aligned}\frac{\partial L}{\partial p(x)}&=(\log p(x)+1)-\lambda_0-\lambda^T f(x)=0\\ \implies\log p(x)&=\lambda^T f(x)+\lambda_0-1\\ \implies p(x)&=\exp\left\{\lambda^T f(x)+(\lambda_0-1)\right\}\end{aligned} \tag{52}$$

结果分析:
我们将结果整理一下：

$$p(x) = \exp\left\{ \underbrace{\lambda^T}_{\eta^T} \underbrace{f(x)}_{\phi(x)} - \underbrace{(-(\lambda_0 - 1))}_{A(\eta)} \right\} \tag{53}$$

这是一个标准的**指数族分布**形式！

- **自然参数**: $\eta = \lambda$ (拉格朗日乘子)
- **充分统计量**: $\phi(x) = f(x)$ (特征函数)
- **对数配分函数**: $A(\eta) = 1 - \lambda_0$ (由归一化条件决定)

结论:
最大熵原理告诉我们，在满足已知事实（期望约束）的前提下，对未知分布不做任何额外假设（熵最大），我们必然得到指数族分布。

这也解释了为什么指数族分布在统计建模中如此重要——它是最“客观”、最少偏见的分布选择。