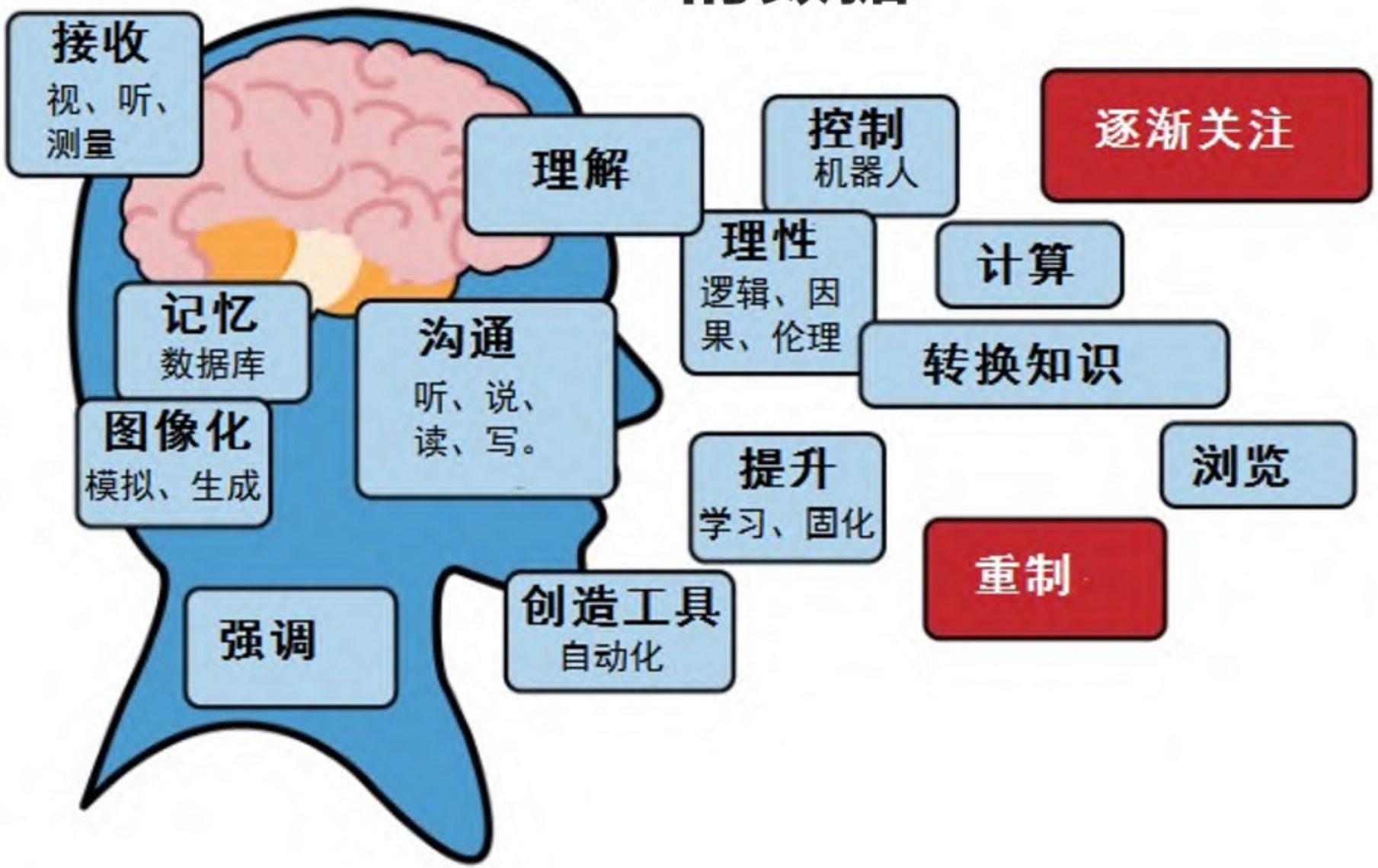
数据、数据、数据

顾立平

01AI的数据


接收
视、听、测量
理解
控制
机器人
逐渐关注
理性
逻辑、因果、伦理
计算
记忆
数据库
沟通
听、说、读、写。
转换知识
图像化
模拟、生成
浏览
提升
学习、固化
重制
强调
创造工具
自动化

01AI的数据

- 存在一个非常简单的数学问题：将给定的数据点集，拟合到适当的函数中（将输入映射到输出）拾取数据中的重要信号并忽略噪声，然后确保此函数在新数据上运行良好。然而，复杂性和挑战来自各种来源：

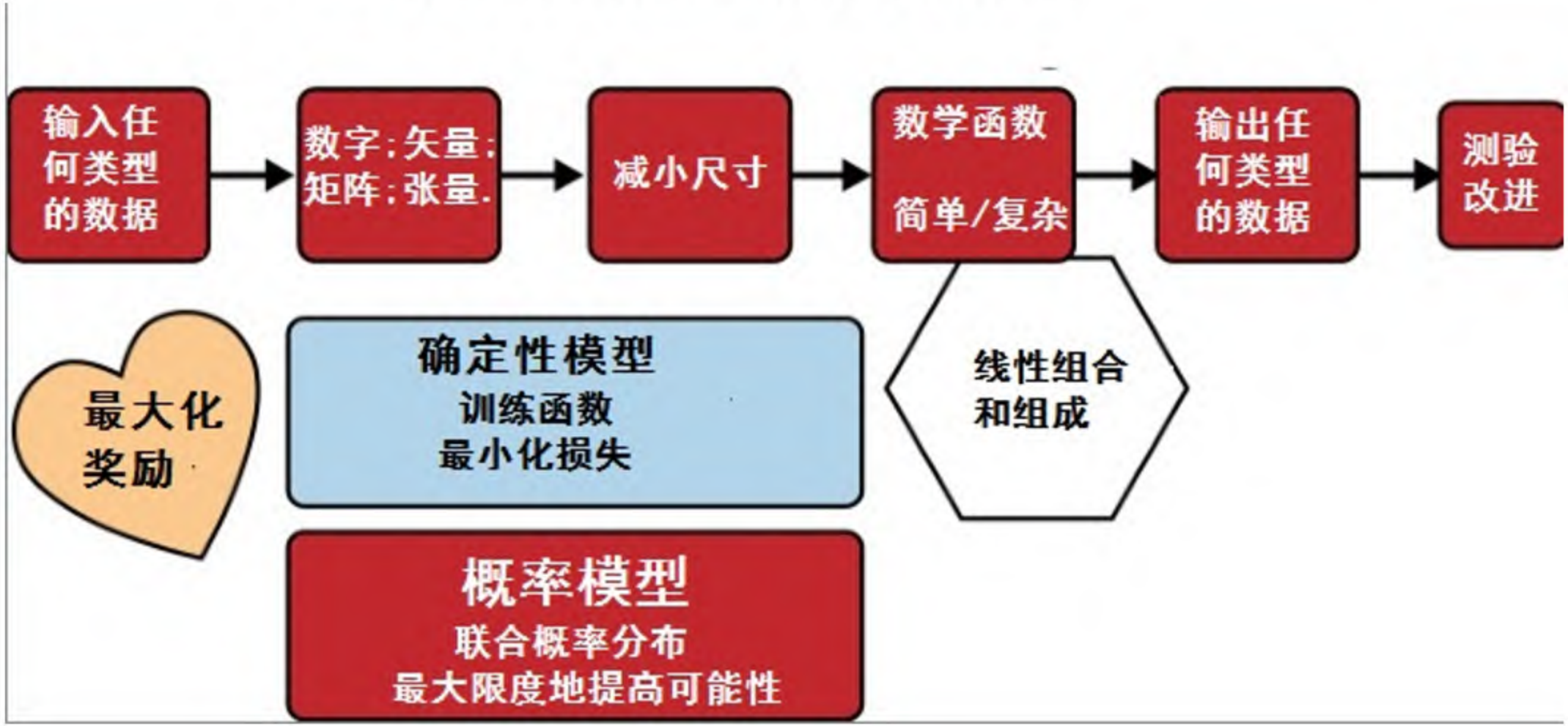
➢ 假设和特征：我们只需观察数据，然后尝试估计生成它的假设函数。我们的函数试图学习那些数据特征，来对于我们的预测、分类、决策或一般目的。人工智能的巨大潜力之一是它的能力：拾取人类没有的能力，搞清楚数据特征之间的微妙交互。

➢ 性能：我们不能轻易衡量我们人工智能系统是否运行良好，并做出正确或适当的预测，以及决策。

➢ 体积：数据的数量实例、观察到的特征和待计算的未知权重可以在数以百万计，所需的计算步骤以十亿计。

➢ 结构：更容易从结构化和标记数据中获得见解，而不是来自非结构化数据的数据。挖掘非结构化数据需要创新技术

02真实数据与模拟数据



02真实数据与模拟数据

- **真实数据**：这些数据是通过使用**测量设备**的真实世界**观测收集**的，传感器、调查、结构化形式，如医学问卷、望远镜、图像设备、网站、股市、受控实验等。这些数据是由于**测量方法的不准确和故障**，通常**不完美和有噪声**和仪器。在数学上，我们不知道确切的函数或概率生成真实数据的分布，但我们可以使用模型、理论和模拟。然后，我们可以测试我们的模型，最后使用他们可以做出预测。

- **模拟数据**：这是使用**已知函数生成的数据**，或从**已知概率分布**。这里，我们有已知的数学函数选项或模型，我们**将数值插入模型**中，以**生成我们的数据点**。例子很多：偏微分的数值解在各种尺度上模拟各种自然现象的方程，例如作为湍流、蛋白质折叠、热扩散、化学反应、行星运动、断裂材料、交通，甚至是增强迪士尼电影动画，例如模拟Moana或Elsa头发中的自然水运动冻结中的移动。

02真实数据与模拟数据

- 真实数据是指通过实际观测收集到的数据，而模拟数据则是通过已知函数或概率分布生成的数据。

- 两种数据类型都对人类的进步和发展有着重要的价值。

- 然而，真实数据往往存在误差和噪声，需要通过模型、理论和模拟来验证其准确性。相比之下，模拟数据则更加可控和精确。

03数学模型：线性与非线性

$$f(x_1,x_2,x_3) = \omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_3$$

- 曲线弯曲为非线性的函数，以及聚集在其周围的数据点弯曲曲线或曲面由非线性函数生成。
- 线性相关性模型描述的是世界的平面度，而非线性函数则可以用来描述更复杂的现实情况。线性函数的公式相对简单，只涉及标量的加减乘除运算；而非线性函数则需要考虑幂、三角函数、指数、对数等微积分函数的嵌入。
- 在数据处理中，我们需要根据不同情况进行选择合适的数学模型，以便更好地理解和预测数据的变化趋势。

03数学模型：线性与非线性

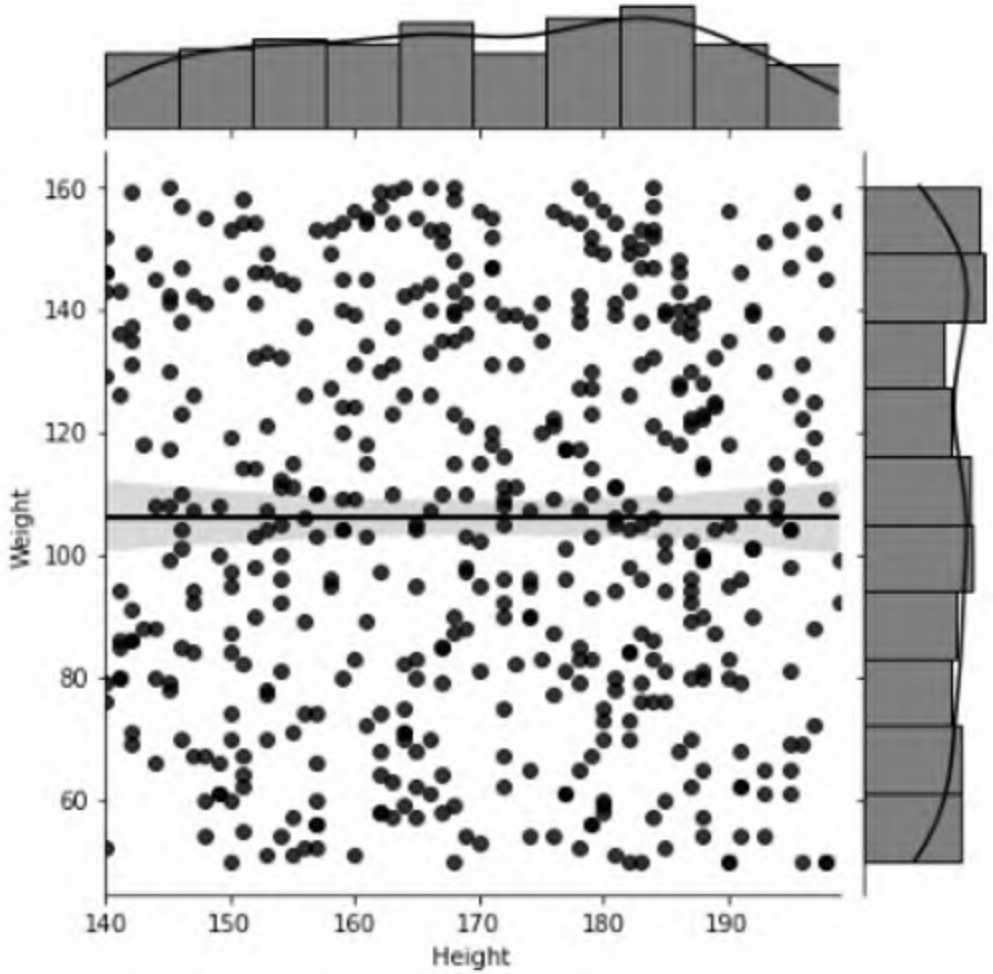
$$f(x_1,x_2,x_3) = \omega_0 + \omega_1\sqrt{x_1} + \omega_2\frac{x_2}{x_3}$$

$$f(x_1,x_2,x_3) = \omega_0 + \omega_1 x_1^2 + \omega_2 x_2^2 + \omega_3 x_3^2$$

$$f(x_1,x_2,x_3) = \omega_1 e^{x_1} + \omega_2 e^{x_2} + \omega_3 \cos(x_3)$$

- 线性相关性模型包括一维直线、二维平面和高维超平面，而线性函数的图形则永远是平坦的。相反，非线性函数的图形则是弯曲的，例如曲线或曲面。
- 线性函数可以用简单的公式来表示，而非线性函数则需要使用幂、乘法或除法等运算符。

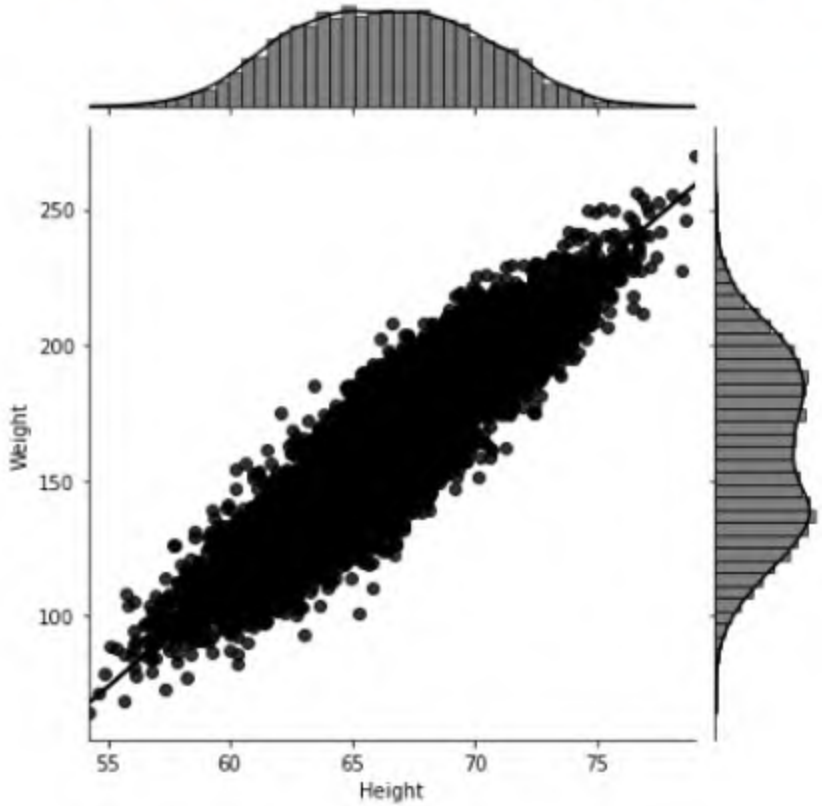
04真实数据示例

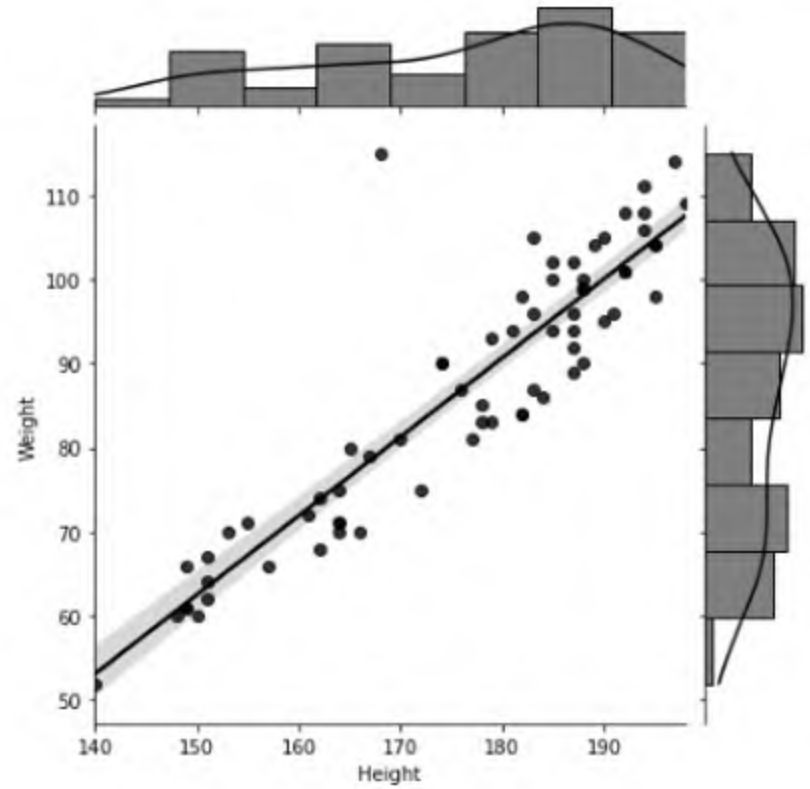


- 在模拟数据的过程中，可以通过设定参数和函数来生成符合特定分布的数据集，从而进行数据分析和建模。
- 这种方式可以帮助我们在缺乏真实数据的情况下，对数据的分布和规律进行研究和预测。

04真实数据示例

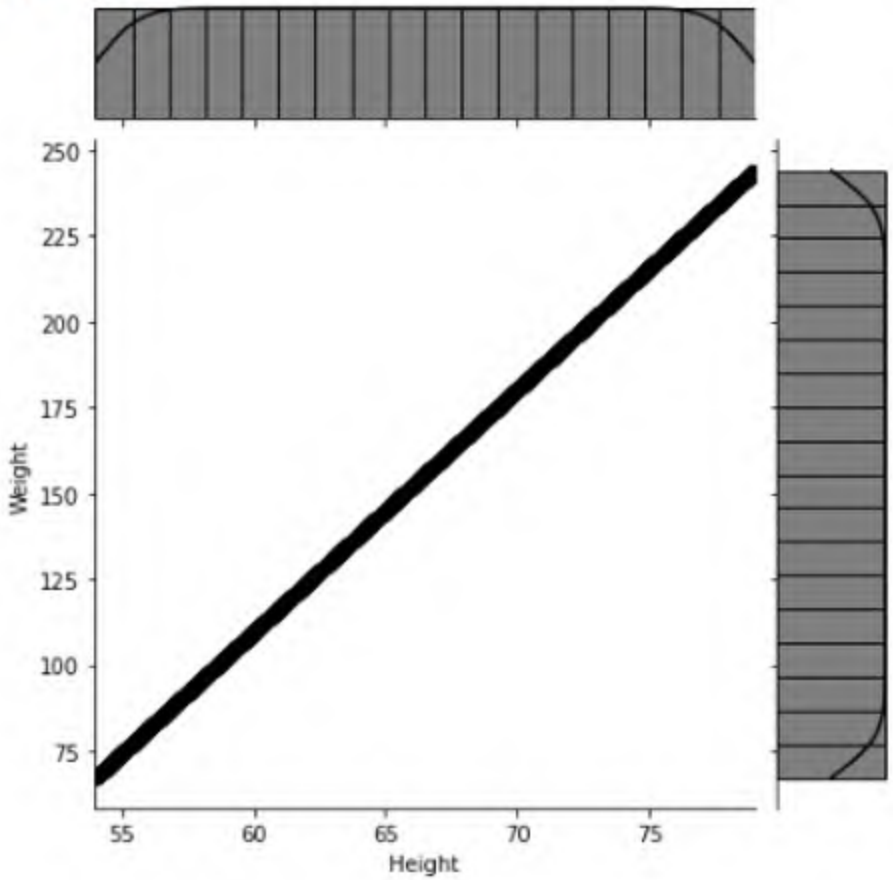
- 两个特征（高度和体重）实现线性相关性。采用钟形正态分布和均匀分布来模拟数据中的随机波动。

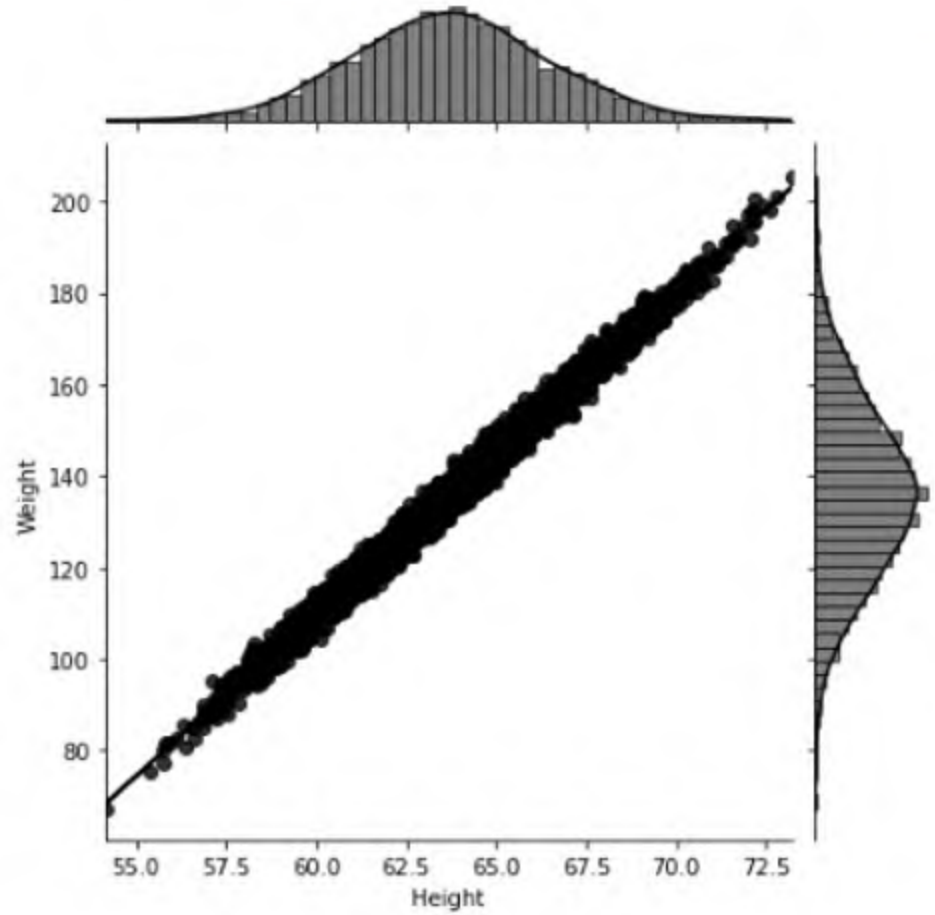




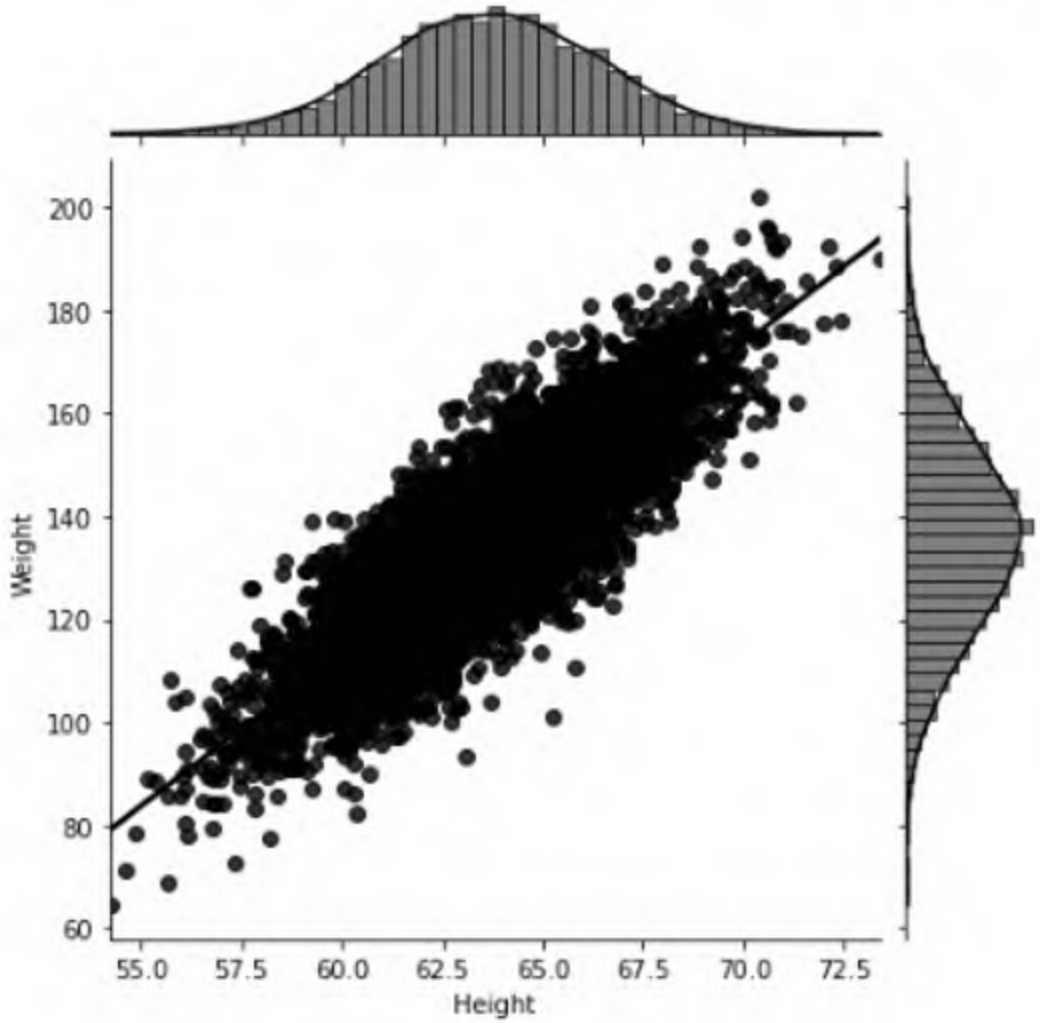
05模拟数据

- 一个是模拟数据的高度-重量模型，另一个是基于真实数据的女性身高-体重数据分布。





05模拟数据



- 在模拟数据时需要考虑概率分布的重要性。
- 我们可以通过调整模型参数来优化模拟结果的观点。

06数学模型：模拟和AI

- 数学模型可以帮助我们更加真实地模拟和理解自然现象，但需要在现实性和简单性之间做出平衡。
- 随着计算能力的提高，我们能够创建和测试更复杂和更真实的数学模型。而人工智能则可以通过学习和泛化来模拟人类智能，并应用于各个领域。
- 此外，人工智能还可以增强现有的数学模型和模拟，使其更加准确和高效。
- 人工智能本身也可以被视为一种数学模型和模拟，其目的是复制人类智能。

07我们从哪里获取数据？

- 不要因为遇到数据所有者的犹豫和抵制而感到惊讶和沮丧，可以通过网络搜索等方式获取公开数据集。
- 在学习数学前要牢记人工智能系统需要数字数据的事实，并且目前全球正朝着数字化的方向发展，科技论文开放获取和科学数据开放共享是主流趋势，加上开源软件运动的扩大，称为：开放科学。

08数据分布、概率和统计词汇

- 当我们进入一个新的领域时，首先需要学习该领域的常用词汇，就像学习一门新的语言一样。通过学习常用的词汇，我们可以更快地适应这个领域，并避免因词汇冲突而产生的困惑。
- 此外，学习词汇还可以帮助我们更好地理解不同领域引用相同概念的不同术语。
- 对于人工智能、机器学习或数据科学等领域，我们需要掌握一些基本的概率和统计学词汇，以便更好地理解和应用相关技术。
- 在阅读这些领域的论文时，不必担心尚未定义的术语，只需先熟悉已知的词汇即可。

09随机变量概率分布

- 随机变量：是指结果不确定、不可预测或随机的变量，而确定性函数则具有确定或确定性的结果。
- 随机变量：可以是从离散集合或连续单元中返回结果，并且可以通过概率分布来描述其可能的结果。
- 概率分布：包括离散随机变量的概率质量函数和连续随机变量的概率密度函数。
- 虽然很少有人知道所有相关随机变量的全联合概率分布，但通过从数据中学习，我们可以更好地理解和处理这些随机变量。

10边际概率

- 边际概率是指在一个联合概率分布中，针对某个随机变量，通过对该变量进行积分或求和得到的概率分布。它是联合概率分布沿着某一维“切掉”之后剩余的概率分布。例如，在二维连续型随机变量 X 和 Y 的联合概率密度函数 f(x,y) 中，X 的边际概率密度函数为：

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy$$

- 即在 y 的取值范围内对 f(x,y) 求积分，得到 X 取某个值的概率分布。同理，Y 的边际概率密度函数为：

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx$$

11均匀分布与正态分布

- 均匀分布是一种概率分布，它的特点是所有可能的取值都有相等的概率被选中。
- 正态分布则是一种连续型概率分布，也被称为高斯分布，它的特点是呈钟形曲线，左右两侧逐渐趋于无穷大。
- 这两种分布都在人工智能应用中有广泛的应用，比如在深度学习中，通常会使用正态分布来初始化神经网络的权重，而在一些决策树算法中，则会使用均匀分布来进行特征的选择。

12条件概率与Bayes定理

- 条件概率是指在一个事件发生的前提下，另一个事件发生的可能性。
- Bayes定理是一种用于计算条件概率的方法，它可以用来更新我们对某个假设的信念，基于新的证据或信息。
- 在机器学习和自然语言处理等领域，条件概率和Bayes定理都有着广泛的应用，比如在文本分类任务中，可以使用朴素贝叶斯算法来预测一个文本属于哪个类别。

13条件概率和联合分布

- 条件概率是指在一个事件发生的前提下，另一个事件发生的可能性。
- 联合分布则是指两个或多个随机变量之间的关系，即它们同时取不同值的概率分布。
- 条件概率和联合分布之间有着密切的关系，根据贝叶斯定理，可以通过已知的联合分布和先验分布来计算出后验分布，从而得到条件概率。
- 在机器学习和自然语言处理等领域，条件概率和联合分布也有着广泛的应用，比如在图像识别任务中，可以使用联合分布来描述像素之间的关系，从而提高识别准确率。

14先验分布、后验分布和似然函数

- 先验分布是指在观察任何数据之前，模型权重的概率分布；
- 后验分布则是指在给定观测数据之后，权重的概率分布。
- 似然函数则是一种编码概率的函数，用于描述观察给定特定权重分布的数据点的真实性。

- 这些概念对于理解概率论的基本原理非常重要，也是很多实际问题求解的基础。

15混合概率分布

- 我们可以混合概率分布，并产生分布的混合。
- 分布混合，是指将多个概率分布进行混合，形成一个新的概率分布的过程。其中最著名的例子就是高斯混合模型，它可以用来建模多峰分布的数据，例如性别分类等。通过将多个高斯分布进行混合，可以更好地拟合数据的分布情况，从而提高模型的准确性。在机器学习和数据挖掘等领域，分布混合物也有着广泛的应用。
- 我们将可能性称为函数，而不是分布。因为概率分布的总和必须为1（或积分如果我们处理的是连续随机变量，则为1），但似然函数不一定等于1（或积分连续随机变量的情况下为1）。

16随机变量的和和积

- 随机变量的和与积，指的是将多个随机变量进行求和或乘积操作得到新的随机变量的过程。这个过程可以帮助我们构建更加复杂的随机事件模型，例如两个骰子的点数之和、两个人的年龄之积等等。
- 在概率论和统计学中，随机变量的和与积也是常用的基本概念之一。

17使用图表示联合概率分布

- 使用图Graph表示联合概率分布是一种有效的分解方式，可以帮助我们更好地理解和分析多个随机变量之间的关系。
- 通过绘制有向图或无向图，我们可以清晰地表达各个随机变量之间的依赖关系，进而计算出它们的联合概率分布。
- 这种方式不仅可以简化计算过程，还可以帮助我们更好地掌握复杂数据集的特点和规律。

18期望值、均值、方差和不确定性

- 期望值和均值用于量化平均值，而方差和标准偏差则用于量化围绕该平均值的扩散，从而编码不确定性。目标是控制方差，以减少不确定性。方差越大，在使用平均值时，可以犯的错误越多，就越可能做出预测（No Good）。
- 当我们有一个随机变量具有相应的概率分布时，我们可以计算期望、方差和标准偏差来描述其特征。
- 对于已经采样或观察到的数据，我们可以计算样本平均值、方差和标准偏差来描述其特征。
- 注意到：当样本大小趋于无穷大时，我们的期望与样本平均值相等。

19协方差和相关性

- 线性关系捕捉方法，主要是协方差和相关性，即：它们是用于描述两个或多个随机变量之间关系的重要工具。
- 协方差矩阵：可以用来描述多个随机变量之间的关系。
- 独立性和零协方差：理解随机变量之间关系的关键点。
- 区分这两个概念是非常重要的，因为它们对数据分析和模型构建有着不同的影响。在统计学中，两个随机变量独立和它们的协方差为零并不是同一个概念，虽然两者之间存在一定的联系。当两个变量的协方差为零时，它们之间不相关，但这并不意味着这两个变量是独立的。独立性意味着两个变量之间没有任何依赖关系，而不仅仅是它们的协方差为零。

20马尔可夫过程

- 马尔可夫过程，是强化学习的一个重要组成部分。
- 马尔可夫过程对于人工智能的强化学习范式是非常重要的。他们以系统的所有可能状态为特征，一组所有可能的动作可以由代理（向左移动、向右移动等）执行。它包含一个状态转移概率矩阵，该矩阵描述了所有状态之间的转移概率，即代理在采取某个操作后将转换到哪个状态的概率分布。
- 马尔可夫过程还涉及奖励函数，这是强化学习中的一个核心概念。奖励函数用于评估代理在某个状态下采取某个动作所获得的即时奖励。在强化学习中，目标是找到一种策略，使得代理在采取一系列动作后能够最大化累积奖励。

21标准化、缩放和/或标准化随机变量或数据集

- 有时，我们需要针对随机变量或数据集进行规范化、缩放和/或标准化处理，以便更好地控制它们的值；例如，要使它们居中于零，或将它们的扩散限制在小于或等于1的范围内，同时保持其固有的可变性（否则就是修改数据了）。
- 这个过程可以通过减去一个数字（移位）并对数据或随机变量进行除以常数（标尺）的操作来实现。常见的例子包括掷硬币、滚动骰子、抽出球等。

22常见示例

- **二项分布**：描述了在独立重复实验中成功次数的概率分布。

✓示例：预测在临床试验中，使用某种疫苗或新药后，出现副作用的患者数量。

- **泊松分布**：用于描述单位时间内（或空间内）稀有事件发生的次数的概率分布。

✓示例：预测一个小时内出生的婴儿数量，或某个时间段内机器故障的次数。

- **几何分布**：描述在独立重复伯努利试验中，首次成功所需试验次数的概率分布。

✓示例：估计一家公司在不经历网络故障的情况下可以运行多少周。

22常见示例

- **指数分布**：用于描述事件之间时间的概率分布，如无线通信中信号到达的间隔时间。

✓示例：预测机器部件故障前的工作时间。

- **Weibull分布**：常用于工程领域，用于描述产品的寿命。

✓示例：近似模拟汽车在使用前停止工作（考虑多个部件及其最薄弱环节）的寿命。

- **对数正态分布**：如果对一个随机变量的对数进行变换后，得到的数据服从正态分布，则该随机变量服从对数正态分布。

✓示例：描述石油储备中的气体体积或股票价格的变化。

22常见示例

- **卡方分布**：是正态分布随机变量的平方和的分布。

✓示例：用于检验数据是否符合正态分布，或检验两个数据集是否独立。

- **帕累托分布**：描述许多自然现象、社会现象和经济现象的分布，如财富分配。

✓示例：描述超级计算机完成一个作业所需的时间，或家庭收入水平。

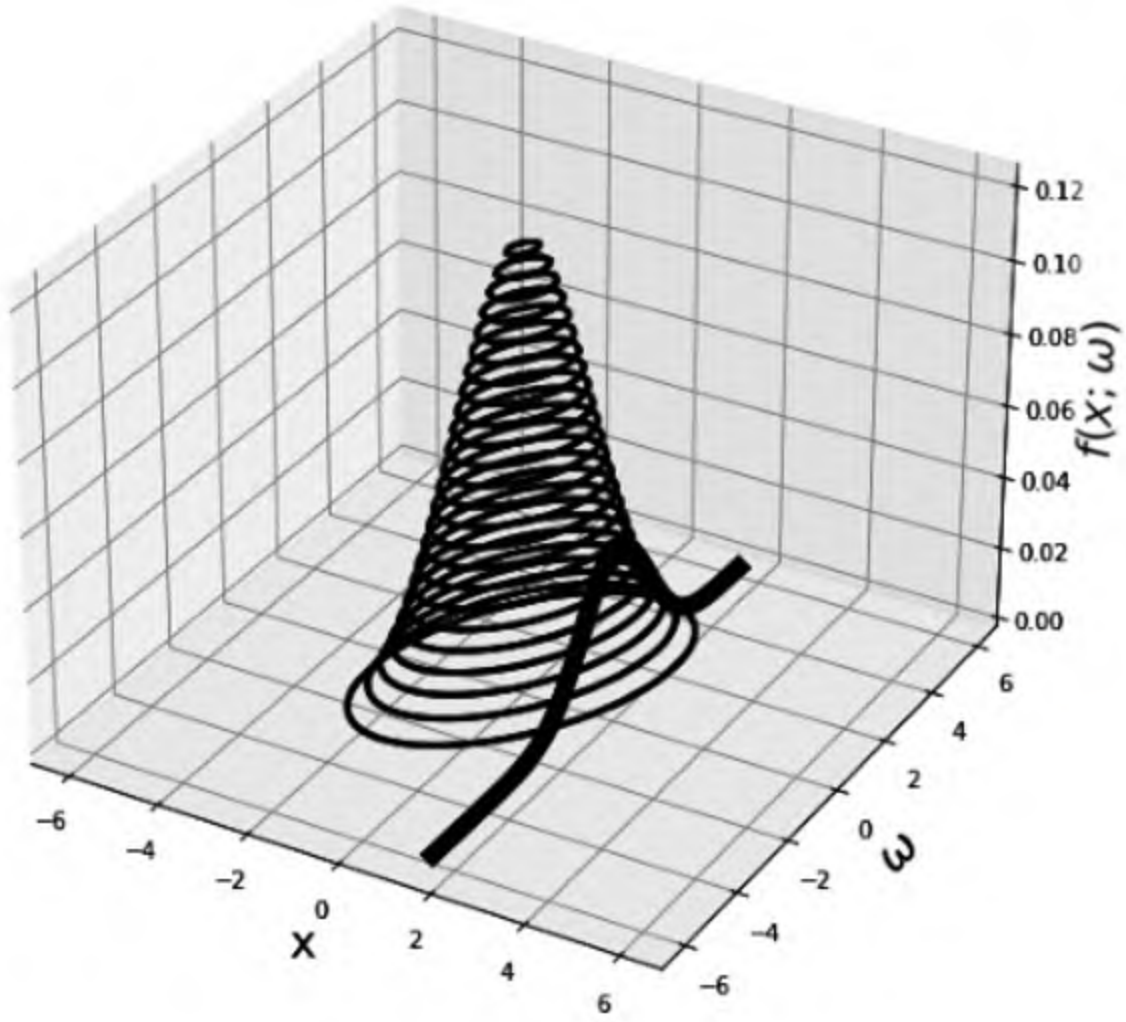
23连续分布与离散分布（密度与离散分布质量）

- 当我们处理连续分布时，使用观测器在某个值附近或周围对数据点进行采样，而不是观察或采样一个精确的值。事实上，当我们的数字在连续分布中，在一个值和下一个值之间，实数具有无限的精度。例如，某人身高228公分，或是227.98989898989976798989898768公分。
- 离散随机变量没有这样的担忧，因为可以很容易地分离相互之间的可能值。例如，当我们掷骰子，数值就是1、2、3、4、5或6。
- 由于这种推理，当我们有一个连续的随机变量时，我们定义其概率密度函数，而不是其概率质量函数，因为连续随机变量的结果是无法准确确定的。
- 对于多个连续随机变量的情况，我们需要使用更高维度的概率密度函数来表示它们的联合分布。

23连续分布与离散分布（密度与离散分布质量）

- 测量理论在数学步骤中的作用，避免了出现悖论的问题。
- 罪魁祸首是实数的无限精度。如果我们允许所有集合都具有概率，在我们可以构造不相交集的意义上，我们遇到了悖论（例如，分形形状的集或通过转换有理数集而形成的数字的集）：其概率加起来超过一个1啊。
- 在数学步骤中，并提供了一个数学框架，我们可以在其中工作具有概率密度函数，而不会遇到悖论。它定义了测量零（它们在我们工作的空间中不占用体积），然后给出了许多定理，使我们几乎可以在任何地方进行计算，也就是说，除了测量零点的集合，使得我们能够在其中安全地处理概率密度函数。

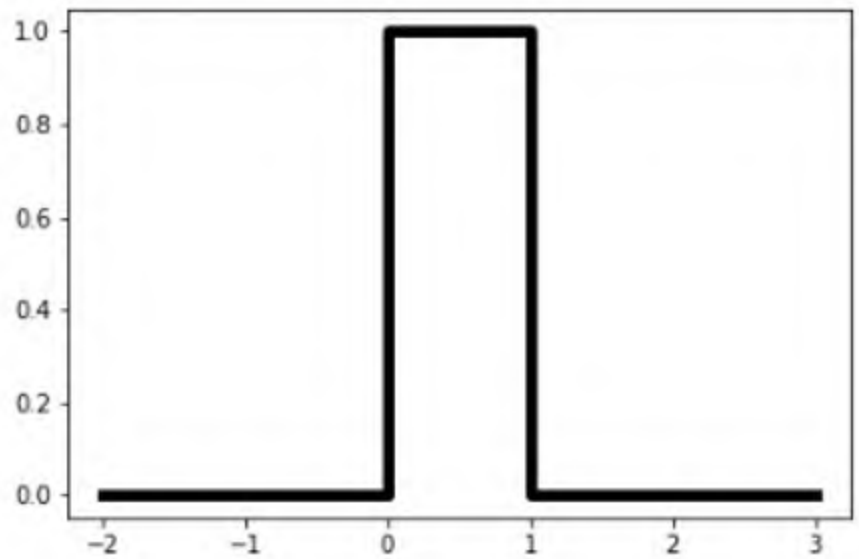
24联合概率密度函数的幂

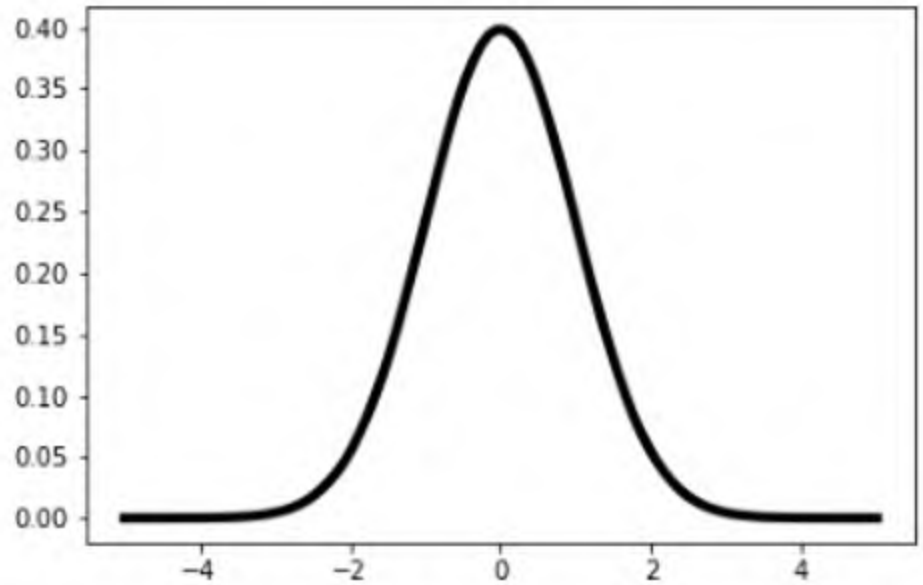


- 如果随机变量是独立的，那么联合概率密度函数就是各个随机变量概率分布的乘积。
- 如果随机变量之间存在依赖关系，那么联合概率密度函数就变得非常重要，因为它可以帮助我们更好地理解和处理数据。
- 在实际应用中，我们可以使用贝叶斯规则来推导出后验概率分布，从而得到更加准确的结果。

25数据分布：均匀分布

- 均匀分布适用于数据在某一区间内均匀分布的情况，而正态分布则更适合模拟人类身高的数据分布。
- 正态分布的特点是样本倾向于聚集在平均值附近，并且随着距离平均值的增加而逐渐变窄。





26数据分布：钟形正态（高斯）分布

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

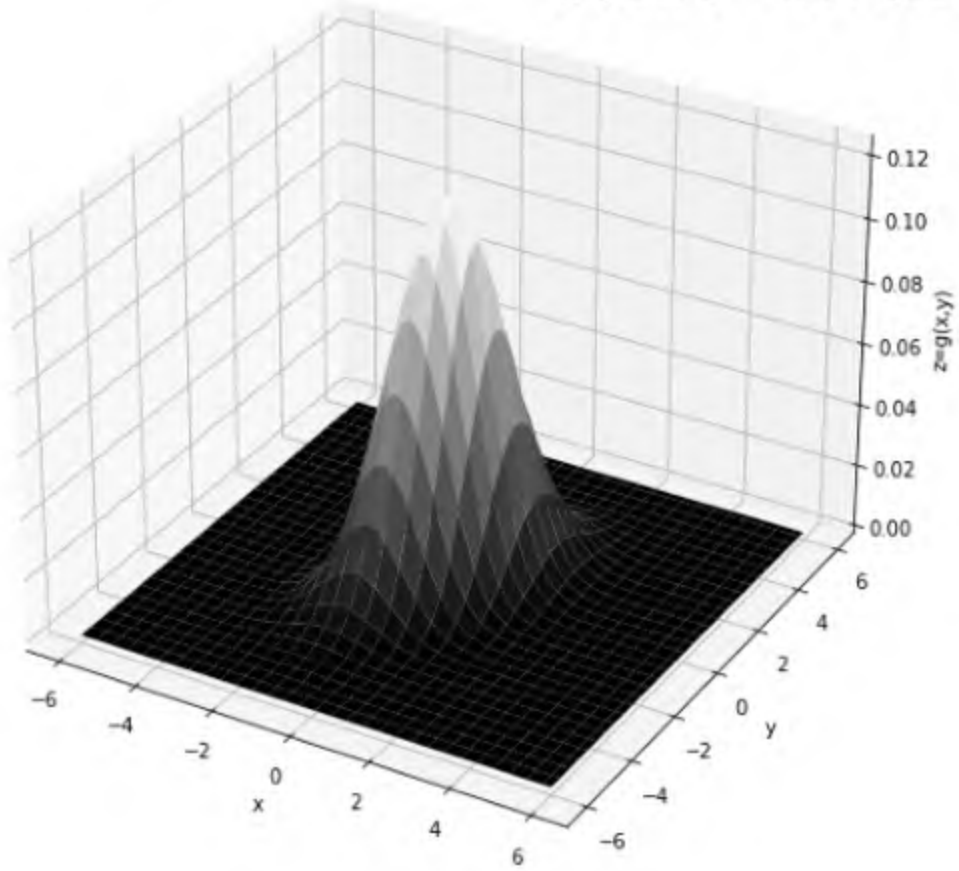
- 对于单变量正态分布，只有平均值μ和标准偏差σ两个参数；
- 对于双变量正态分布，则需要考虑两个随机变量的平均值、标准差以及它们之间的相关性。
- 在实际应用中，我们可以使用线性代数的方法来计算多变量正态分布的概率密度函数。通过绘制图形和采样点，我们可以更好地理解正态分布的特点和规律。

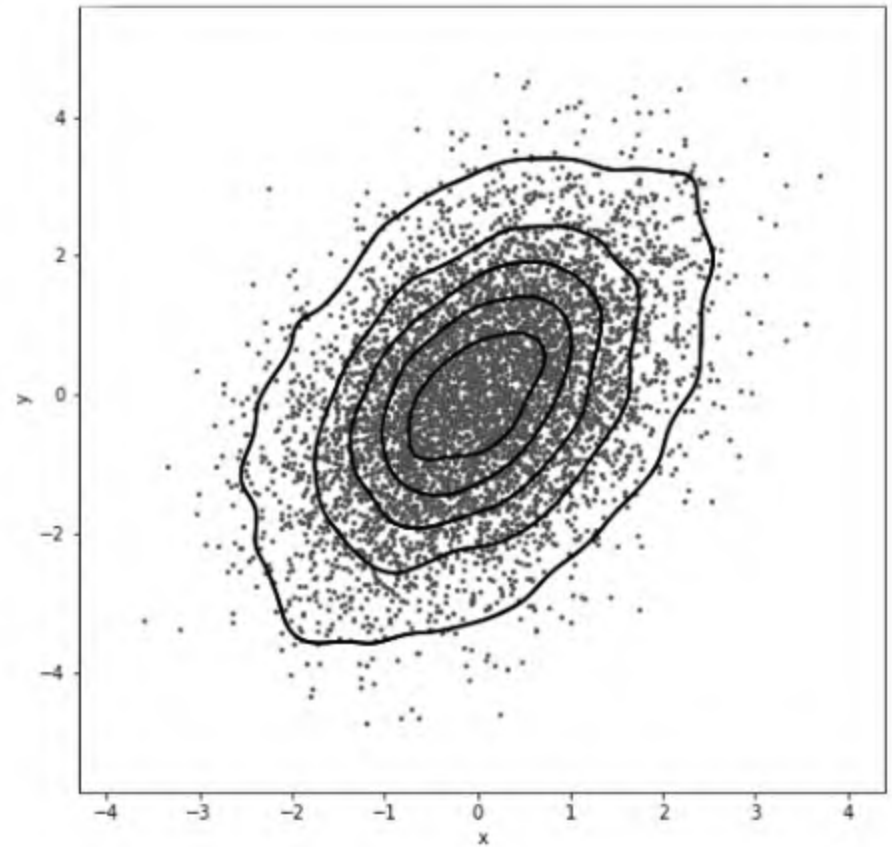
26数据分布：钟形正态（高斯）分布

$$g(x,y;\mu_1,\sigma_1,\mu_2,\sigma_2,\rho) = \frac{1}{\sqrt{(2\pi)^2 \det\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}}} e^{-\left\{\frac{1}{2}(x-\mu_1 \;\; y-\mu_2)\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}^{-1}\begin{pmatrix}x-\mu_1\\ y-\mu_2\end{pmatrix}\right\}}$$

26数据分布：钟形正态（高斯）分布

$$g(x,y;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^2 \det(\Sigma)}} e^{-\left\{\frac{1}{2}(u-\mu)^T \Sigma^{-1}(u-\mu)\right\}}$$





27数据分布：其他重要和常用的分布

- **二项分布**用于预测重复实验成功次数的概率，泊松分布则用于预测在给定时间内事件发生的次数。
- **几何分布**用于预测在执行独立试验时成功，每个试验的概率为成功的概率；
- **指数分布**则用于预测事件发生的等待时间，如等待地震发生的时间或机器部件故障之前的时间。
- **威布尔分布**广泛应用于工程预测领域，用于计算产品寿命并确定最薄弱的环节。
- **对数正态分布**在处理平均值较低、较大的倾斜数据时很有用，通过取对数后可得到正态分布的数据。

28“分布”一词的各种用法

- 常见的概率分布，包括卡方分布、帕累托分布、学生t分布、贝塔分布、柯西分布、伽马分布和负二项分布。其中，卡方分布是正态分布平方和的分布，用于计算随机变量的方差；帕累托分布则适用于一些实际应用场景，如机器学习计算、家庭收入水平等。
- 经验分布是指根据实际观测数据得出的一种概率分布，通常用于描述某一特征的频率分布情况。概率质量函数是指离散型随机变量取不同值的概率分布函数，即每个可能取值对应的概率值。累积分布函数是指随机变量小于或等于某个特定值的概率分布函数，通常用于计算某个区间内的概率。概率密度函数则是连续型随机变量的概率分布函数，表示在某一区间内取值的可能性大小。

28“分布”一词的各种用法

- 超几何分布和负超几何分布都是离散型概率分布，用于描述从有限总体中随机抽取若干次，每次抽样有放回或不放回的情况下的概率分布。
- 它们都涉及到一个总体容量N、一个成功的容量K和一个被选中的容量n的关系。而“分布”这个词语在这里有多种用法，包括经验分布、概率质量函数、累积分布函数等等。
- 具体来说，“分布”可以用来表示一组数据的分布情况，也可以用来表示某种随机变量的概率分布函数。

29A/B测试

- A/B测试是一种常用的实验设计方法，用于比较两个或多个版本的产品、网站、广告等的效果。
- 它的基本思想是将用户随机分成两组或多组，分别展示不同版本的产品或接受不同的营销策略，然后对比各组的表现指标，如点击率、转化率、留存率等，从而确定哪个版本更优秀。
- A/B测试在互联网产品开发、市场推广等领域广泛应用，是数据驱动决策的重要手段之一。

30把数据拟合到函数

- 将数据拟合到函数中，以便进行预测和决策。这种方法可以应用于计算机视觉、自然语言处理等领域。
- 我们可以先选择一个合适的函数形式，比如线性回归、多项式回归、支持向量机等等，然后根据已有数据来确定函数的参数，使得函数能够最好地拟合数据。
- 一旦我们得到了这个函数，就可以用它来预测未来的结果或者做出决策。
- 注意到：为了保证拟合结果的准确性，我们需要有足够的高质量数据，并且需要对不同的函数形式和参数组合进行充分的比较和评估。

谢谢！

gulp@mail.las.ac.cn