概率图 – 概率论与图论

作者：Calvin
QQ: 179209347
Mail: 179209347@qq.com

介绍

笔记简介：

- 面向对象：深度学习初学者
- 依赖课程：**线性代数，统计概率**，优化理论，图论，离散数学，微积分，信息论

知乎专栏：

https://zhuanlan.zhihu.com/p/693738275

Github & Gitee 地址：

https://github.com/mymagicpower/AIAS/tree/main/deep_learning

https://gitee.com/mymagicpower/AIAS/tree/main/deep_learning

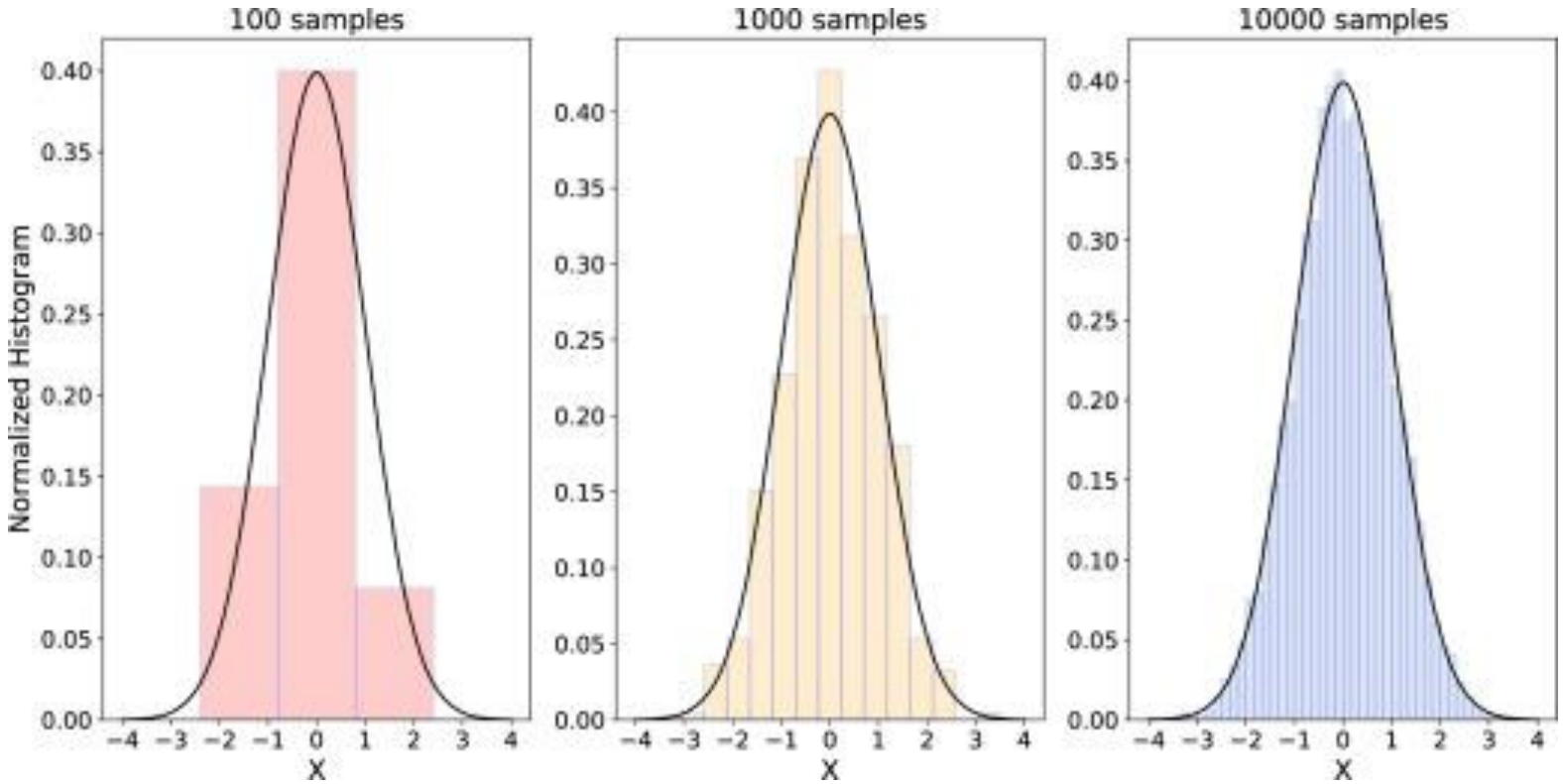
*** 版权声明：**

- 仅限用于个人学习
- 禁止用于任何商业用途

概率论

随机现象：随机现象是指在一定条件下，无法准确预测其具体结果的事件或现象。这些事件的结果可能在一定范围内呈现多样性，并且在每次试验中可能出现不同的结果。

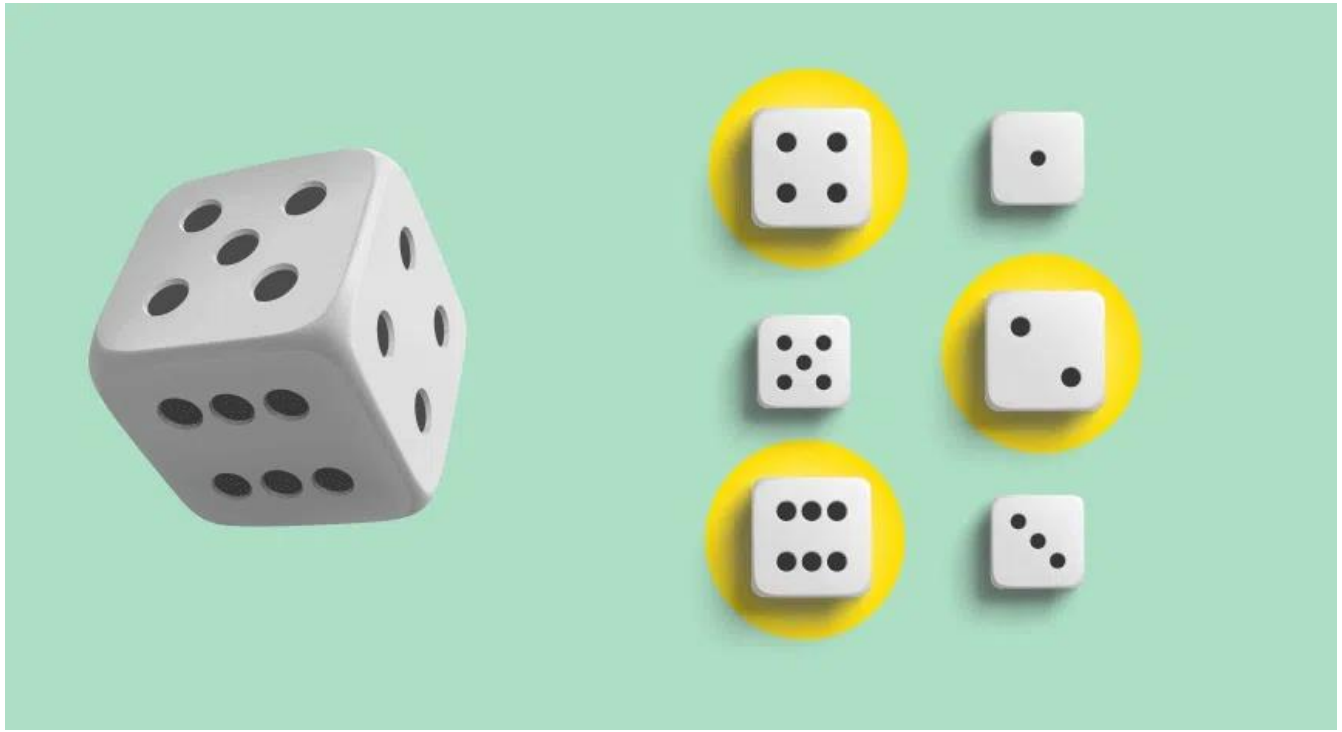
统计规律性：是指在大量的观察或试验中，随机现象的行为趋于呈现某种稳定的模式或趋势。虽然单个事件的结果可能是随机的，但是当这些事件被重复进行很多次，并且数据量足够大时，其中的一些统计特征就会显示出明显的规律性。



概率论 - 样本空间

样本空间（Ω）： 在随机试验中可能出现的所有可能结果的集合。它包含了试验的所有可能的基本结果，每个结果称为一个样本点。样本空间通常用符号 "Ω" 表示。

例如：
对于一次投掷一个六面骰子的试验，样本空间可以定义为 {1, 2, 3, 4, 5, 6}，其中每个数字代表了骰子可能停留的六种不同的结果。



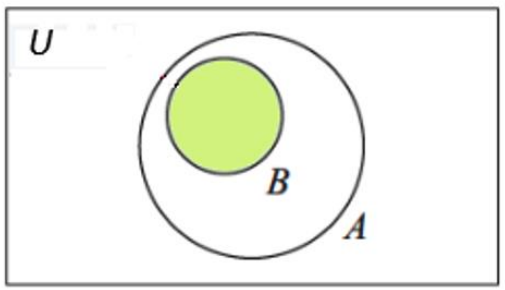
概率论 - 随机事件

随机事件：某些样本点组成的集合，常用 A 、 B 、C...表示。

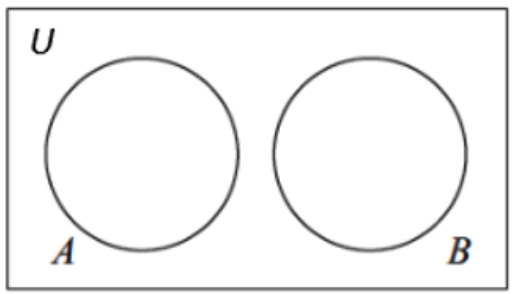
随机变量：随机变量是描述随机事件结果的数学量。它可以是一个数值，也可以是一个函数。随机变量可以是离散的，即取有限或无限个可能值中的一个，也可以是连续的，即可以取任意一个数值。
例如，投掷一枚硬币可以是一个随机事件，因为我们无法确定硬币会正面朝上还是反面朝上。我们可以定义一个随机变量X来表示投掷硬币的结果，其中X=1表示正面，X=0表示反面。
常用大写字母 X、Y、Z ...表示。

事件间的关系：

包含关系： $A \subset B$,一个事件包含于另一个事件，表示一个事件的发生意味着另一个事件一定发生。A 发生必然导致B 发生。
相等关系： $A = B \Leftrightarrow A \subset B$ 且 $B \subset A$ 。
互斥关系： $A \cap B = \emptyset$， A 和 B 不可能同时发生。



B is proper subset of A　　$B \subset A$

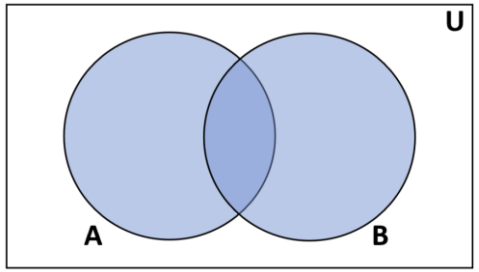


A and B are disjoint sets

概率论 - 事件的运算

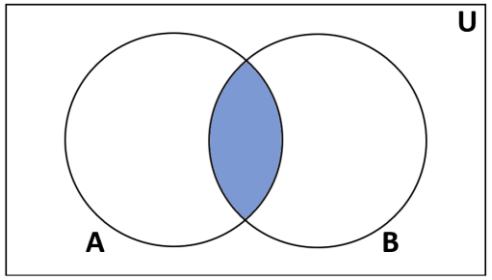
事件的运算:

并（Union）：表示两个或多个事件中至少有一个发生的情况。
•记号：$A \cup B$
•解释：$A \cup B$ 包含了事件 A 或事件 B 或两者同时发生的所有情况。



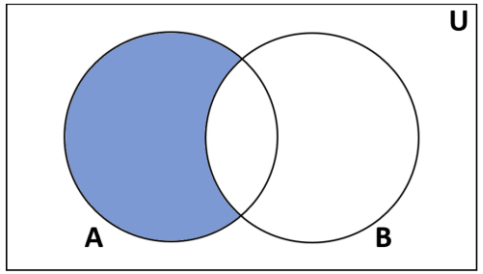
{1, 2, 3} ∪ {3, 4, 5, 6} = {1, 2, 3, 4, 5, 6}

交（Intersection）：表示两个事件同时发生的情况。
•记号：$A \cap B = AB$
•解释：$A \cap B = AB$ 包含了事件 A 和事件 B 同时发生的所有情况。



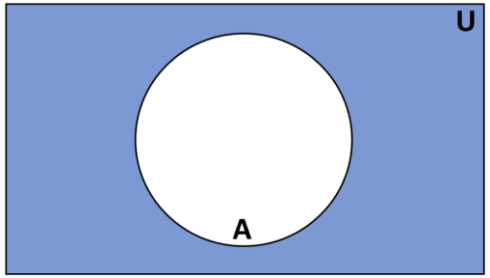
{1, 2, 3} ∩ {3, 4, 5, 6} = { 3 }

差（Difference）：表示一个事件发生而另一个事件不发生的情况。
•记号：A−B 或 A\B
•解释：A−B 包含了事件 A 发生而事件 B 不发生的所有情况。



{1, 2, 3} – {3, 4, 5, 6} = {1, 2 }

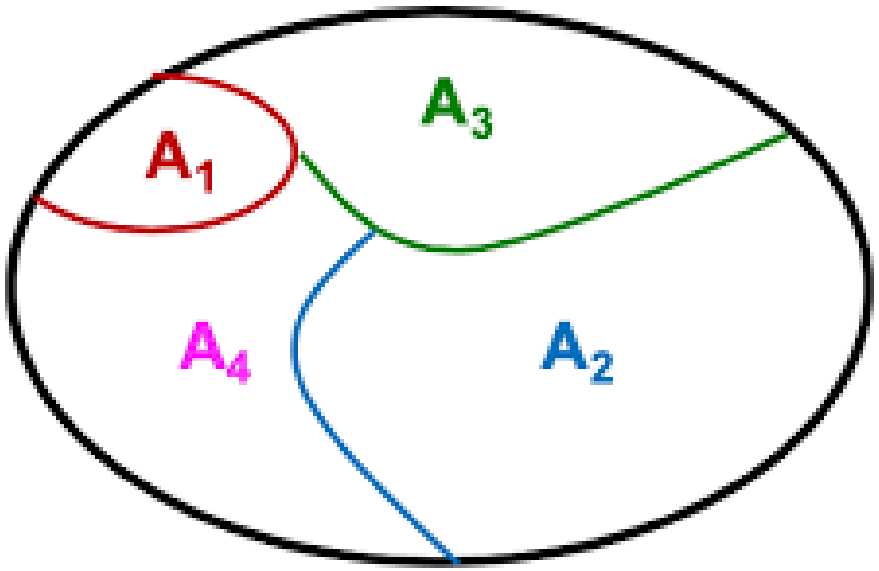
补（Complement）：表示一个事件不发生的情况。
•记号：A′ 或 $\bar{A}$
•解释：$\bar{A}$ 包含了事件 A 不发生的所有情况。



概率论 - 样本空间的分割

样本空间的分割:

若 $A_1, A_2, \ldots, A_n$ 满足
1. A_i 间互不相容;
2. $A_1 \cup A_2 \cup \ldots\ldots \cup A_n = \Omega$
则称 $A_1, A_2, \ldots, A_n$ 为 Ω 的一组分割。



概率的定义

概率的直观定义:
指某一事件发生的可能性大小。在数学上，概率通常用一个介于0到1之间的数字表示，0表示不可能事件，1表示必然事件，其他数字则表示事件发生的可能性大小。

概率的统计定义:
概率的统计定义是指通过大量实验或观察，对事件发生的频率进行估计，并将这个频率作为概率的近似值。
例如，如果一个事件在大量实验中发生的次数占总实验次数的比例稳定在某个值，那么这个值就是该事件的概率。这种定义适用于频率论的概率观点，其中概率被理解为事件在无限次试验中发生的频率极限。

概率的公理化定义:
非负性公理：$P(A) \geq 0$;
正则性公理： $P(\Omega) = 1$;
可列可加性公理： $A_1, A_2, \ldots\ldots, A_n$ 互不相容, 则

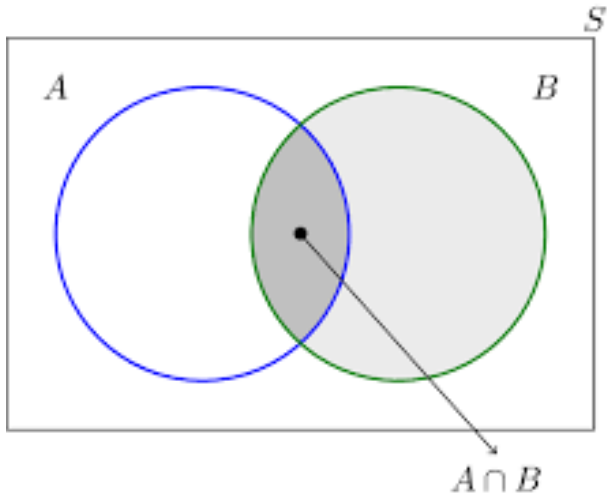
$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

条件概率 - 条件概率的三个重要公式 （1/3）

条件概率定义: 对于事件A、B，若 $P(B)>0$，则称 $P(A|B)=P(AB)/P(B)$ 为在 B 出现的条件下，A 出现的条件概率。

乘法公式:

若 $P(B)>0$, 则 $P(AB)=P(B)P(A|B)$;
若 $P(A)>0$, 则 $P(AB)=P(A)P(B|A)$.
若 $P(A_1A_2\ .....A_n-1)>0$, 则 $P(A_1A_2\ .....A_n)=P(A_1)P(A_2|A_1)\ .....P(A_n|A_1A_2\ .....A_{n-1})$



$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$

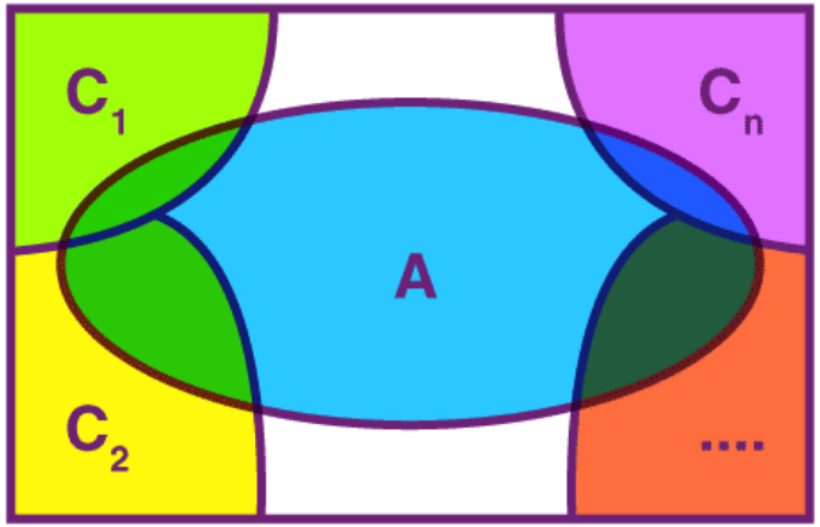
条件概率 - 条件概率的三个重要公式 （2/3）

全概率公式：

若事件 $C_1, C_2, \ldots\ldots, C_n$ 是样本空间 Ω 的一组分割，且 $P(C_i)>0$。

利用全概率公式，可以通过已知条件概率 $P(A \mid C_i)$ 和每个 C_i 的概率 $P(C_i)$ 来计算事件 A 的概率 $P(A)$ 。

$$P(A) = \sum_{i=1}^{n} P\left(AC_i\right) = \sum_{i=1}^{n} P\left(C_i\right) P(A \mid C_i)$$



条件概率 - 条件概率的三个重要公式 （3/3）

贝叶斯公式:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

- $P(B)$ 为先验概率
- $P(B|A)$ 为后验概率

其中，$P(B|A)$ 表示在事件 A 发生的条件下事件 B 发生的概率， $P(A|B)$ 和 $P(B)$ 的含义同上， $P(A)$ 是事件 A 的概率。贝叶斯定理可以用来在已知事件 A 发生的条件下，推断事件 B 发生的概率。

若事件$B_1, B_2, \cdots\cdots, B_n$ 是样本空间 Ω 的一组分割,且$P(A) > 0, P(B_i) > 0$,则:

$$P(B_i \mid A) = \frac{P(AB_i)}{P(A)} = \frac{P(B_i)P(A \mid B_i)}{P(A)}$$
$$= \frac{P(B_i)P(A \mid B_i)}{\sum_{j=1}^{n} P(B_j)P(A \mid B_j)}, i = 1,2,\ldots,n$$

- $P(B_i)$为先验概率
- $P(B_i \mid A)$ 为后验概率

事件的独立性

事件的独立性:
对于两事件，若其中任何一个事件的发生不影响另一个事件的发生， 则这两事件是独立的，即，若事件 A 与 B 满足: $P(AB) = P(A)P(B)$, 则称 A 与 B 相互独立。

$$P(A|B) = P(A)$$
$$P\,(AB)/P\,(B) = P(A)$$
$$P(AB) = P(A)P(B)$$

条件独立性:
在给定 C 的条件下，若事件 A 与 B 满足 $P(AB|C) = P(A|C)P(B|C)$, 则 称 A 与 B 在给定 C 的条件下相互独立。

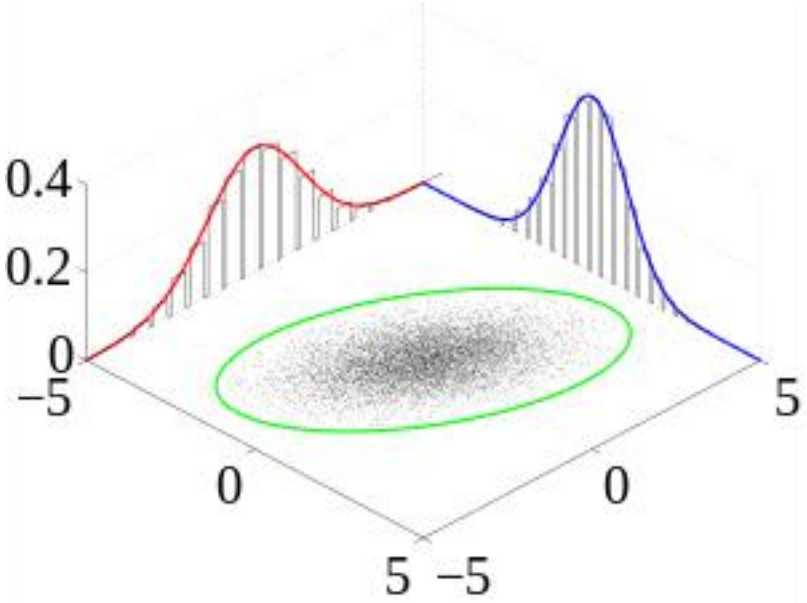
联合概率分布

联合概率分布:
联合概率分布是指在统计学和概率论中描述两个或多个随机变量同时取值的概率情况。

- 联合概率分布可以用概率密度函数（对于连续型变量）或概率质量函数（对于离散型变量）来表示。
- 在实际应用中，联合概率分布可以用来描述多个随机变量之间的关系，比如相关性、独立性等。

$$p(\mathbf{X}) = p(X_1, X_2, \ldots, X_N)$$

如果有两个随机变量 X 和 Y，它们的联合概率分布描述了在给定X和Y的取值情况下，它们同时取到这些值的概率。

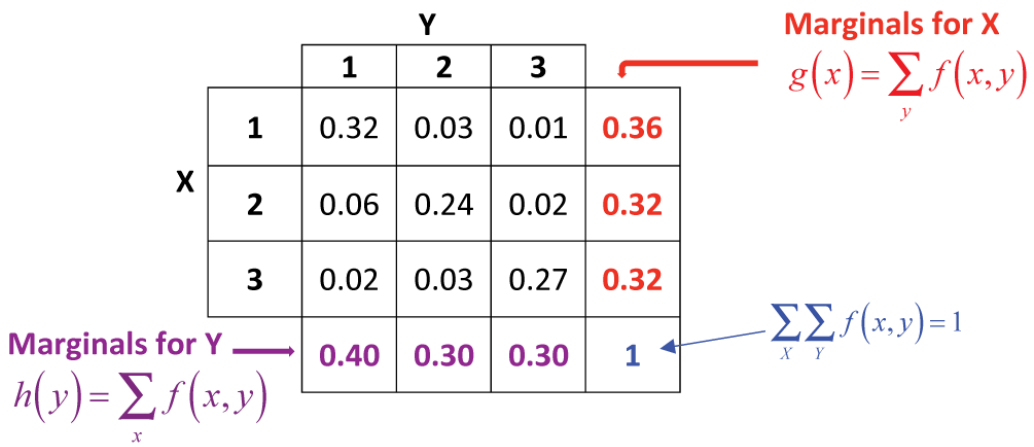


边缘概率

边缘概率:
指的是一个事件在给定一些部分信息的情况下的概率。具体来说，边缘概率是指在考虑某个或某些变量的条件下，对其他变量的概率进行的边缘化或求和。这在多变量统计分析中是非常常见的。

$$p(\mathbf{X}_\alpha) = \sum_{\mathbf{X}\setminus\mathbf{X}_\alpha} p(\mathbf{X})$$

举个例子，假设有两个随机变量X和Y，它们之间存在某种关系。边缘概率就是指在给定X的条件下，求Y的概率分布，或者在给定Y的条件下，求X的概率分布。



$$f_X(x) = \int f_{X,Y}(x, y)dy$$

$$f_Y(y) = \int f_{X,Y}(x, y)dx$$

Calcworkshop.com

似然函数

令X_1、X_2、X_3、…、X_n 为来自具有参数 θ 的分布的随机样本。假设我们已经观察到$X_1 = x_1$、$X_2 = x_2$、⋯、$X_n = x_n$ 。

1.如果X_i是离散的，则似然函数定义为：

$$L(x_1, x_2, \cdots, x_n; \theta) = P_{X_1 X_2 \cdots X_n}(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} P(x_i|\theta).$$

2.如果X_i是联合连续的，则似然函数被定义为：

$$L(x_1, x_2, \cdots, x_n; \theta) = f_{X_1 X_2 \cdots X_n}(x_1, x_2, \cdots, x_n; \theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

在某些问题中，使用下式给出的对数似然函数更容易：

$$\ln L(x_1, x_2, \cdots, x_n; \theta) = \sum_{m} \ln P\,(x|\theta) \qquad \ln L(x_1, x_2, \cdots, x_n; \theta) = \sum_{m} \ln f\,(x|\theta)$$

最大似然估计（MLE）与 最大后验概率（MAP）

Suppose we have data $\mathcal{D} = \{x^{(i)}\}_{i=1}^{N}$

Maximum Likelihood Estimate (MLE)

$$\boldsymbol{\theta}^{\text{MLE}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})$$

Maximum *a posteriori* (MAP) estimate

$$\boldsymbol{\theta}^{\text{MAP}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N} p(\mathbf{x}^{(i)}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Prior

MLE和MAP有什么区别

最大似然估计是最大后验估计的一种特殊情况。具体来说，MLE是当您使用统一先验进行MAP估计时得到的结果。这两种方法都是在我们想要回答以下形式的问题时出现的： “给定一些数据，场景 Y 的概率是多少， X 即 $P(Y|X)$ 。
这种形式的问题通常使用贝叶斯定律来回答。

$$\underbrace{P(Y|X)}_{\text{posterior}} = \frac{\overbrace{P(X|Y)}^{\text{likelihood}}\overbrace{P(Y)}^{\text{prior}}}{\underbrace{P(X)}_{\text{probability of seeing the data}}}.$$

MLE

如果我们做最大似然估计，我们不考虑先验信息。上述等式简化为：

$$P(Y|X) \propto P(X|Y).$$

在这种情况下，我们可以拟合一个统计模型，通过最大化可能性 $P(X|Y)$ 来正确预测后验概率 $P(Y|X)$ 。因此叫，最大似然估计。

MAP

如果我们知道关于 Y 的概率，我们可以将其以先验的形式纳入方程中。在这种情况下:

$$P(Y|X) \propto P(X|Y)P(Y)$$

然后，我们通过考虑似然概率和我们对 Y 的先验信念来找到后验概率。因此叫做， “最大后验概率” 。

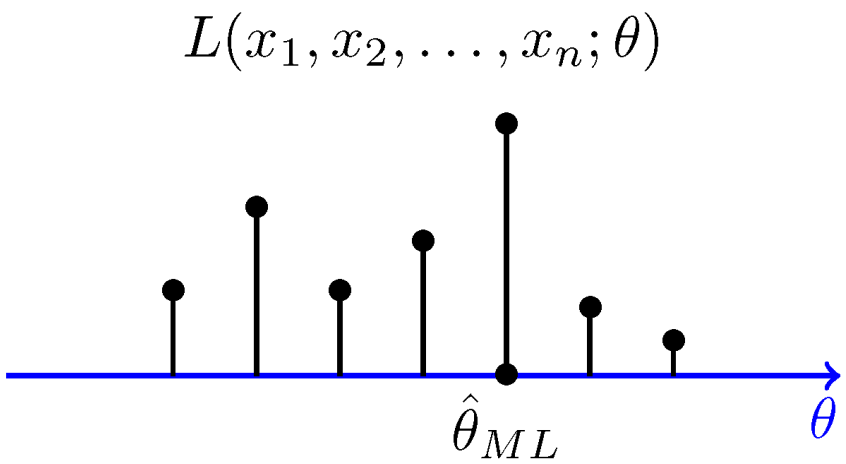
最大似然估计（Maximum Likelihood Estimation，MLE）

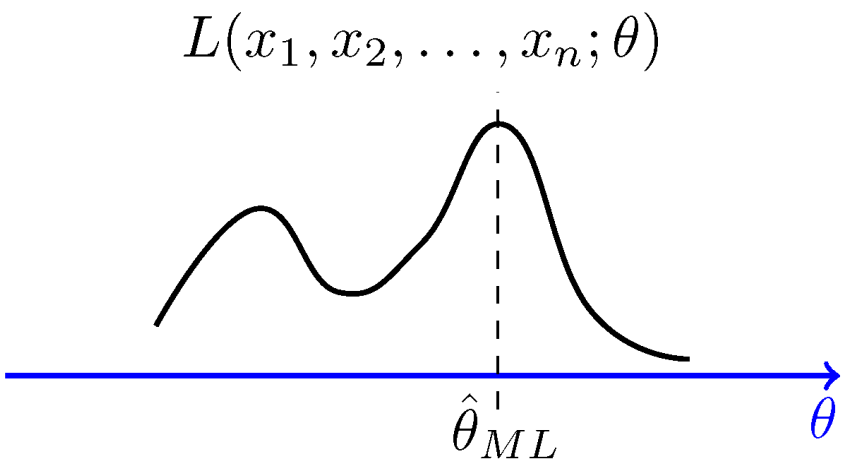
最大似然估计举例说明:

令X_1、X_2、X_3、…、X_n 为来自具有参数θ的分布的随机样本。假设我们已经观察到$X_1 = x_1$ 、$X_2 = x_2$ 、··、$X_n = x_n$ 。 $\hat{\theta}_{ML}$ 表示 θ 的最大似然估计，其为使似然函数最大化的值

$$L(x_1, x_2, \cdots, x_n; \theta).$$

下图说明了如何找到最大似然估计值作为似然函数的 θ 的最大值。图中显示了两种情况：
在第一个图中， θ 是一个离散值参数。在第二个图中， θ 是一个连续值参数。在这两种情况下， θ 的最大似然估计是使似然函数最大化的值。





最大似然估计（MLE）

最大似然估计（Maximum Likelihood Estimation，MLE）是统计学中一种常用的参数估计方法。它的核心思想是：在给定观测数据的情况下，通过调整模型参数的取值，使得这些数据出现的概率最大化。换句话说，MLE试图找到能够最好地解释已知数据的参数值：

$$\theta_{MLE} = arg\,max_{\theta} P(X|\theta)$$
$$= arg\,max_{\theta} \prod_{i} P(x_i|\theta)$$

$$\theta_{MLE} = arg\,max_{\theta} f(X|\theta)$$
$$= arg\,max_{\theta} \prod_{i} f(x_i|\theta)$$

由于取一些小于1的数的乘积会随着这些数的数量趋于无穷而接近0，因此计算是不实际的。因此，我们将在对数空间中工作，因为对数是单调增加的，所以最大化函数等于最大化该函数的对数。

$$\theta_{MLE} = \arg max_{\theta} \ln P(X|\theta)$$
$$= arg\,max_{\theta} \ln \prod_{i} P(x_i|\theta)$$
$$= \arg max_{\theta} \sum_{i} \ln P(x_i|\theta)$$

$$\theta_{MLE} = \arg max_{\theta} \ln f(X|\theta)$$
$$= arg\,max_{\theta} \ln \prod_{i} f(x_i|\theta)$$
$$= \arg max_{\theta} \sum_{i} \ln f(x_i|\theta)$$

为了使用这个框架，我们只需要导出模型的对数似然，然后使用我们最喜欢的优化算法（例如**梯度下降**）将其关于θ最大化。

最大似然估计（MLE） – 正态分布例子

假设我们已经观察到随机样本X_1，X_2，X_3,…, X_n，其中$X_i \sim N(\theta_1, \theta_2)$，所以有：

$$f_{X_i}(x_i; \theta_1, \theta_2) = \frac{1}{\sqrt{2\pi\theta_2}} e^{-\frac{(x_i-\theta_1)^2}{2\theta_2}}.$$

求θ_1和θ_2的最大似然估计。

似然函数：

$$L(x_1, x_2, \cdots, x_n; \theta_1, \theta_2) = \frac{1}{(2\pi)^{\frac{n}{2}}\theta_2^{\frac{n}{2}}} \exp\left(-\frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)^2\right).$$

似然函数对数形式：

$$\ln L(x_1, x_2, \cdots, x_n; \theta_1, \theta_2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\theta_2 - \frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i-\theta_1)^2.$$

最大似然估计（MLE） – 正态分布例子

似然函数对数形式：

$$\ln L(x_1, x_2, \cdots, x_n; \theta_1, \theta_2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\theta_2 - \frac{1}{2\theta_2}\sum_{i=1}^{n}(x_i - \theta_1)^2.$$

对 θ_1 和 θ_2 求导,并将它们设为零:

$$\frac{\partial}{\partial\theta_1}\ln L(x_1, x_2, \cdots, x_n; \theta_1, \theta_2) = \frac{1}{\theta_2}\sum_{i=1}^{n}(x_i - \theta_1) = 0$$

$$\frac{\partial}{\partial\theta_2}\ln L(x_1, x_2, \cdots, x_n; \theta_1, \theta_2) = -\frac{n}{2\theta_2} + \frac{1}{2\theta_2^2}\sum_{i=1}^{n}(x_i - \theta_1)^2 = 0.$$

求解上述方程，得到 θ_1 和 θ_2 的最大似然估计：

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\theta}_1)^2.$$

伯努利（ Bernoulli ）最大似然估计（MLE）

P 为伯努利（ Bernoulli ） 分布：

$$P(X=1)=\theta, P(X=0)=1-\theta$$

独立同分布的采样：

$$D=\{x[1], x[2], \ldots, x[M]\}$$

$$P(x[m]|\theta)=\begin{cases}\theta & x[m]=x^1\\ 1-\theta & x[m]=x^0\end{cases}$$

伯努利（Bernoulli）最大似然估计（MLE）

似然函数：

$$L(\theta: D) = P(D|\theta) = \prod_{m=1}^{M} P(x[m]|\theta)$$

投掷硬币为例子，样本数据，M_1 正面朝上，M_2 反面朝上，似然函数为：

$$L(\theta: M_1, M_2) = \theta^{M_1}(1-\theta)^{M_2}$$

最大化似然函数（对数形式）：

$$l(\theta: M_1, M_2) = M_1 \log\theta + M_2 \log(1-\theta)$$

得到 θ 的最大似然估计：

$$\hat{\theta} = \frac{M_1}{M_1 + M_2}$$

贝叶斯公式

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

先验概率（Prior Probability）：$P(\theta)$
它代表了我们相信的参数值分布。

后验概率（Posterior Probability）：$P(\theta|X)$
它代表利用观察数据计算了等式右边之后的参数值分布。

似然概率（Likelihood Probability）：$P(X|\theta)$
似然概率是指在给定某个假设（或参数）成立的情况下，观察到某个特定数据的概率。它反映了数据在假设条件下的“合理性”。

为什么忽略了 $P(X)$？

$P(X)$ 是一个归一化常数，它确保了计算得到的后验分布的总和等于 1。
在某些情况下，我们并不关心归一化，因此可以将贝叶斯定理写成：

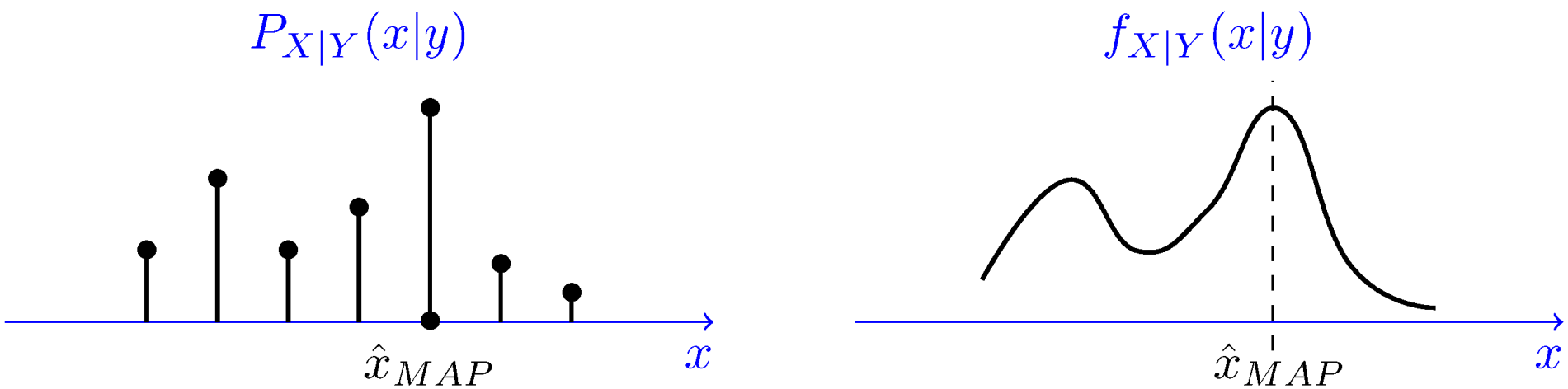
$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

$$\propto P(X|\theta)P(\theta)$$ 我们忽略了归一化常数 $P(X)$

最大后验概率（MAP）

最大后验概率（Maximum a posterior，MAP):
最大后验概率（MAP）是指在贝叶斯统计学中，给定观察到的数据，确定使后验概率最大化的状态。换句话说，它是在考虑了观察数据的情况下，所能获得的最有可能的状态。在贝叶斯推断中，后验概率是基于先验概率和观察数据计算得到的，表示在观察到数据后对参数或未知变量的信念程度。

$$\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathcal{X}} p(\mathbf{X})$$



最大后验概率（MAP）

根据贝叶斯公式，我们可以得到后验概率作为似然概率和先验概率的乘积：

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

$$\propto P(X|\theta)P(\theta) \quad \text{我们忽略了归一化常数 } P(X)$$

我们将MLE公式中的可能性替换为后验，我们可以得到：

$$\begin{aligned}\theta_{MAP} &= \arg\max_{\theta} P(X \mid \theta)P(\theta) \\ &= \arg\max_{\theta} \log P(X \mid \theta) + \log P(\theta) \\ &= \arg\max_{\theta} \log \prod_{i} P(x_i \mid \theta) + \log P(\theta) \\ &= \arg\max_{\theta} \sum_{i} \log P(x_i \mid \theta) + \log P(\theta)\end{aligned}$$

比较MLE方程和MAP方程，唯一不同的是MAP中包含先验 $P(\theta)$ ，否则它们是相同的。这意味着，可能性现在被来自先验的一些权重加权。

最大后验概率（MAP）

如果在MAP估计中使用最简单的先验，比如均匀先验。这意味着可以在所有可能的θ值上分配相等的权重。这意味着可能性由一些常数等效加权。由于是常数，它可以从MAP方程中忽略，因为它不会对最大值有贡献。
假设可以为 θ 分配六个可能的值。现在，我验 $P(\theta)$ 在分布中处处都是 1/6。因此，可以在MAP估计中忽略该常数。

$$\theta_{MAP} = \arg\max_{\theta} \sum_{i} log\ P\ (x_i|\theta) + log\ P\ (\theta)$$

$$= \arg\max_{\theta} \sum_{i} log\ P\ (x_i|\theta) + const$$

$$= \arg\max_{\theta} \sum_{i} log\ P\ (x_i|\theta)$$

$$= \theta_{MLE}$$

如果选择一个先验不是均匀的，比如高斯的，那么先验不再是一个常数了。概率可能很高或很低，但不会相同，因为这取决于分布的区域。
因此，可以清楚地得出结论：当先验是一致的时，MLE是MAP的特殊情况。

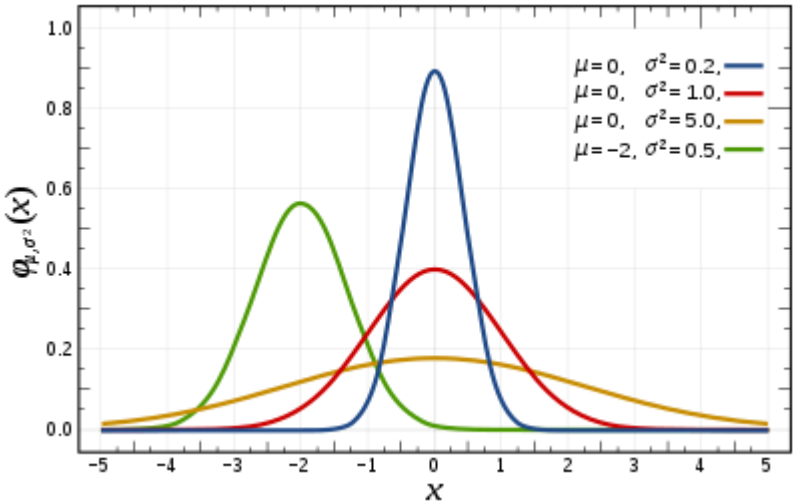
正态分布 – 一元正态分布

正态分布或**高斯分布**是实值随机变量的一种连续概率分布。其概率密度函数的一般形式为：

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

参数 μ 是分布的均值或期望，参数 σ 是标准差。方差是 σ^2 。

符号：$X \sim \mathcal{N}(\mu, \sigma^2)$



标准正态分布

最简单的正态分布称为标准正态分布或单位正态分布。这是特殊情况 μ=0 和 σ=1，它由概率密度函数（或密度） 描述：

$$\varphi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}$$

符号：$X \sim \mathcal{N}(0,1)$

正态分布 – 独立多元正态分布

假设n个变量互不相关，服从正态分布（维度不相关多元正态分布），根据联合概率密度公式有：

$$f(x) = p(x_1, x_2 \ldots x_n) = p(x_1)p(x_2) \ldots p(x_n) = \frac{1}{(\sqrt{2\pi})^n \sigma_1 \sigma_2 \cdots \sigma_n} e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2} \cdots - \frac{(x_n-\mu_n)^2}{2\sigma_n^2}}$$

令：$z^2 = \frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} \cdots + \frac{(x_n-\mu_n)^2}{\sigma_n^2}, \sigma_z = \sigma_1 \sigma_2 \cdots \sigma_n$

则有：

$$f(z) = \frac{1}{(\sqrt{2\pi})^n \sigma_z} e^{-\frac{z^2}{2}}$$

转换成矩阵形式：

$$z^2 = z^{\mathrm{T}} z = [x_1 - \mu_1, x_2 - \mu_2, \cdots, x_n - \mu_n] \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix} [x_1 - \mu_1, x_2 - \mu_2, \cdots, x_n - \mu_n]^{\mathrm{T}}$$

正态分布 – 独立多元正态分布

各个维度的变量： $x = [x_1, x_2, \cdots, x_n]^{\mathrm{T}}$
各个维度的均值： $E(x) = \mu_x = [\mu_1, \mu_2, \cdots, \mu_n]^{\mathrm{T}}$,
各个维度的方差： $\sigma(x) = [\sigma_1, \sigma_2, \cdots, \sigma_n]^{\mathrm{T}}$

则有：

$$x - \mu_x = [x_1 - \mu_1, x_2 - \mu_2, \cdots, x_n - \mu_n]^{\mathrm{T}}$$

Σ 代表变量 X 的协方差矩阵， i行j列的元素值表示 x_i 与 x_j 的协方差。

$$\sum = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

由于各变量之间是相互独立的，所以只有对角线上 (i = j) 存在元素，其他都为 0，即 **Σ** 是一个对角阵，且 x_i 与它本身的协方差就等于方差。根据对角矩阵的性质，它的逆矩阵：

$$\Sigma^{-1} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 & \cdots & 0 \\ 0 & \frac{1}{\sigma_2^2} & \cdots & 0 \\ & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & \frac{1}{\sigma_n^2} \end{bmatrix}$$

正态分布 – 独立多元正态分布

由于对角矩阵的行列式等于对角元素的乘积，所以有：

$$\sigma_z = |\sum|^{\frac{1}{2}} = \sigma_1\sigma_2.\ldots\ldots\sigma_n$$

$$z^{\mathrm{T}}z = (x-\mu_x)^{\mathrm{T}}\sum{}^{-1}(x-\mu_x)$$

多元高斯正态分布形式为：

$$f(z) = \mathcal{N}(\mathbf{x}|\mu,\mathbf{\Sigma}) = \frac{1}{(\sqrt{2\pi})^n\sigma_z}e^{-\frac{z^2}{2}} = \frac{1}{(\sqrt{2\pi})^n|\sum|^{\frac{1}{2}}}e^{-\frac{(x-\mu_x)^{\mathrm{T}}(\sum)^{-1}(x-\mu_x)}{2}}$$

其中 μ 是一个 n 维均值向量， $\mathbf{\Sigma}$ 是 n x n 的协方差矩阵，并且 $|\mathbf{\Sigma}|$ 表示 $\mathbf{\Sigma}$ 的行列式。

符号：$X \sim \mathcal{N}(\mu,\mathbf{\Sigma})$

多元标准正态分布：

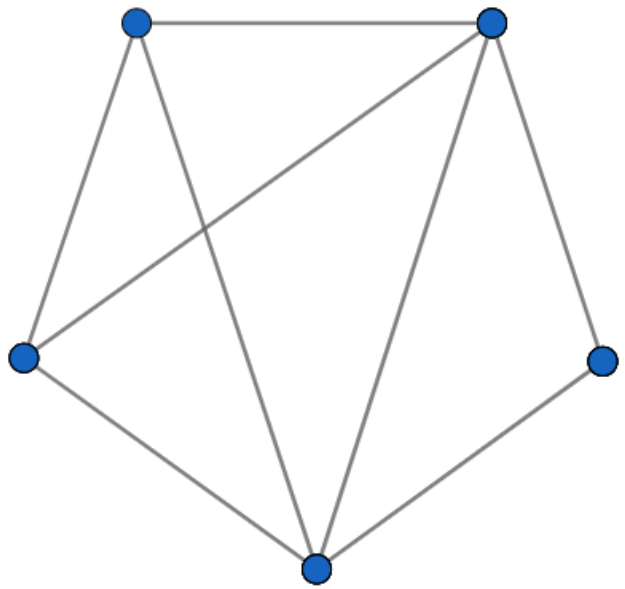
$$f(z) = \mathcal{N}(\mathbf{x}|\mu,\mathbf{\Sigma}) = \frac{1}{(\sqrt{2\pi})^n}e^{-\frac{x^{\mathrm{T}}x}{2}}$$

符号：$X \sim \mathcal{N}(\mathbf{0},\mathbf{I})$

图论

在图论中，图被定义为一个由节点（顶点）和连接这些节点的边（或弧）组成的数学结构。

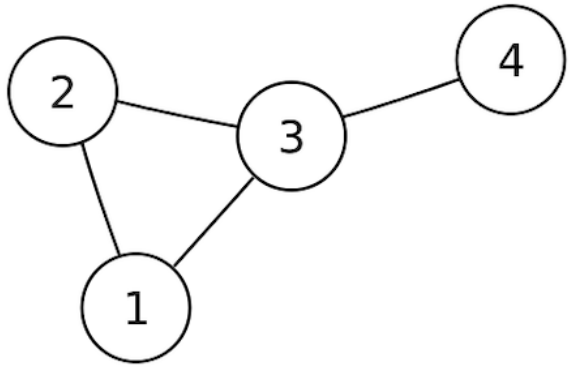
节点（顶点）（Vertices）： 图中的节点是图的基本组成单元。节点可以代表任何实体，如人、地点、物体等。在数学表示中，通常使用字母集合来表示节点的集合。
边（Edges）（或弧）： 边是节点之间的连接线。边可以是有向的（从一个节点指向另一个节点），也可以是无向的（不区分方向）。在有向图中，边通常用箭头表示；在无向图中，边通常用直线表示。边可以带有权重，表示连接的强度或距离。在数学表示中，通常使用字母集合来表示边的集合。



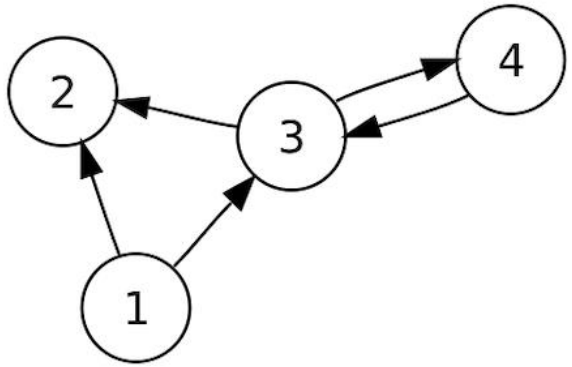
图论 - 图的类型

图的类型： 根据边的性质和节点之间的连接方式，图可以分为多种类型。常见的图类型包括：

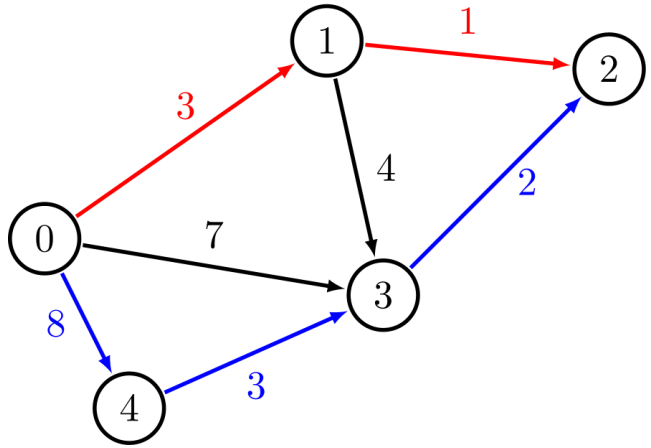
- 无向图（Undirected Graph）：边没有方向，连接两个节点。
- 有向图（Directed Graph）：边有方向，从一个节点指向另一个节点。
- 加权图（Weighted Graph）：边带有权重，表示连接的强度或距离。
- 无权图（Unweighted Graph）：边没有权重。
- 多重图（Multigraph）：允许多条边连接同一对节点。
- 无环图（Acyclic Graph）：没有环的图。
- 简单图（Simple Graph）：没有自环或重复边的图。



无向图（Undirected Graph）



有向图（Directed Graph）



加权图（Weighted Graph）

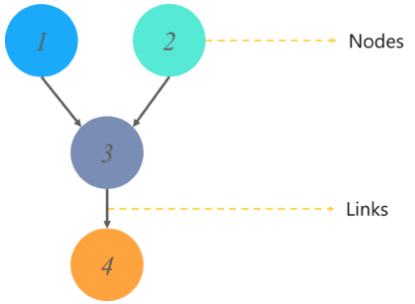
概率图模型

概率图模型（Probabilistic Graphical Models，简称PGM）是一类使用图来表示随机变量及其条件依赖关系的概率模型。它们结合了概率论和图论的优点，使得复杂的概率分布可以通过局部的概率分布来简化描述和计算。

常见的概率图模型

1.贝叶斯网络（Bayesian Networks）

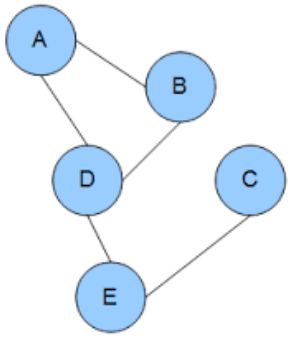
- **结构**: 由有向无环图（DAG）构成，其中节点代表随机变量，边表示条件依赖关系。
- **条件独立性**: 每个节点只依赖其直接父节点，从而简化联合概率分布的计算。
- **应用**: 疾病诊断、语音识别、图像分析等。



贝叶斯网络（Bayesian Networks）

2.马尔可夫随机场（Markov Random Fields，MRF）

- **结构**: 由无向图构成，节点代表随机变量，边表示直接的相互作用。
- **条件独立性**: 每个节点在给定其邻居节点的情况下与其他节点条件独立。
- **应用**: 图像复原、社交网络分析、自然语言处理等。



马尔可夫随机场（Markov Random Fields，MRF）

概率图模型

概率图模型的基本组成部分：

1.节点（Node）

- 每个节点都对应一个随机变量，如：观察变量，隐变量或未知参数等；

2.边（Edge）

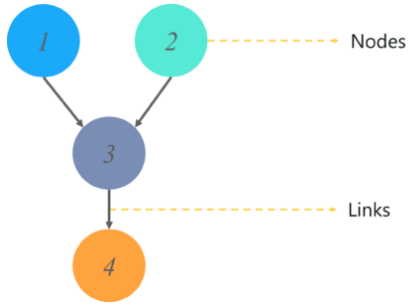
- 代表随机变量之间的依赖关系（有向或无向）。

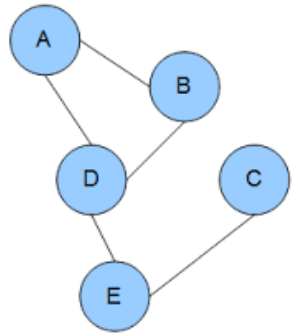
3.条件概率分布（Conditional Probability Distribution，CPD）

- 定义在有向图中的每个节点及其父节点之间的概率关系。

4.潜在函数（Potential Function）

- 定义在无向图中的每个节点及其邻居之间的概率关系。





概率图模型 – 表示（Representation）- 图结构

有向图模型（Directed Graphical Models）

贝叶斯网络（Bayesian Networks）：

- 用有向无环图（DAG）表示随机变量之间的依赖关系。
- 节点表示随机变量，边表示条件依赖关系。
- 每个节点都有一个条件概率分布，条件于其父节点。

例如，一个简单的贝叶斯网络可以表示为：

$$A \rightarrow B \rightarrow C$$

$$P(A,B,C) = P(A)P(B|A)P(C|B)$$

无向图模型（Undirected Graphical Models）

马尔可夫随机场（Markov Random Fields, MRFs）：

- 用无向图表示随机变量之间的相互依赖关系。
- 节点表示随机变量，边表示变量之间的相互影响。
- 用势函数（potential function）表示每个团（clique）的联合概率。

例如，一个简单的MRF可以表示为：

$$\mathrm{A} -\!- \mathrm{B} -\!- \mathrm{C}$$

$$P(A,B,C) = \frac{1}{Z}\phi(A,B)\phi(B,C)$$

概率图模型 – 推理（ Inference ）

推理（ Inference ）：指在已知部分变量的情况下，计算其他变量的概率分布或期望值。

常见的推理方法包括:

1. 精确推理（Exact Inference)

变量消除（Variable Elimination）：

- 通过逐步消除变量来计算目标变量的边缘概率分布。
- 对于贝叶斯网络，可以通过消除非目标变量来获得边缘概率。

信念传播（Belief Propagation, BP）：

- 在树形图中，通过消息传递算法来计算边缘概率分布。
- 针对一般图，可以使用近似信念传播（Loopy Belief Propagation）。

2. 近似推理（Approximate Inference)

蒙特卡罗方法（Monte Carlo Methods）：

- 通过随机采样来近似计算概率分布。
- 常用方法包括马尔可夫链蒙特卡罗（MCMC）、重要性采样（Importance Sampling）。

变分推理（Variational Inference）：

- 将复杂的概率分布近似为更简单的分布，通过优化方法进行推理。
- 常用技术包括变分贝叶斯（Variational Bayesian）、期望最大化（Expectation-Maximization）。

推理问题的分类

推理（Inference）：指在已知部分变量的情况下，计算其他变量的概率分布或期望值。

已知联合概率分布:

$$p(\mathbf{x}) = p(x_1, x_2, \ldots, x_N) = \frac{1}{Z}\prod_c \phi_c(\mathbf{x}_c)$$

三类概率推理

边缘概率	条件概率	最大后验概率
$p(x_A)$	$p(x_A \mid x_B)$	$x_A^* = \arg\max_{x_A} p(x_A \mid x_B)$
marginal distribution	conditional distribution	maximum a posteriori

where x_A , x_B are non-overlapping subset of $x_1, x_2, x_3, \ldots, x_n$

推理问题的分类

Likelihood of evidence: $p(e_1, e_2, e_3, \ldots)$

Posterior belief: $p(z_1, z_2, z_3, \ldots \mid e_1, e_2, e_3, \ldots)$

Prediction: $p(e_1, e_2, e_3, \ldots \mid z_1, z_2, z_3, \ldots)$

Classification: $p(y \mid e_1, e_2, e_3, \ldots)$

Maximum a posterior: $\arg\max_{z} p(z_1, z_2, z_3, \ldots \mid e_1, e_2, e_3, \ldots)$

Missing data: $p(m_1, m_2, m_3, \ldots \mid e_1, e_2, e_3, \ldots)$

MAP: $p(\theta_1, \theta_2, \theta_3, \ldots \mid e_1, e_2, e_3, \ldots)$

MLE: $\arg\max_{\theta} p(e_1, e_2, e_3, \ldots \mid \theta_1, \theta_2, \theta_3, \ldots)$

概率图模型 – 学习（Learning）

学习（Learning）： 学习是指根据数据来估计概率图模型中的参数和结构。学习可以分为参数学习和结构学习。

参数学习（Parameter Learning）

最大似然估计（Maximum Likelihood Estimation, MLE）：

- 给定数据集，寻找使得数据概率最大的参数。
- 对于贝叶斯网络，使用最大似然估计可以得到条件概率分布。

贝叶斯估计（Bayesian Estimation）：

- 使用贝叶斯方法，通过先验分布和数据更新参数的后验分布。

结构学习（Structure Learning）

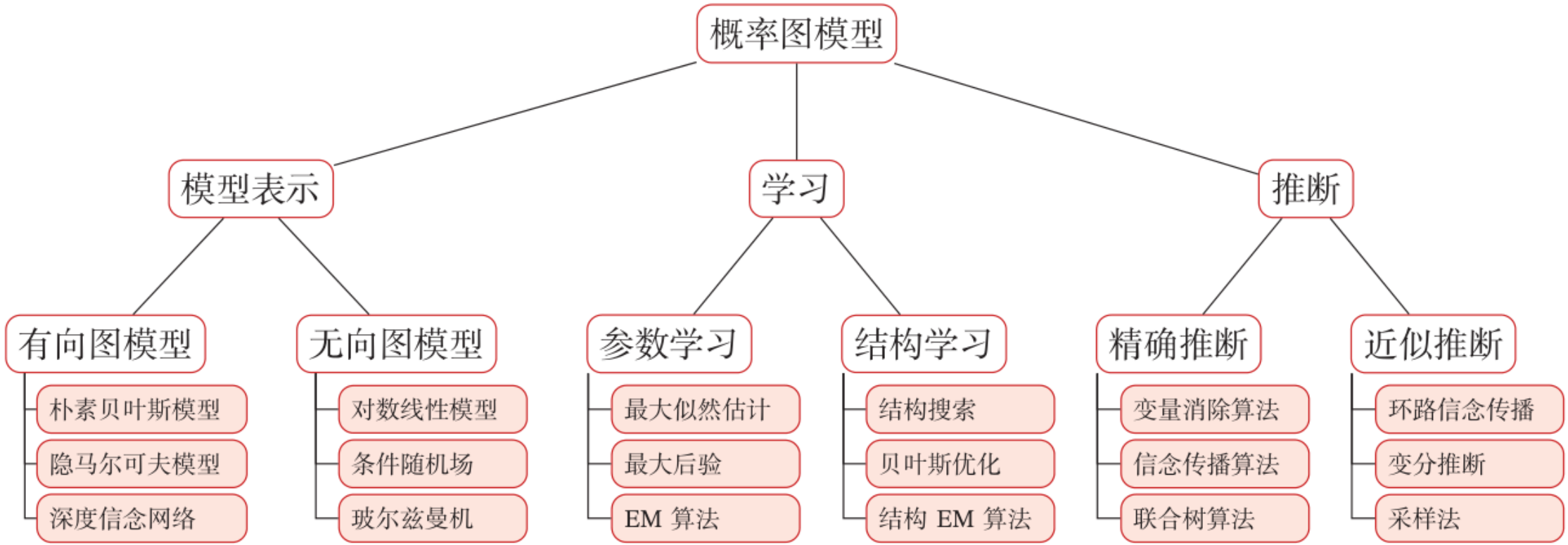
贝叶斯网络的结构学习：

- 寻找最佳的网络结构（DAG），使得数据的概率最大化。
- 常用方法包括评分搜索方法（如BIC、AIC）、约束方法（如独立性检验）。

马尔可夫随机场的结构学习：

- 通过估计图的结构及势函数来表示数据中的依赖关系。
- 常用方法包括最大似然估计和正则化方法。

概率图模型分类



来源：邱锡鹏《神经网络与深度学习》


Thank
You