第一章 绪论

数据

1.数据的定义：是对客观事物、现象或过程的某种记录和表示，它以符号、数字、文字、图像、声音等形式存在，能够被存储、传输、处理和分析，用于反映信息和支持决策。

2.数据的类型：数据可以是**定性的**（例如描述性的，如“味道好闻”）或**定量的**（例如具体的数值，如体重70kg）。

3.数据的用途：把“客观记录”转化为“有价值的洞察和行动”，它支撑决策、推动预测、发现规律、优化运营，并不断创造新的应用和产业。

4.数据的来源和普遍性：数据可以来源于多种渠道，如**日常交易**、**IoT设备**、**互联网活动**等。数据在我们的日常生活中无处不在。

对这些大量数据进行**有效管理**和**深入分析**是一个复杂且具有挑战性的任务。

⇩

为了系统地研究和分析数据，深入理解其含义，并将这些信息作为决策制定和问题解决的有效工具，**数据科学**应运而生。

数据科学

数据采集 → 数据清洗 → 数据表示（特征工程）→ 数据评估 → 建模 → 模型评估 → 应用

- 数据科学是一个**综合性学科**，它融合了**统计学**、**计算机科学**、**信息科学**以及相关学科的技术和理论，专注于从**结构化**和**非结构化数据**中提取知识。

数据收集（把“问题相关的原始信息”变成“可分析的数据资产”）

数据清洗（处理缺失，异常，统一格式，去重等）

数据表示（转化为合适的结构，例如向量、张量、特征矩阵、图邻接矩阵）

数据评估（判断数据是否可靠、完整、可用）

建模（构建模型，能够解释数据、发现规律、或者做预测。）

模型评估和应用（评估模型泛化能力，部署模型）

数据科学的历史

- **1962年**，约翰·图基（John Tukey）在其论文《数据分析的未来》中首次提出了“数据分析”（Data Analysis）的概念，强调了统计学与计算机技术的融合，为现代数据科学奠定了基础。
- **1974年**，彼得·诺尔（Peter Naur）在其著作《计算方法简明调查》中使用了“数据科学”（Data Science）一词，定义其为“处理数据的科学”，并将其视为计算机科学的替代表达。
- **1985年**，C.F. Jeff Wu 在北京的中国科学院讲座中首次提出将统计学更名为“数据科学”，以突出统计学与新兴数据处理技术的融合趋势。
- **2001年**，威廉·S. 克利夫兰（William S. Cleveland）在国际统计学会会议上提出将数据科学作为独立学科的构想，强调数据分析、计算方法与信息技术的交叉整合。
- **2008年**，DJ Patil 和 Jeff Hammerbacher 在 LinkedIn 和 Facebook 分别首次使用“数据科学家”（Data Scientist）这一职位名称，标志着数据科学作为职业角色的正式确立。

约翰·图基

彼得·诺尔

威廉·S. 克利夫兰

Jeff Wu

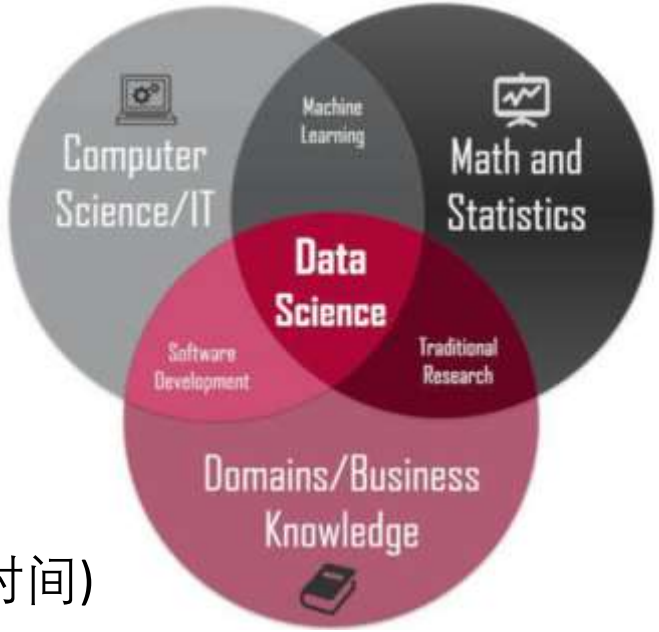
DJ Patil

数据科学家

- 数据科学家1）在统计学方面的能力超越了普通软件工程师，2）在计算机科学的技能上也超出了一般统计学家。这样的跨学科专长要求数据科学家不仅**精通数据分析所需的统计技术**，同时也能够**通过编程实践这些技术**。此外，3）他们还需对其服务的行业有深入的理解。

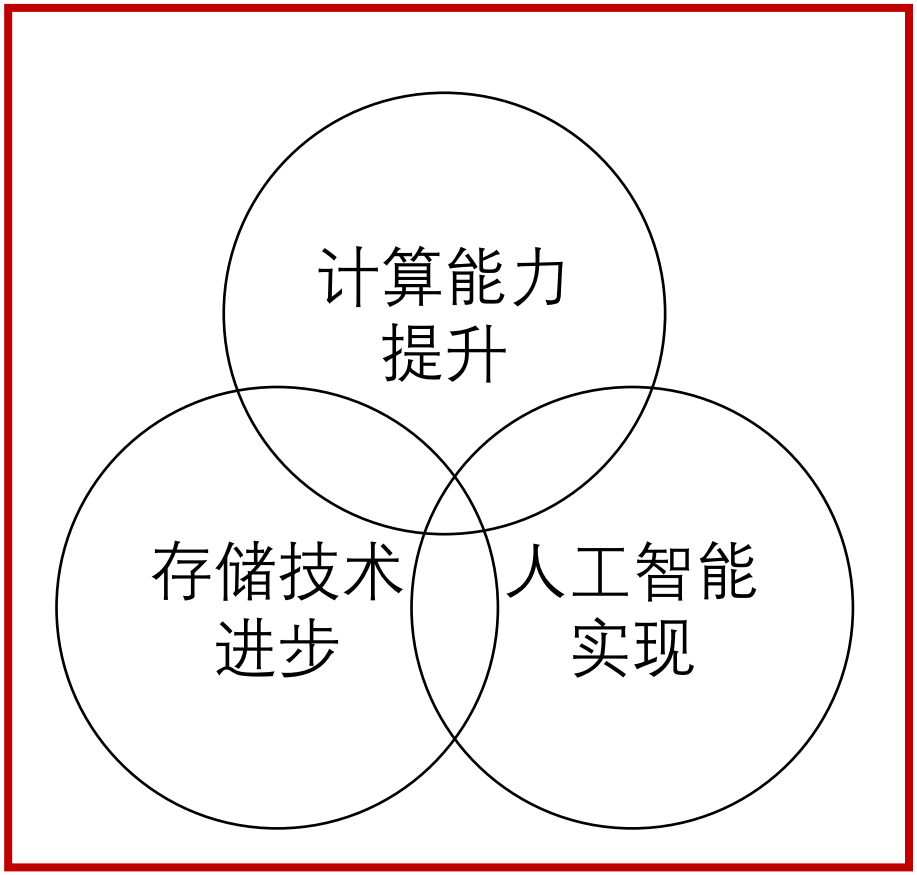
数据科学家的工作（来自一位知乎网友）：

- 商业逻辑理解与思考（占10%时间）
- 数据检查与清洗（占25%时间）
- 特征工程（占20%的时间）
- 数据建模（占5%时间）
- 与客户,同事，或者上级沟通（占20%时间)
- 写模型文档，数据分析文档等。（占15%时间)



从数据到大数据?

- 大数据指的是分布在多个系统上的大规模、非集中化的原始数据集。这些数据以高速度从各种来源产生，并且呈现出多样的格式和结构。

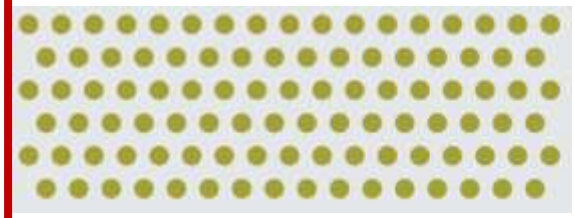


- 网络与传感技术的发展（互联网，物联网（IoT）、移动设备和传感器产生了前所未有的数据规模）
- 分布式与并行计算架构（Hadoop、Spark 等大数据框架使得海量数据可以被高效存储和处理。）
- 数字化转型（金融、医疗、制造、教育等行业）
- 政策与基础设施支持（国家级数据战略、云计算平台、5G 基础设施建设）
- …

大数据的四个维度

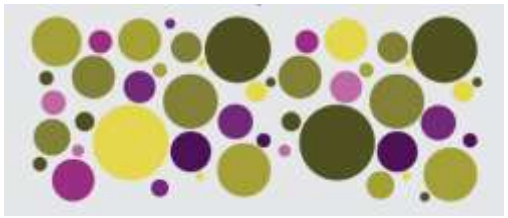
- 四个维度（4”V”： Volume, Variety, Velocity, Veracity, Value ）

体量（Volume）



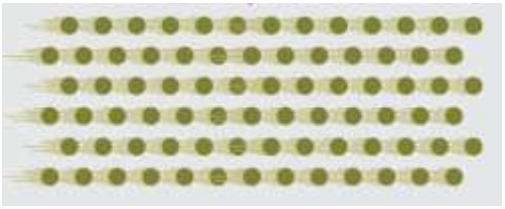
数据规模庞大，达到 TB、PB 甚至 EB 级别。

多样性（Variety）



数据来源和类型多样，涵盖结构化、半结构化和非结构化数据。

速度（Velocity）



数据生成、传输与处理的速度极快，要求实时或准实时分析。

真实性（Veracity）

数据质量参差不齐，存在噪声、不确定性和可信度问题。

大数据的四个维度

- **Volume（体量）**

体量是大数据的一个核心特征，它指的是**巨量的数据。**然而，定义“巨大的数据量”是相对的，这在不同的行业、领域和地区之间会有所不同。随着技术的进步，今天被视为巨大的数据量在未来可能会显得更加普遍。

- 随着随身设备、物联网、云计算和云存储的发展，人和物的行为轨迹几乎都能被记录，数据因此呈现爆炸式增长。
- 移动互联网时代，每个人都成为数据的制造者，短信、微博、照片、视频等不断产生。
- 与此同时，传感器、监控、刷卡机、收款机、ETC、网络点击等自动化设施和流程记录源源不断地产生数据。
- 来自电信、互联网、政府、金融、商业、交通等各类机构的大量数据最终汇聚，形成了庞大的“数据之海”。

大数据的四个维度

- **Variety（多样性）**

大数据的多样性指的是大数据具有各种不同类型的数据，包括结构化数据（如数据库中的数据）、半结构化数据（如电子邮件）和非结构化数据（如视频和图片）。

- 北京市交通智能化分析平台的数据来源丰富。除了来自路网摄像头和传感器的数据外，还包括公交、轨道交通、出租车等多种交通工具的运营数据。例如，4万辆浮动车每天生成2000万条记录，而交通卡刷卡记录、手机定位数据、出租车运营数据以及电子停车收费系统的数据，每天也会产生数百万条数据点。这些数据**种类繁多**，覆盖了交通领域的各个方面。
- 电子商务平台的数据来源广泛，涵盖了结构化数据（如产品信息、用户行为）和非结构化数据（如图片、评论、社交媒体内容等），展示了大数据的**多样性**（Variety），即不同来源和不同类型的数据在平台上同时存在。

大数据的四个维度

- **Velocity（速度）**

数据的生成、处理和分析需要足够“快”。这主要是数据的实时生成以及业务流程和决策中的迫切需求。提升速度能够显著缩短数据的时延，对于时间敏感型业务至关重要，例如实时欺诈监测或高频交易等场景。

- 汽车导航系统能够**实时**响应交通状况变化，**快速**重新规划行车路线，确保驾驶者最有效率地抵达目的地。
- 航空公司根据航班的搜索量和余票情况，**即时**调整票价，优化利润。
- 酒店通过**实时**更新多个在线平台（如携程、飞猪、去哪儿）上的房间信息，确保各平台显示的剩余房间数量一致。
- 银行借助实时监测系统，**迅速**识别并应对信用卡盗刷行为，保障客户资金安全。
- 公安机关通过**即时**监控系统，**快速**识别并处理电信诈骗活动，保护公众免受欺诈。

大数据的四个维度

- **Veracity（真实性）**

  数据真实性。大数据的真实性（Veracity）指的是数据的准确性、可信度和可靠性。这一特征强调的是数据本身的质量，包括数据的来源、完整性和上下文的正确性。在大数据环境中，因为数据量大且来源多样，数据的质量可能参差不齐。

- 电商平台聚集了成千上万的消费者评论和评分。大数据的真实性挑战在于如何确保这些评论和评分是真实可靠的，而不是虚假或误导性的。例如，一些评论可能来自真正的消费者，而另一些可能是由制造商或竞争对手伪造的。
- 在社交媒体平台上，用户发布的内容、点赞、分享和评论等构成了海量的数据。市场研究人员和品牌经理利用这些数据来分析消费者行为和市场趋势。然而，数据真实性的挑战在于确保所分析的数据能够代表真实的用户意见和行为，而不是由虚假账户或自动化的文本生成工具产生的误导信息。（这在人工智能和生成模型高速发展的今天尤其重要）

大数据的来源

1. **社交数据**：包括微信聊天记录、抖音短视频、各种在线评论、搜索记录以及通过社交媒体平台上传和分享的图片、视频等内容。
2. **机器数据**：来自工业设备、机械传感器以及追踪用户行为的网络日志。典型来源包括医疗设备、智能穿戴设备、交通摄像头、汽车、智能家居以及物联网设备等。
3. **交易数据**：指线上和线下交易过程中产生的所有数据。包括发票、付款订单、存储记录、交货收据等，广泛存在于电商平台、零售商和支付系统中。
4. **网络数据**：主要来源于各种静态网页，这些数据包括网站内容、用户浏览行为、点击流等，能够反映出用户的兴趣和需求。
5. **数据库数据**：包括医疗数据库、学术数据库等，这些结构化的数据库内容为各行业提供了大量的数据支持。
6. **政府和公共数据**：来自政府机构和公共部门的数据，涵盖了人口普查、经济统计、交通流量、公共健康等方面的信息，具有重要的社会价值。
7. **卫星与遥感数据**：通过卫星、无人机及其他遥感技术采集的数据，广泛应用于地理信息系统、环境监测、农业、气候变化研究等领域。
8. **生物医学和基因组数据**：医疗和健康领域产生的数据，包括临床试验数据、病人健康记录、基因组数据等，这些数据为精准医疗和疾病研究提供了基础。

大数据涉及到的技术

数据管理

- 数据存储
- 数据清洗
- 数据评估
- 数据备份与恢复
- …

数据流

- 数据采集
- 数据传输
- 数据同步
- …

数据分析

- 统计建模
- 机器学习
- 可视化技术
- 自然语言处理
- …

大数据的关键处理步骤

麦肯锡全球研究所的报告描述了大数据的主要组成部分和生态系统，主要包括：

1.数据分析技术：如机器学习、自然语言处理等用于解析数据的高级技术。
2.大数据技术：包括云计算、数据库等支撑大数据存储和处理的技术。
3.数据可视化：通过图表、图形及其他数据展示方式，直观地呈现数据分析结果。

大数据涉及到的技术

数据管理

- 数据存储
- 数据清洗
- 数据评估
- 数据备份与恢复
- …

数据流

- 数据采集
- 数据传输
- 数据同步
- …

数据分析

- 统计建模
- 机器学习
- 可视化技术
- 自然语言处理
- …

大数据的关键处理步骤

麦肯锡全球研究所的报告描述了大数据的主要组成部分和生态系统，主要包括：

1.数据分析技术：如机器学习、自然语言处理等用于解析数据的高级技术。
2.大数据技术：包括云计算、数据库等支撑大数据存储和处理的技术。
3.数据可视化：通过图表、图形及其他数据展示方式，直观地呈现数据分析结果。

大数据的应用

1. 理解客户、满足客户服务需求

大数据的应用可以极大地增强企业对客户行为的理解和预测能力，进而显著提升客户体验。

- 微信聊天记录来推荐商品
- 通过淘宝搜索记录做个性化推荐
- **大数据杀熟**（基于用户的历史数据来对不同用户展示不同价格的）

大数据的应用

2. 改善生活

大数据同样深入到我们日常生活的每个角落。比如，智能手表和手环等设备持续生成数据，来监测健康指标；交友平台依靠算法分析个人偏好、行为习惯以及交互模式，帮助用户找到兴趣爱好更接近的朋友。

某手环记录的数据（通过数据仪表板来展示）

大数据的应用

3. 提高医疗和研发

通过分析健康监测数据和临床数据（医疗大数据），医生能够制定个性化的治疗方案，及时预测疾病风险。例如，医院通过监控早产儿的生命体征，提前发现并处理潜在的健康问题，从而提高医疗的准确性和救治的成功率。

大数据的应用

4. 提高体育成绩

大数据分析技术能够提升训练效果和竞技表现。例如，通过视频分析技术，我们能够精确追踪足球比赛中球员的运动轨迹，活动热区以及战术表现。同时，装配在运动器材上的传感器（22年的世界杯的足球内部就装有传感器）能够提供详细的比赛性能数据，帮助裁判做出正确的判罚，也能够帮助教练员分析如何提高技巧和战略。

大数据的应用

5. 优化机器和设备性能

大数据分析能提升机器和设备的智能化。电车在行驶过程中实时记录加速度、刹车力度、电池充电状态和定位信息。即使车辆停止，关键信息如胎压，电池状态，周边状况也会持续传输到手机端，以便进行故障预警和性能监控。这些数据能够帮助汽车厂家分析故障原因，了解用户的驾驶习惯，并进一步开发自动驾驶功能等。

大数据的应用

6. 改善安全和执法

大数据的运用能够改善安全措施和执法工作。1）企业通过大数据技术分析网络流量模式，有效预防和响应网络攻击。2）执法机构利用大数据工具分析犯罪模式和行为趋势，更迅速地追踪并捕捉罪犯。3）信用卡公司则运用大数据算法实时监控交易活动，以便快速识别和阻止欺诈性交易。

大数据的应用

7. 改善我们的城市

大数据在城市管理和优化方面也发挥着重要的作用。利用实时交通流数据，交通管理者能够优化交通信号、减少拥堵，并提高公共交通的效率。社交媒体和天气数据的分析帮助应急管理部门在自然灾害发生时迅速做出反应，同时指导居民避开危险区域。

大数据的应用

8. 金融领域

1.风险控制： 中国人民银行的征信系统使得金融机构能够通过征信系统来评估和控制贷款风险，从而准确预测借贷违约的可能性，并作出更明智的放贷决策。

2.保险定价： 在保险行业，特别是车险领域，大数据分析允许保险公司根据车主的事故历史、职业、年龄、性别等多种因素来定制个性化的保险产品。

3.高频交易（HFT）:在股票市场，高频交易利用大数据算法在极短的时间内分析市场数据和外部信息，来作出交易决策。这些算法能够捕捉到微小的价格变动，并在秒级别内自动执行买卖订单，利用市场波动赚取利润。

课堂思考

1. 写出大数据的四个特点
2. 为什么会实现“数据”到“大数据”的跨越（写出至少两个原因）

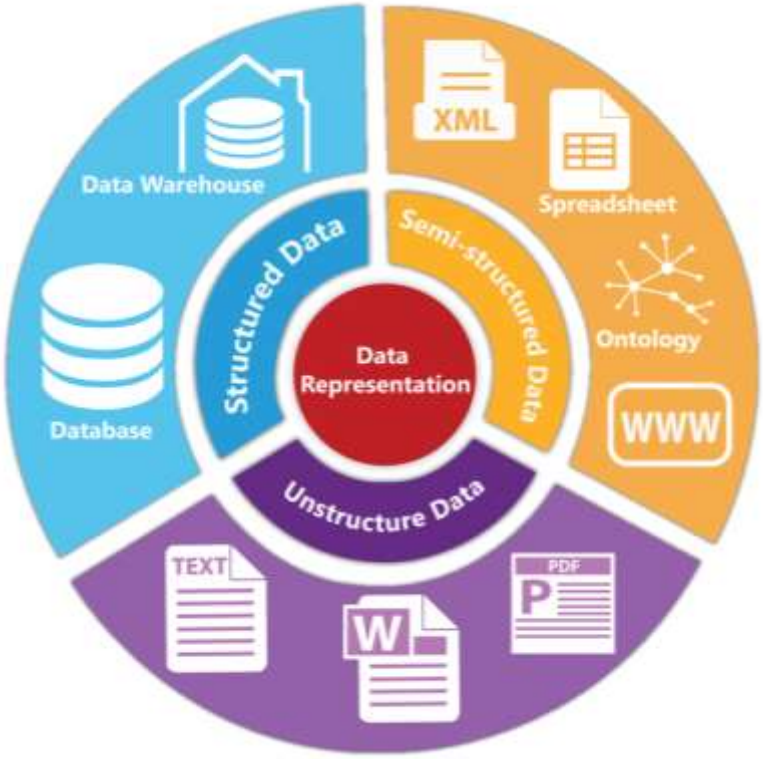
第二章 数据分析

数据类型

- **结构化数据**，结构化数据指的是以固定格式组织的数据，使其易于存储、查询和分析。这种数据类型通常在数据库中管理，但不限于此。具体场景包括：企业ERP（企业资源计划系统），财务系统，医疗HIS（医院信息系统）数据库，教务系统等。

- **非结构化数据**指的是不遵循固定格式或数据模型的数据，如文本、图像、视频和音频文件，以及各种办公文档和社交媒体内容。由于其结构不规则，处理和分析这类数据相比结构化数据更为复杂。

- **半结构化数据**，半结构化数据是一种介于结构化数据和非结构化数据之间的数据类型。这类数据部分遵循结构化数据的模型，但同时保留了一定的灵活性和不规则性。常见的半结构化数据格式包括JSON、XML等。



数据类型的例子

结构化数据

结构化数据遵循严格的格式和组织结构，通常存储在关系数据库中，易于查询和分析。例子包括：

1.数据库表格：如客户信息、销售记录、库存数据等。

2.电子表格：如Excel中的财务报表。

非结构化数据

非结构化数据没有固定的格式或组织结构，通常需要更复杂的方法来处理和分析。例子包括：

1.文本文件：如新闻文章、微博、社交媒体更新。

2.图像和视频：如照片、腾讯视频，B站。

3.音频文件：如音乐录音、播客。

半结构化数据

半结构化数据包含一定的组织结构，但这些结构没有严格的格式规定，介于结构化和非结构化数据之间。例子包括：

1.JSON和XML文件：广泛用于网络数据传输。

2.HTML文档：网页的标记语言，包含文本内容和结构标签。

3.日志文件：如服务器日志，包含日期、时间戳和事件描述。

半结构化数据的例子

```json
{
  "person": {
    "name": "John Doe",
    "age": 30,
    "email": "johndoe@example.com",
    "address": {
      "street": "123 Main St",
      "city": "New York",
      "zip": "10001"
    },
    "phone_numbers": [
      { "type": "home", "number": "123-456-7890" },
      { "type": "work", "number": "987-654-3210" }
    ]
  }
}
```

JSON文件例

```xml
<person>
  <name>John Doe</name>
  <age>30</age>
  <email>johndoe@example.com</email>
  <address>
    <street>123 Main St</street>
    <city>New York</city>
    <zip>10001</zip>
  </address>
  <phone_numbers>
    <phone type="home">123-456-7890</phone>
    <phone type="work">987-654-3210</phone>
  </phone_numbers>
</person>
```

XML文件例

数据类型的观点

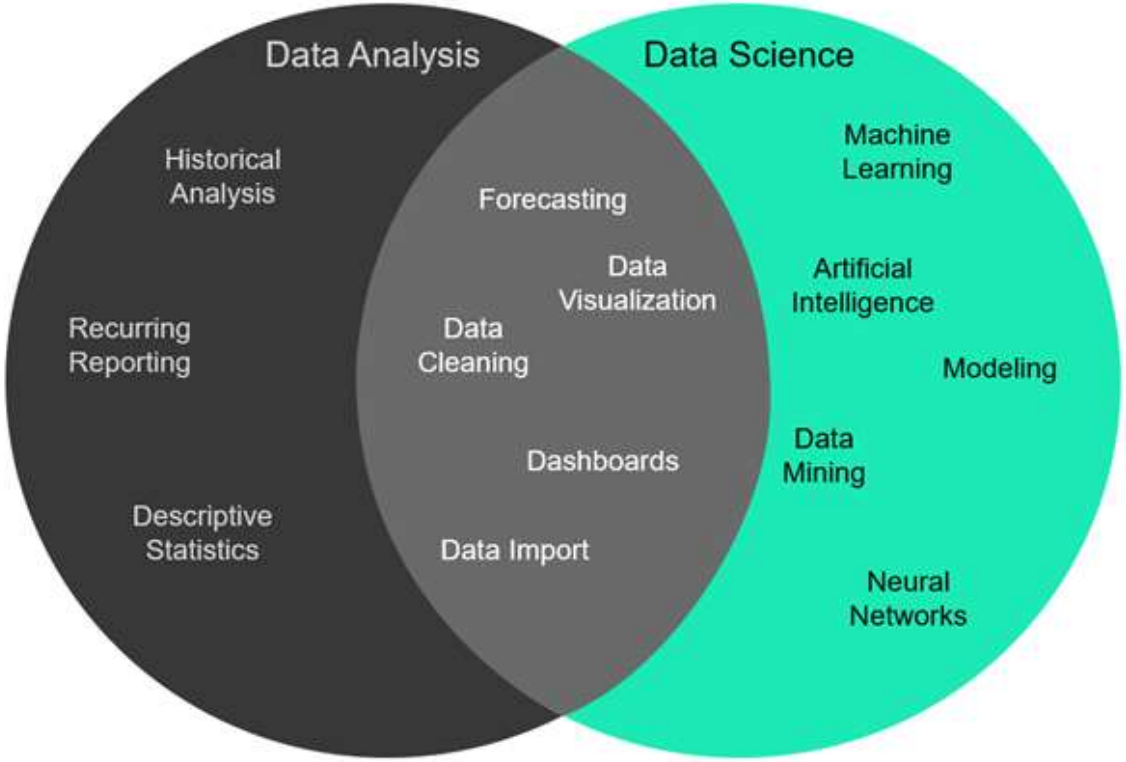
- **非结构化数据比结构化数据多很多。**大数据时代之前，数据多以结构化形式存在，它们具有明确的格式和组织结构。随着计算机、互联网和数字媒体的普及，非结构化数据，比如文本、图像、音频和视频的数量急剧增加，并已成为数据宇宙中的主要组成部分。

- **非结构化数据比结构化数据要难理解的多。**非结构化数据通常不遵循固定的格式或结构，使得它难以标准化和理解。相比之下，结构化数据具有明确的组织结构和格式，因此更易于管理和分析。处理非结构化数据通常需要如自然语言处理、图像识别和机器学习等技术，来提取有用信息。

数据分析vs.数据科学

数据分析：主要是对已有数据（如历史数据）进行整理、统计和解读，从中发现趋势或规律，以支持决策。例如，通过分析销售数据判断某个季度的销售高峰期或客户流失情况。

数据科学：在数据分析的基础上，更强调通过算法、机器学习和数据挖掘建立模型，预测未来趋势或发现隐藏模式，从而提供更智能的决策支持。例如，根据用户行为数据推荐个性化产品。



数据科学过程

1. 问题定义（Problem Definition）

- 明确业务或研究目标，确定数据科学要解决的问题。
- 例如：“预测下个月销售额”或“识别潜在欺诈交易”。

2. 数据收集（Data Collection）

- 获取相关数据来源，包括结构化数据（数据库）、半结构化数据（JSON、日志）、非结构化数据（文本、图片、音频等）。

3. 数据清洗与预处理（Data Cleaning & Preprocessing）

- 处理缺失值、异常值、重复数据。数据转换、归一化、特征工程。

4. 探索性数据分析（EDA, Exploratory Data Analysis）

- 使用可视化、统计描述发现数据规律、趋势和异常。

数据科学过程

5. 建模（Modeling）

- 选择合适的算法（回归、分类、聚类、深度学习等）建立预测或识别模型。
- 进行训练、调参和交叉验证。

6. 评估与验证（Evaluation & Validation）

- 使用指标评估模型性能（如准确率、召回率、RMSE）。验证模型在新数据上的泛化能力。

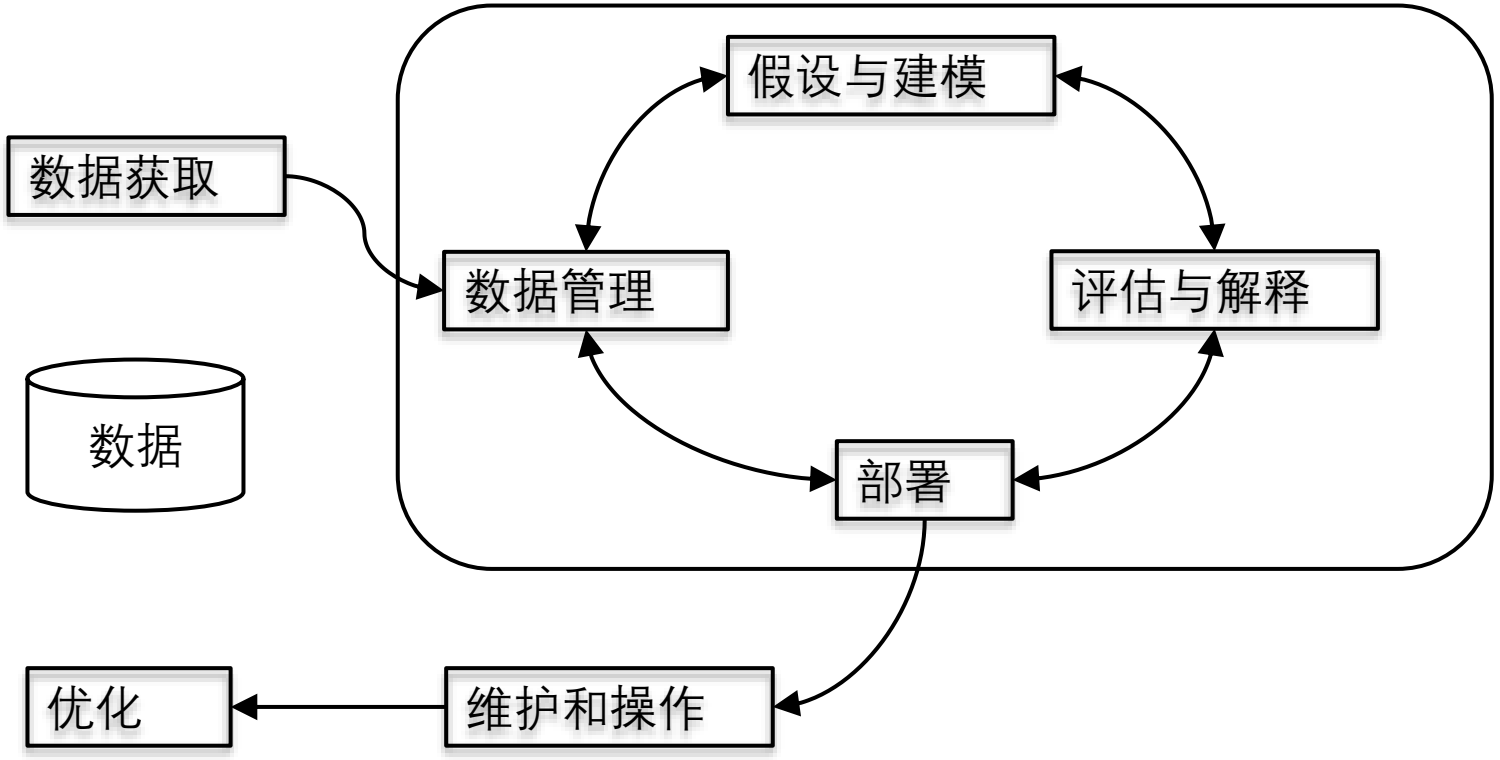
7. 部署与应用（Deployment & Application）

- 将模型或分析结果应用到实际业务中，实现决策支持或自动化系统。
- 例如，将销售预测模型接入电商后台，指导库存管理。

8. 监控与优化（Monitoring & Optimization）

- 持续跟踪模型效果，更新数据和算法以保持准确性和适应性。

数据科学过程



数据分析过程

1. 问题定义（Problem Definition）

明确分析目标：想要回答的问题是什么。例如："哪款产品在上季度销量最好？"、"客户流失的原因是什么？"

2. 数据收集（Data Collection）

获取相关数据，通常是已有的结构化数据（数据库、报表）、日志数据或问卷调查数据。

3. 数据清洗与整理（Data Cleaning & Preparation）

处理缺失值、重复数据和异常值。转换数据格式或合并不同数据源，确保数据质量。

4. 探索性数据分析（EDA, Exploratory Data Analysis）

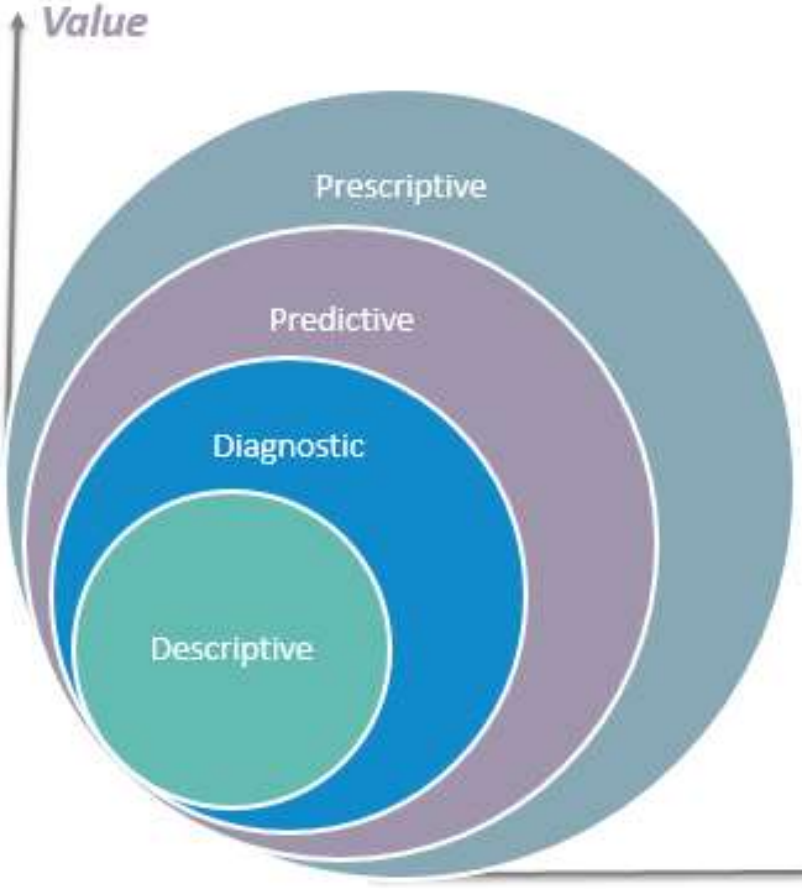
使用统计描述（均值、中位数、标准差等）、可视化（柱状图、折线图、散点图等）理解数据特征。发现趋势、异常、关联关系或潜在规律。

5. 结论与报告（Insights & Reporting）

根据分析结果总结关键发现，提出决策建议。通常以报表、图表或文字报告呈现。

数据分析的类型

4种数据分析的类型



- 1.描述性分析（Descriptive analysis)
  -在发生什么？
- 2.诊断性分析（Diagnostic analysis）
  - 为什么会发生？
- 3.预测性分析（Predictive analysis）
  -未来可能发生什么？
- 4.指导性分析（Prescriptive analysis）
  -我需要做些什么？

层层递进

描述性分析

在发生什么？

- **数据的目的性描述**：描述性分析是为了获得对数据的初步感知。帮助理解数据集的基本特性，但不涉及深入的原因或预测性分析。
- **利用统计学知识**：描述性分析依赖于基础统计学概念，如均值、中位数、众数、标准差等来帮助总结和解释数据的关键特征。
- **数据的可视化展示**：除了统计量之外，描述性分析还涉及数据的可视化。例如，使用条形图、饼图、散点图等来展示数据分布和模式。

例子

1.年度销售报告：分析公司在过去一年中每月或每季度的销售额，以及不同产品或服务的销售情况。

2.客户满意度调查结果：对顾客满意度调查数据进行汇总，展示顾客对公司服务的总体评价，包括平均满意度评分、反馈的常见问题等。

3.员工绩效数据：汇总员工的绩效评分，包括各部门或团队的平均绩效指标，以及表现最好的员工。

诊断性分析

为什么会发生？

- 诊断性分析**主要关注数据背后的原因和因果关系。**诊断性分析深入解释“为什么会这样”。诊断性分析通过深入挖掘数据中的模式、趋势和异常，帮助识别问题的根本原因和数据间的关系。

例子

1.公司销售额下降分析：一家公司注意到其最近一个季度的销售额比去年同期有显著下降。诊断性分析会审查销售渠道的效率、评估市场竞争情况、分析客户满意度。然后找到造成销售下降的具体原因，比如市场需求变化、新竞争者的出现或客户服务问题。

2.医院急诊室就诊率激增：一家医院发现其急诊室在过去几个月的就诊率急剧上升。通过诊断性分析，医院分析了病人的共同症状、他们来自的地区，并评估了近期的环境或社区健康事件。这些分析帮助揭示了可能的公共卫生问题，如流感爆发或环境污染，从而促使医院和公共卫生部门采取相应的应对措施。

预测性分析

未来可能发生什么？

- 预测性分析使用历史数据来预测未来趋势或结果。它比描述性和诊断性分析更复杂，因为它不仅涉及数据的理解和解释，还需要对未来进行合理的预测。

例子

1.销售预测：企业使用历史销售数据和市场趋势分析来预测未来的销售额。帮助做出库存管理、预算规划和市场策略调整等决策。

2.股市趋势预测：金融分析师利用过往的股市数据和经济指标来预测股市的未来走势，帮助投资者做出更明智的投资决策。

3.天气预测：气象学家使用历史天气数据和气候模型来预测未来的天气情况，这一点对农业、航运和公共安全等多个领域至关重要。

指导性分析

我们应该怎么做?

- 指导性分析**利用描述性、诊断性和预测性分析的结果来提供具体的行动建议**。简而言之，指导性分析将数据分析转化为实际可行的决策和行动。这种分析形式是“数据驱动决策”的核心。

例子

- **企业资源优化**：一家制造公司通过描述性和诊断性分析发现其生产线上某些环节效率低下。预测性分析进一步指出这种低效可能导致未来的产量下降，也会造成成本增加。在这基础上，指导性分析则提出具体的改进措施，比如重新配置生产流程、引入自动化技术或进行员工培训等。

- **市场营销策略**：一家零售商通过分析消费者的购买数据（描述性分析），了解了哪些产品类别最受欢迎（诊断性分析），并预测了未来的购买趋势（预测性分析）。基于这些信息，指导性分析提供了具体的营销策略建议，如调整产品线、制定针对性的促销活动或优化库存管理等。

第三章 基本学习算法

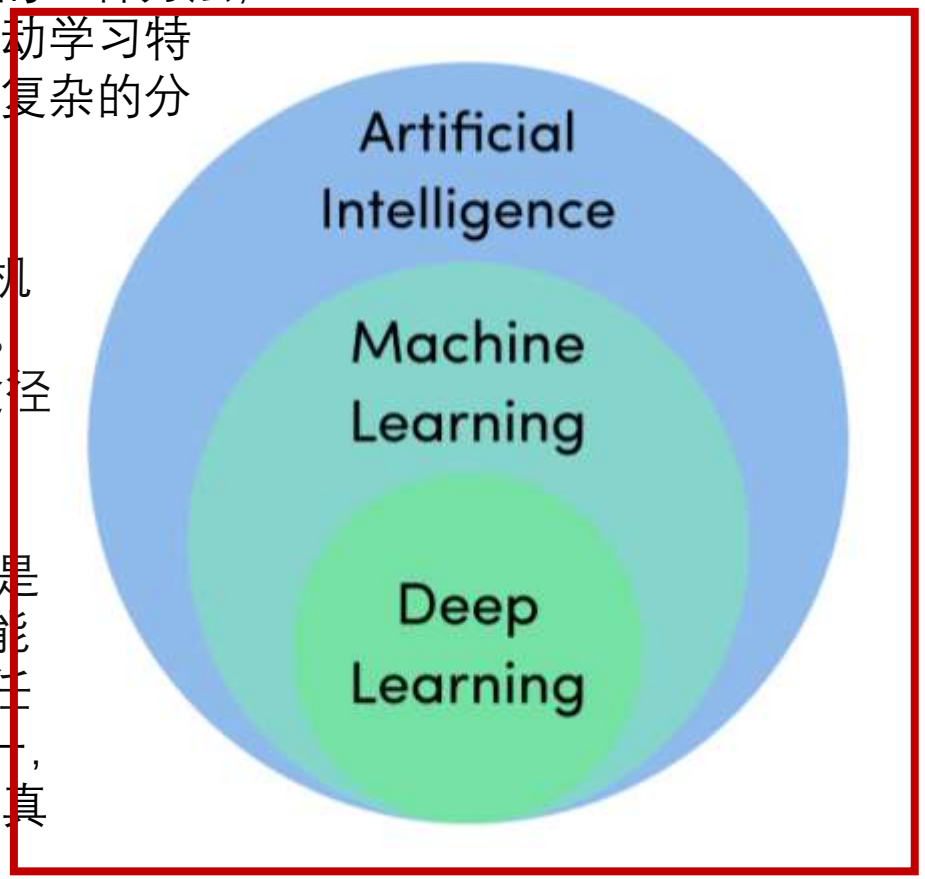
机器学习/深度学习/人工智能

- 机器学习指的是**利用算法使计算机能够从数据中学习，而不需要预设固定的规则。**这些算法使计算机能够自动作出决策或预测。

- 深度学习（Deep Learning）是机器学习的一种方法，它通过构建多层神经网络，从数据中自动学习特征表示和模式，实现比传统机器学习更复杂的分类、预测或生成任务。

- **人工智能**是一个广泛的概念，指的是使机器模仿人类智能行为的各种技术和方法。**机器/深度学习**是实现人工智能的关键途径之一。

- AGI（Artificial General Intelligence）指的是具有广泛认知能力的人工智能，这种智能在理论上能够像人类一样执行任何智能任务。实现 AGI 是人工智能的终极目标之一，但当前大多数 AI 系统仍属于狭义 AI，离真正的 AGI 还有很长的路要走。



机器学习的分类

- **基于学习方式的分类**

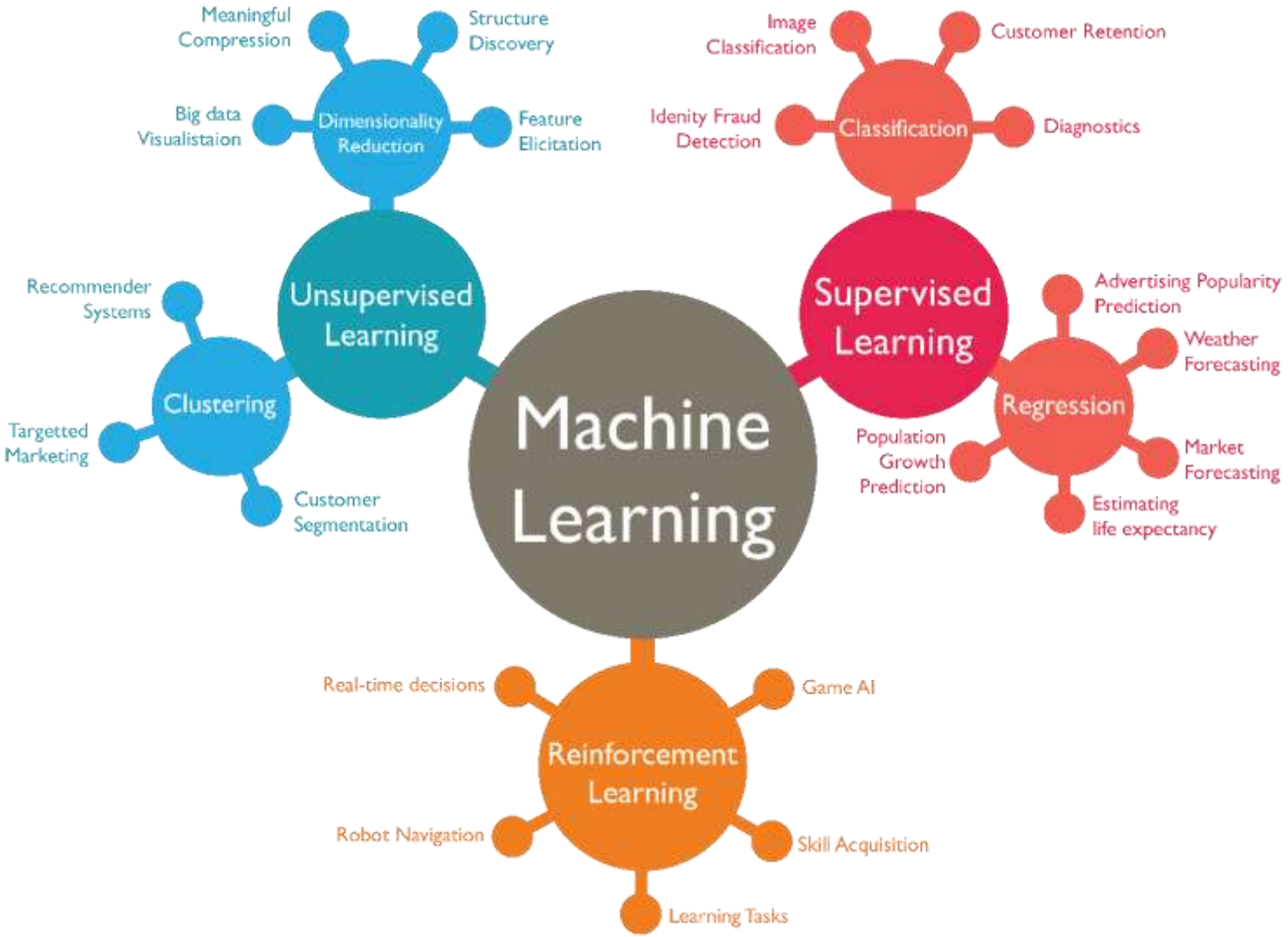
1.监督学习：这是一种**基于标注数据进行学习的方法**。在监督学习中，算法通过分析一组已经标记了类别的样本来调整和优化模型参数，从而提高其性能。这种学习方法主要包括两种类型：分类（确定样本的类别）和回归（预测连续数值）。

2.无监督学习：与监督学习不同，**无监督学习不依赖于标注数据**。它处理的是未经标记的数据集，目的是发现数据中的模式和结构。常见的无监督学习方法包括聚类（将数据分组为不同的类别）和降维（减少数据的维度，复杂性，同时保留重要信息）。

3.强化学习：强化学习也是一种机器学习方法。主要是通过学习的主体在环境中尝试不同的动作并根据结果获得奖励或惩罚来学习。这种学习的目标是使智能体能够在给定的任务中实现最大化的累积奖励。强化学习在处理需要连续决策和适应动态环境的问题时特别有效，如**游戏玩法、自动驾驶和机器人导航**等领域的应用。

机器学习的分类

- **基于学习方式的分类**

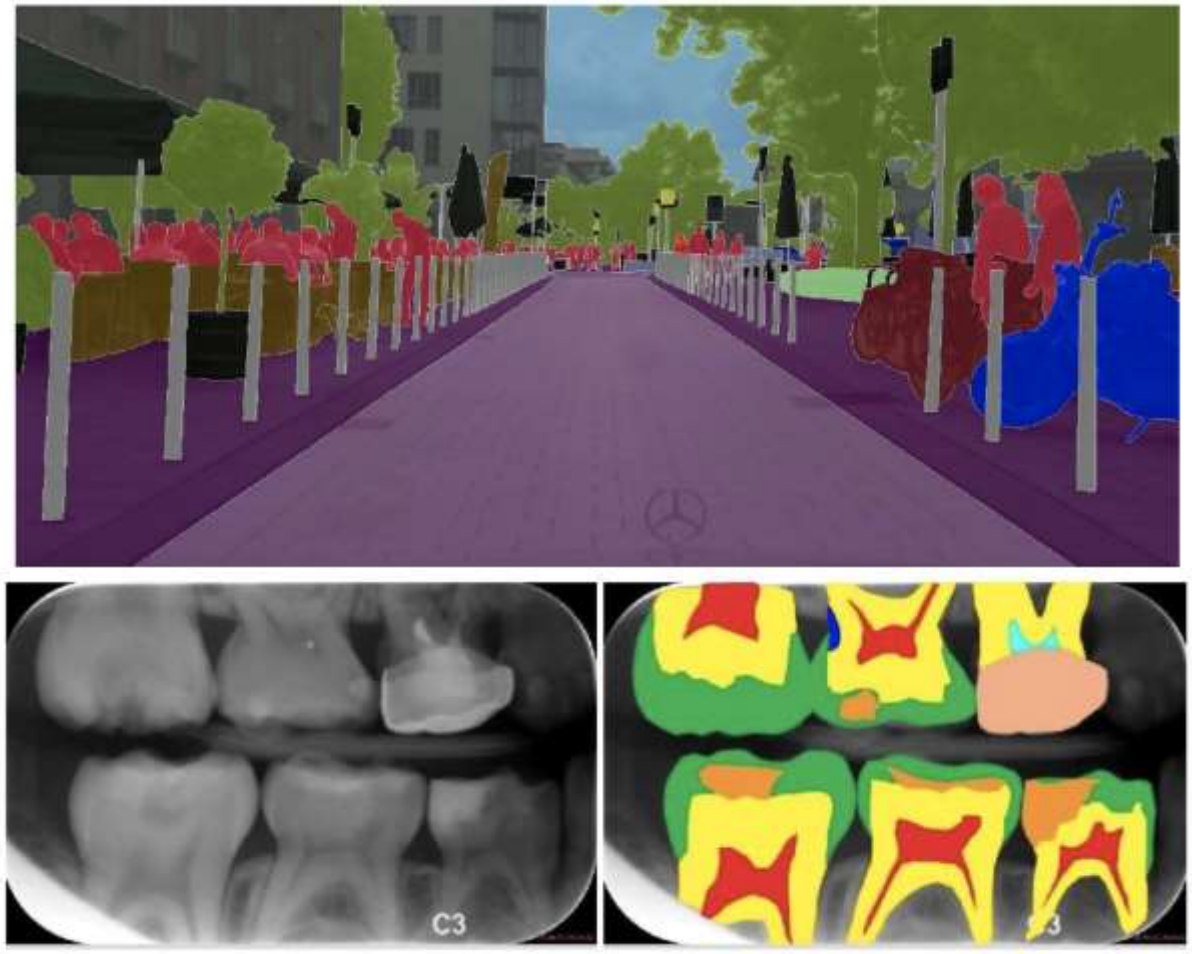


监督学习举例

监督学习举例：图像分类

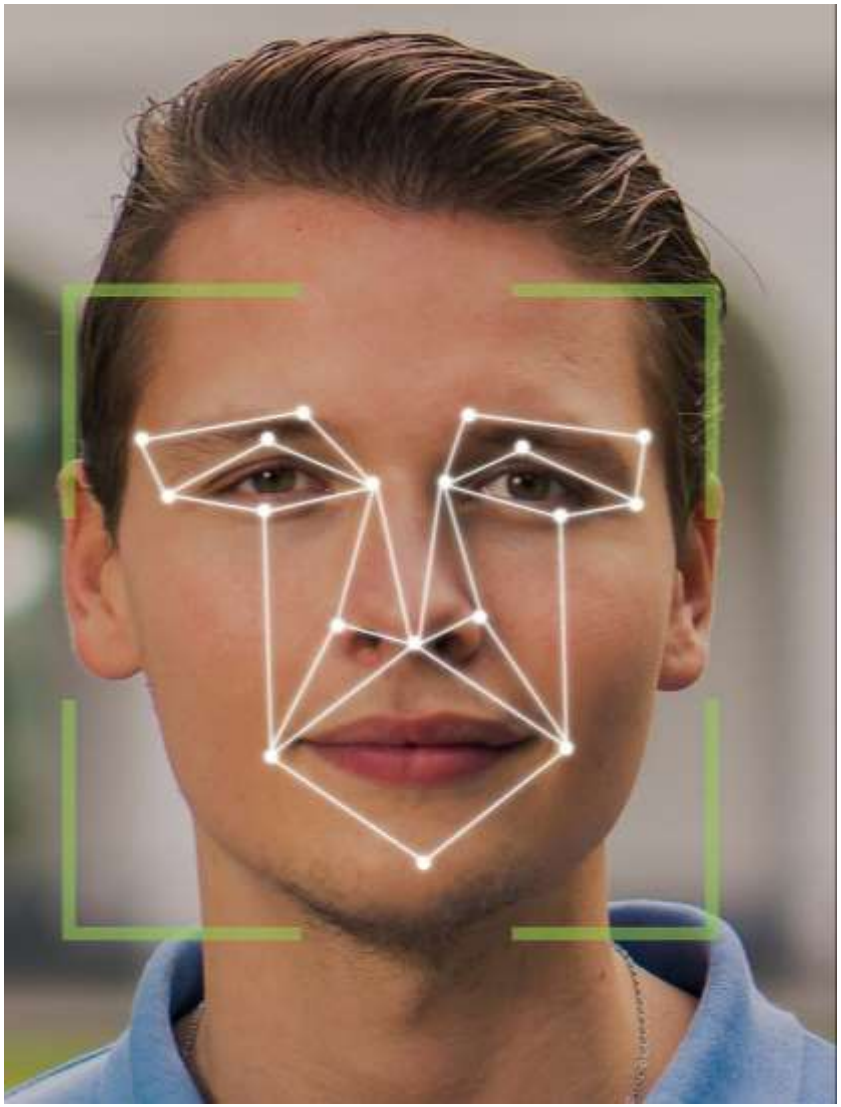
- 分类模型用于将输入数据分配到预定义的类别中
- 猫/狗的分类

监督学习举例：图像（语义）分割



图像语义分割：不仅将图像分割成不同的区域，还赋予这些区域具体的类别标签。语义分割需要理解图像中每个像素属于的类别。

监督学习举例：人脸识别



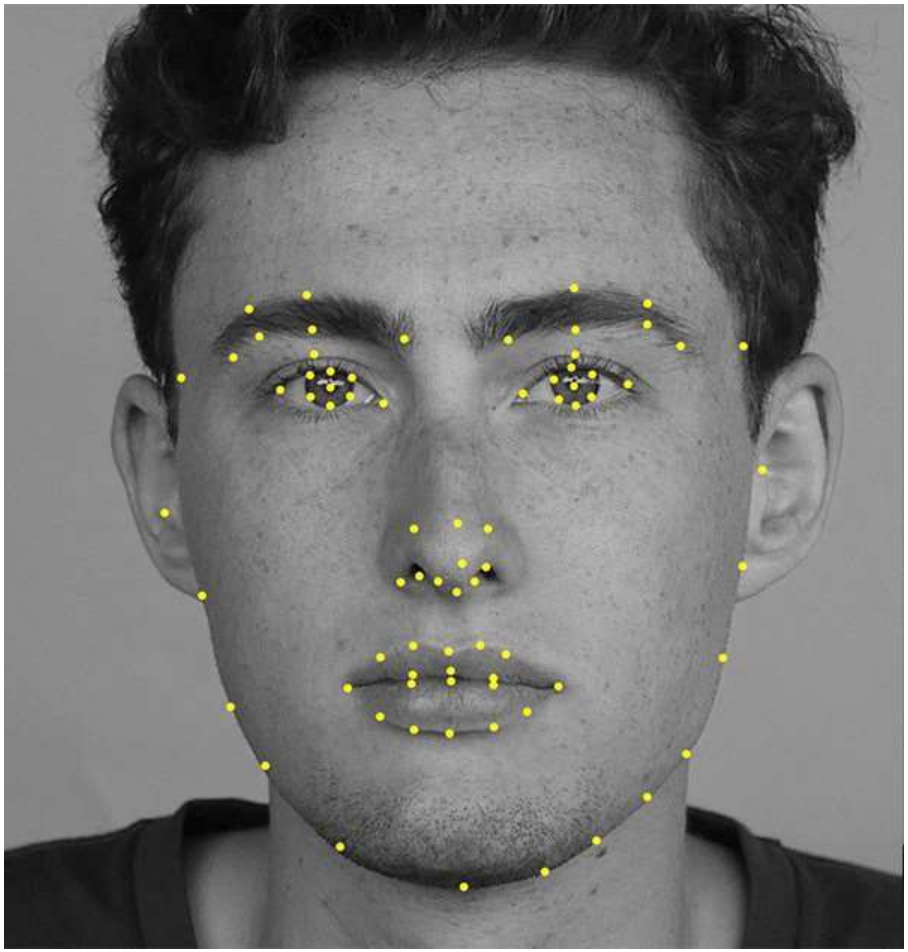
人脸识别是回归/分类？

人脸识别通常是**分类模型**的应用。在人脸识别任务中，模型的目标是确定输入的面部图像是否与已知个体的面部图像匹配，或者从一组已知个体中识别出特定个体。这涉及到将输入图像分配到一个或多个类别（即个体的身份）中，因此是分类任务的一种。

监督学习举例：姿态估计

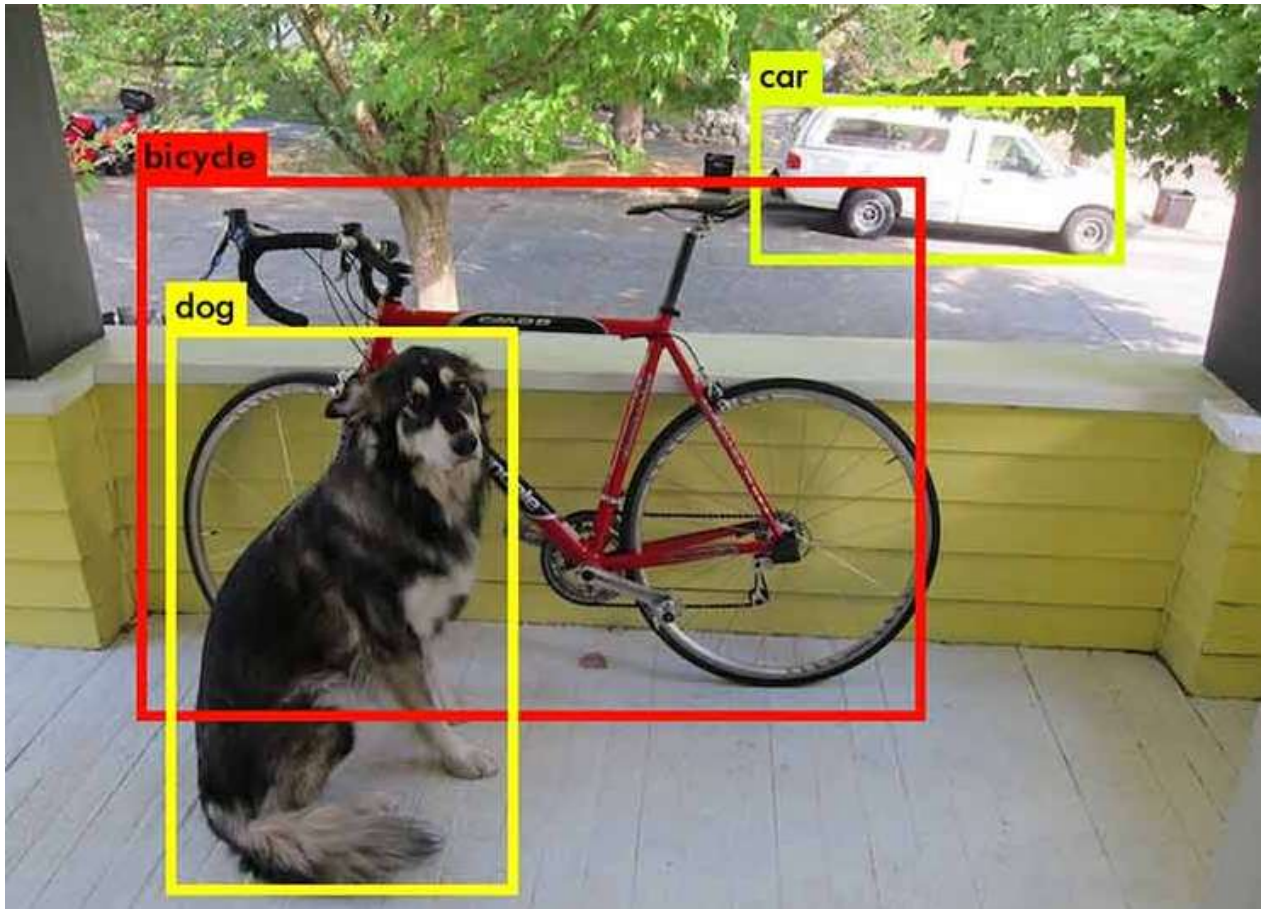
姿态估计属于关键点检测：用于识别并定位图像中特定兴趣点或部位的确切位置。关键点检测属于回归模型。

监督学习举例：人脸特征点检测



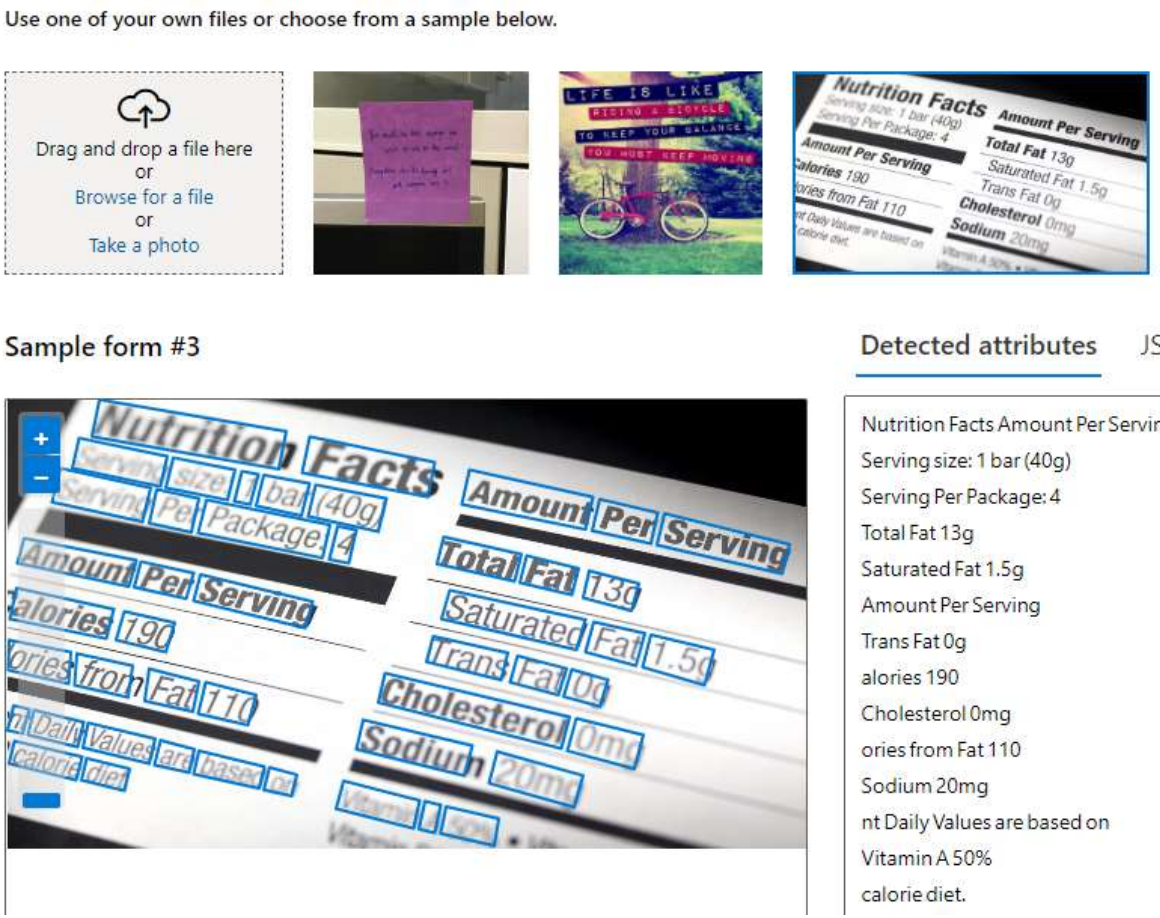
回归模型/分类模型？

监督学习举例：物体检测



物体检测任务是指在图像或视频中识别并定位一个或多个特定物体，并对每个检测到的物体进行分类和边界框定位。（分类/回归？）

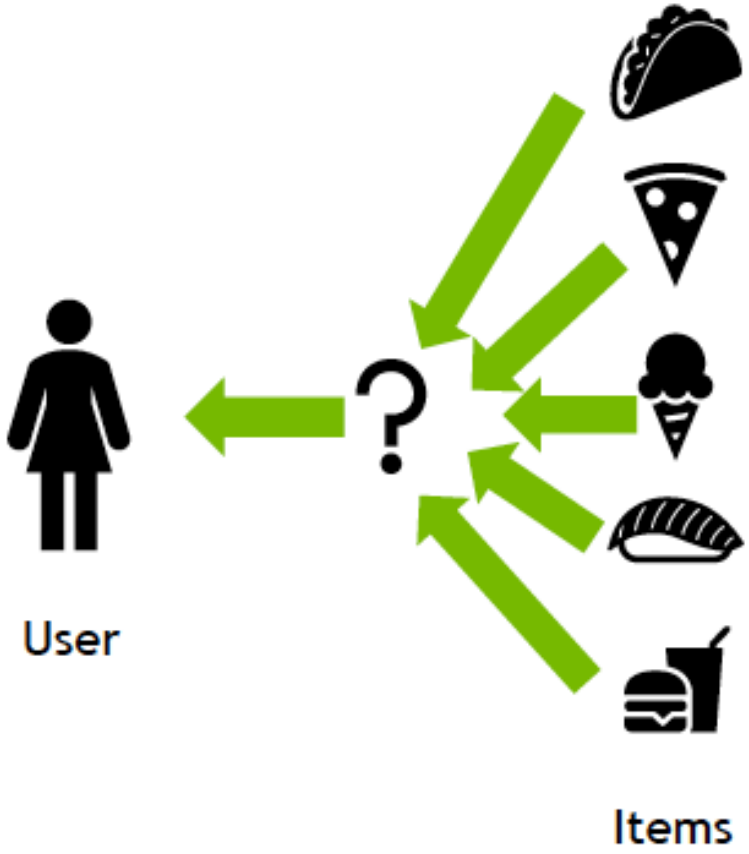
监督学习举例：光学字符识别（OCR）



在OCR中，目标是识别图像中的文字并将每个字符或单词分类为相应的字母、数字或符号。

监督学习的例子：推荐系统

- 根据用户的历史行为、兴趣或偏好，为其推荐个性化的商品、内容或服务。



监督学习的例子：推荐系统

- **推荐系统作为分类任务**

如果推荐系统的目标是**预测用户是否会对某个商品感兴趣**，这就属于一个二分类问题。例如：预测用户是否会点击某个商品，预测用户是否会购买某个商品，预测用户是否会观看某个视频。

在这种情况下，推荐系统的输出是一个类别（感兴趣/不感兴趣），而模型的任务是将用户-商品对分类为“正样本”（用户感兴趣）或“负样本”（用户不感兴趣）。

- **推荐系统作为回归任务**

如果推荐系统的目标是**预测用户对某个商品的评分或兴趣强度**，这就属于回归问题。例如：在电影推荐中，预测用户对某部电影的评分是1到5星。又例如计算用户对某商品的兴趣程度，用一个连续值来表示，可能介于0到1之间，表示购买或点击的概率。

在这种情况下，模型的输出是一个连续值，用来表示用户对商品的兴趣程度或预计的评分。

无监督学习的例子

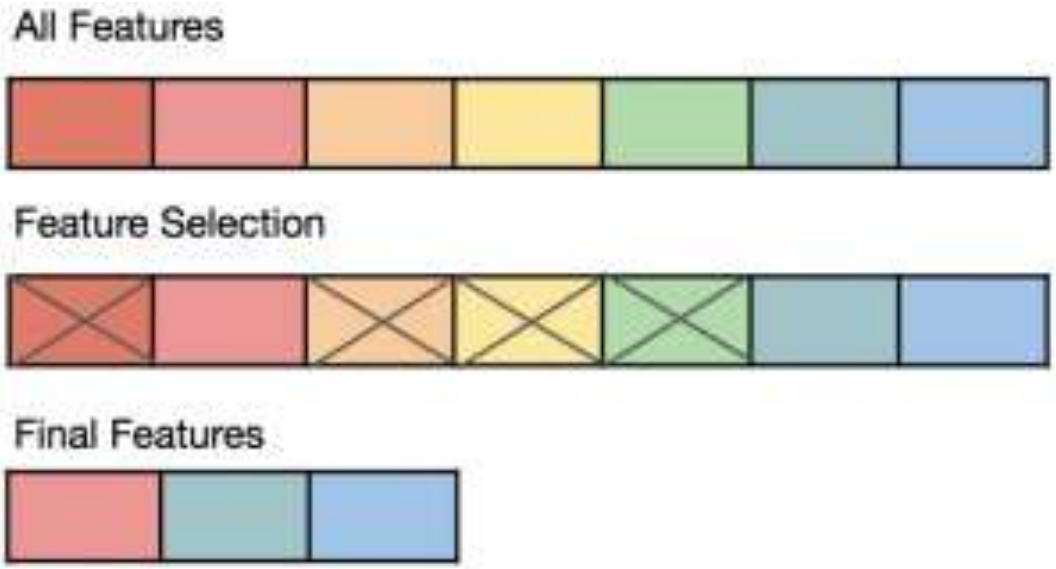
无监督学习的例子：客户细分

无监督学习被用于分析客户数据，发现具有相似行为或特征的客户群体。

“频繁小额消费群体”、“高价值客户群体”、“偶尔大额购物群体”

无监督学习的例子：维度缩减（降维）

在高维数据分析中，无监督学习常用于减少数据的维度，同时保留最重要的信息。这对于数据可视化和进一步的数据分析都很重要。



特征选择的主要目的是减少数据集中的特征数量，以改善模型的性能，减少计算成本，提高模型的可解释性。它通过去除不相关或冗余的特征来实现。

无监督学习的例子：维度缩减（降维）

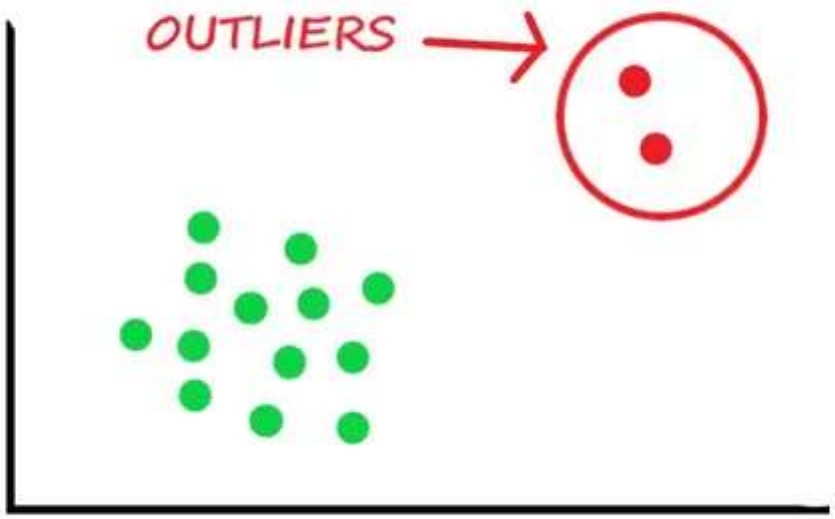
例：在图像处理或计算机视觉领域，图像通常具有非常高的维度。例如，一张分辨率为100×100的灰度图像包含10,000个像素，每个像素可以看作一个特征，因此整个图像的数据维度是10,000维。这会导致计算资源的高消耗，并且增加算法的复杂性和训练难度。为了高效处理图像数据，往往需要进行维度缩减。

应用：

- **图像分类**：在降维后的数据上进行图像分类任务（如识别数字手写体、识别猫狗等），降维可以提升分类模型的计算速度。
- **图像搜索**：在图像数据库中搜索相似图片时，可以在低维空间中快速找到相似的图片，提高检索效率。
- **可视化**：将图像数据从高维降到2D或3D空间，便于可视化，帮助我们了解不同图像之间的关系。

无监督学习的例子：异常检测

在多个行业中，如金融、网络安全或制造业，无监督学习用于识别异常行为或数据点。这对于预防欺诈、监测网络入侵或识别生产线上的缺陷至关重要。



无监督学习的例子：异常检测

例：在网络安全领域，入侵检测系统（IDS）用于识别网络中的异常行为，检测潜在的网络攻击。无监督学习可以用于构建入侵检测系统。

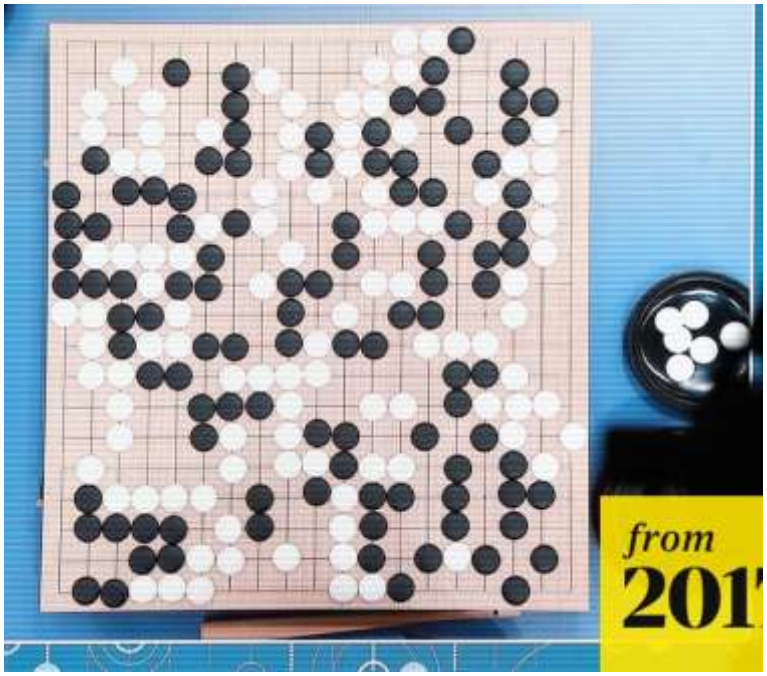
- **数据收集**：收集正常网络流量数据，包括流量大小、连接频率、端口和协议类型等。
- **模型训练**：使用无监督学习算法，比如K均值聚类或DBSCAN，训练模型以识别正常网络流量模式，建立网络流量的“正常”基线。
- **异常检测**：当新的网络流量数据与“正常”模式差异较大时，即被视为潜在的异常行为。例如，一个IP地址的访问频率突然增高或出现未见过的访问模式可能表明入侵企图。

强化学习的例子

强化学习的例子：游戏玩家

强化学习技术在许多游戏中获得了成功的应用，例如AlphaGo用于围棋，或OpenAI Five用于电子竞技游戏《Dota 2》。

OpenAI Five 是由 OpenAI 开发的人工智能程序，专门用于玩多人在线视频游戏《Dota 2》。该程序通过进行相当于超过10,000年时长的游戏对抗（强化学习）来进行学习和优化。



AlphaGo 是由 DeepMind 开发的人工智能程序，专门设计用于玩复杂的棋类游戏围棋。 AlphaGo利用强化学习技术，通过大量的游戏数据和自我对弈进行学习。

强化学习的例子：自动驾驶汽车

自动驾驶技术是强化学习的一个重要应用领域，因为它需要处理复杂的、动态变化的环境，并做出快速而安全的决策。

强化学习的例子：机器手

Google 使用深度学习和持续反馈教会机器人捡起不同形状的物体。

既可以是监督学习也可以是无监督学习的应用举例

生成模型

- 生成模型既可以是监督学习也可以是无监督学习

监督学习的生成模型：监督学习的生成模型是学习输入数据和输出标签之间的关系，从而能够生成新的数据点。例子：图像到图像的转换，文本到图像，图像到文本的生成，机器翻译。

无监督学习的生成模型：无监督学习中的生成模型的目标是学习原有数据的分布和结构，以便能够生成新的数据点，这些点与训练数据在统计特性上是一致的。典型的无监督生成模型包括生成对抗网络（GANs）和变分自编码器（VAEs）。它们通过学习训练数据的分布来生成新的数据实例，比如新图片或文本。

生成模型

对旧照片，旧电影进行修复

生成模型

图片上色并提高分辨率

生成模型

AI换脸

生成模型

- Midjourney
- Stable Diffusion
- …

图像生成

生成模型

- Pika Labs
- Synthesia
- DeepBrain AI
- …

视频生成

课堂思考

- 请分别举出一个结构化数据，非结构化数据的例子

- 请说出下边几种情况分别属于什么学习方式？（备选：无监督/监督/强化学习）

  -给狗和猫进行分类
  -给猫和狗进行聚类
  -让机器人以试错的方式学习抓取物体
  -大数据降维可视化
  -人脸识别
  -物体检测

- 请用图的方式解释人工智能，机器学习，深度学习之间的关系

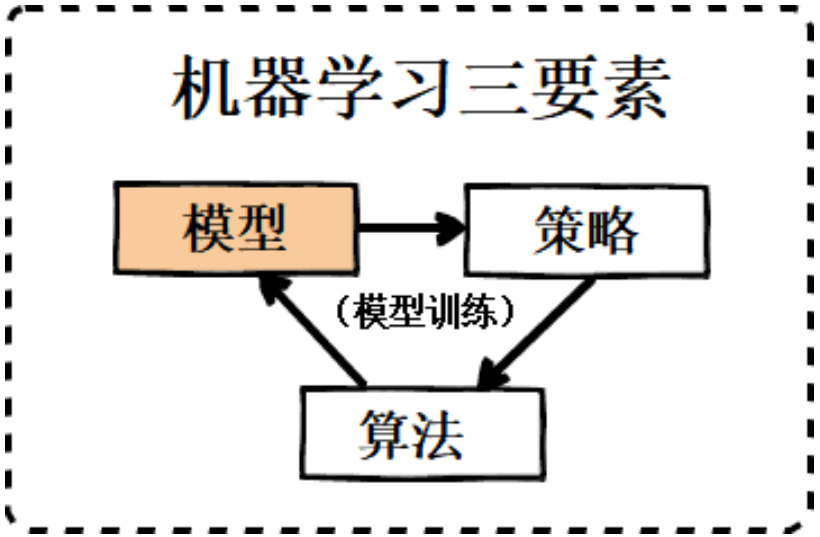
学习算法、线性回归

学习算法三要素

1.模型（Model）：模型指的是数据和预测之间的数学关系。模型中有参数，不同的参数组合构成了不同的模型，这使得可能的模型数量有多个甚至是无限多。这些不同的模型构成了所谓的“假设空间”。这意味着，假设空间包含了所有可能的模型，每个模型都是对现实问题的一种潜在解释。

2.策略（Strategy）：策略定义了一个“好”的模型的标准。这个标准一般是用损失函数来描述的，损失函数衡量的是模型预测值与实际值之间的差异。一个“好”的模型是指在其损失函数上表现最佳的模型，即错误最小。

3.算法（Algorithm）：算法是指在假设空间中找到最优模型的方法。这涉及到求解一系列数学问题，以确定模型参数的最佳值。



机器学习算法举例

问题阐述：想通过某个房产所在的位置（变量x_1），物业（变量x_2），楼层（变量x_3），是否精装修（变量x_4）来预估这个房产的价值（y）。

模型：基于房产价值与其属性之间可能存在的线性关系的假设，我们将构建一个线性回归模型来预测房产价值，即$y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4$，其中$w_0$，$w_1$，$w_2$，$w_3$，$w_4$是模型参数。

策略：在这个例子中，一个“好”的模型的标准是预测房价与实际房价的差异越小越好。于是策略就是找到一组参数，它能够最小化预测房价与实际房价之间的差异$\sum_{i=1}^{n}(Y^{(i)} - y^{(i)})^2$。

算法：为了找到这一组参数，我们可以使用解析方法（通过精确的数学演算找到问题的精确解的方法）或者数值方法（数值方法通过近似和迭代来求解问题，依赖于计算机进行大量计算，以逼近问题的解）来找到使得差异最小的参数。

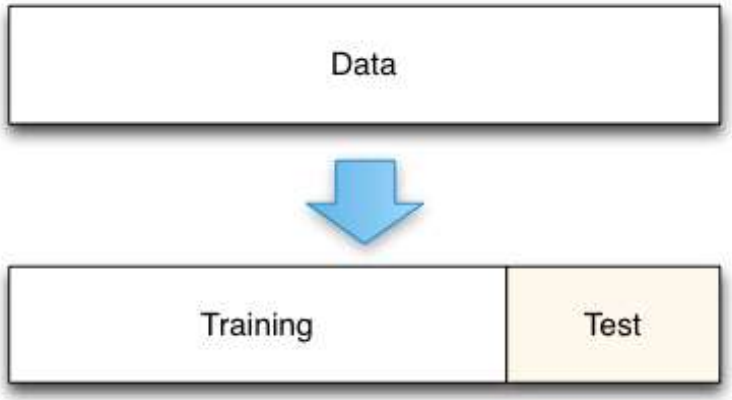
训练集/测试集

- 数据集主要被分为两个部分：训练集、测试集。

1.训练集用于训练模型。在这个阶段，模型学习识别数据中的模式和关系。通过调整模型参数，模型尝试最大限度地减少误差，并在这些已知数据上表现出良好的拟合度。

2.测试集用来评估模型的**泛化能力**。泛化是指模型对新、未见过的数据的处理能力。测试集是**独立于**训练过程的，因此提供了对模型性能的公正评估。

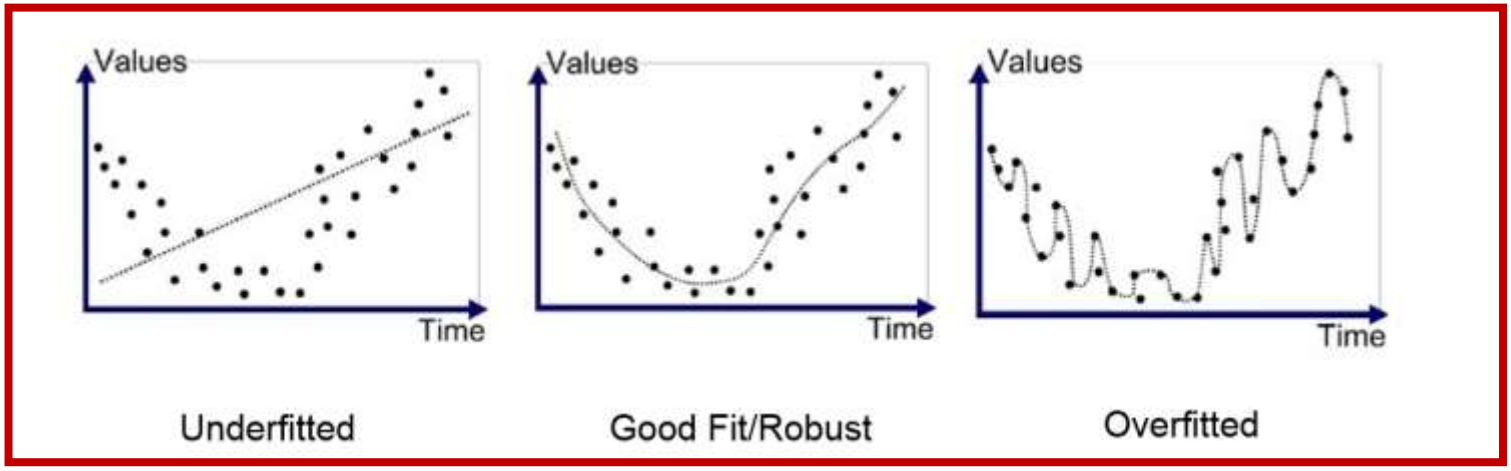
如果一个模型在训练集上表现良好但在测试集上表现不佳，则可能表明模型过拟合了训练数据，未能有效地泛化（到其他数据集上）。



过拟合/欠拟合

我们不仅希望模型能够在训练集上获得较低的误差，还期望模型能够在测试集上也拥有较低的误差（即具有良好的泛化能力）。**测试集上的误差，被称为泛化误差**，是评估模型泛化能力的重要指标。模型泛化能力不足时，通常表现为**过拟合（overfitting）**或**欠拟合（underfitting）**的问题。

- **过拟合**发生在模型对训练数据学习得太过彻底，以至于它开始捕捉到数据中的噪声和偶然的模式，而非真正的趋势。
- **欠拟合**则是发生在模型未能充分学习数据中的趋势和模式，导致在训练集和测试集上都表现不佳。
- 为什么会发生过拟合（**模型太过复杂**），欠拟合（**模型太过简单**）。



特征向量/空间/矩阵

1.特征向量：在机器学习中，每个数据点通常由一系列特征（或属性）表示。这些特征被组合成一个向量，称为特征向量。

2.特征矩阵：当我们将数据集中的所有样本的特征向量汇集起来，形成的矩阵就是特征矩阵。在这个矩阵中，每一行代表一个样本的特征向量，而每一列代表一个特定的特征。

3.特征空间：特征空间是一个抽象的概念，它指的是所有可能的特征向量所构成的空间。在这个空间中，每个维度代表一个特征。这个空间中都有一个唯一的位置，由其特征向量确定。

学习数据的模式和关系之前**通常都需要将数据表示为特征向量和特征矩阵的形式**，这样机器学习算法才可以有效地处理和分析数据

*在使用机器学习算法前，一般会将数据集转化为特征矩阵的形式

特征空间/特征矩阵

以房价预测的例子举例：

1. **特征向量：**每个房产都由一组特征表示：位置（x_1），所在小区（x_2），楼层高度（x_3），以及是否精装修（x_4）每个房产的这四个特征共同构成了一个特征向量（x_1，x_2，x_3，x_4）。
2. **特征矩阵：**收集了多个房产的数据后，可以将每个房产的特征向量作为特征矩阵的一行。在这个矩阵中，每一行代表一个房产，每一列代表一个特定的特征（如位置、小区等）。
3. **特征空间：**特征向量（x_1，x_2，x_3，x_4）所在的空间就是特征空间。在这个空间中，每个维度对应一个特征。在这个例子中，特征空间是一个四维空间，其中每一个维度代表一个房产特征。每个房产样本在特征空间中占据一个唯一的点，这个点由其特征向量确定。

Features（特征）

Samples（样本）

$\boldsymbol{x_1^{(1)}}$	$\boldsymbol{x_2^{(1)}}$	$\boldsymbol{x_3^{(1)}}$	$\boldsymbol{x_4^{(1)}}$
$\boldsymbol{x_1^{(2)}}$	$\boldsymbol{x_2^{(2)}}$	$\boldsymbol{x_3^{(2)}}$	$\boldsymbol{x_4^{(2)}}$
$\boldsymbol{x_1^{(3)}}$	$\boldsymbol{x_2^{(3)}}$	$\boldsymbol{x_3^{(3)}}$	$\boldsymbol{x_4^{(3)}}$
…	…	…	…
$\boldsymbol{x_1^{(n)}}$	$\boldsymbol{x_2^{(n)}}$	$\boldsymbol{x_3^{(n)}}$	$\boldsymbol{x_4^{(n)}}$

线性回归

线性模型

（1）线性模型通过属性的加权线性组合预测变量，即

$$f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + b \ ,$$

其中 $\boldsymbol{x}$ 是特征向量，$w_1, w_2, \ldots, w_d$等都是特征权重，b 是截距项。

（2）在矩阵符号中，上述模型可表示为

$$f(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x} + b \ ,$$

其中 $\boldsymbol{w}$是权重向量，x 是特征向量，T表示转置。

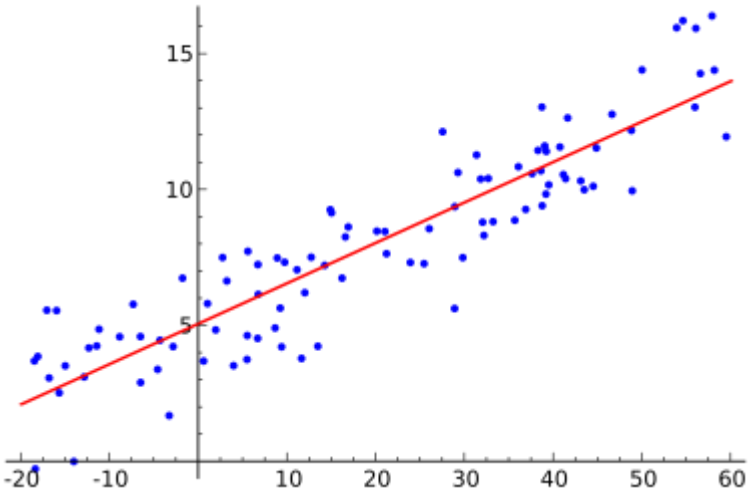
特点

- 线性模型以其简洁的形式而广受欢迎。尽管模型结构基础，但它是多种非线性模型（比如逻辑回归，多项式回归）的理论基础。
- 由于 w 直观地表达了各属性在预测中的重要性，因此线性模型有很好的可解释性。例：$f_{\text{体感温度}}(x) = 0.1 x_{\text{风速}} + 0.4\, x_{\text{温度}} + 0.5\, x_{\text{湿度}} + \mathrm{b}$

线性回归模型

回归：根据输入特征所取的值，模型产生一个连续的预测值作为输出。

线性回归（Linear Regression）是一种预测技术，它通过建立输入与输出之间的线性关系模型来预测实际数值结果。



蓝点：数据真实值
红线上的点：预测模型给出的数据预测值

通常用于预测模型：

- 通常用于预测如产品销量、股票价格、行业趋势等连续变量
- 通过房产的位置，面积大小，楼层数等预测房产的价格
- ……

线性回归模型的**策略**

- 如何定义一个足够“好”的线性回归模型呢?

  一个好的线性回归模型，它的预测值和真实值一定比较接近。**均方误差（MSE）**可以反映预测值和真实值的差异，故可以用**均方误差**作为衡量线性回归模型“优/劣”的标准。其定义为

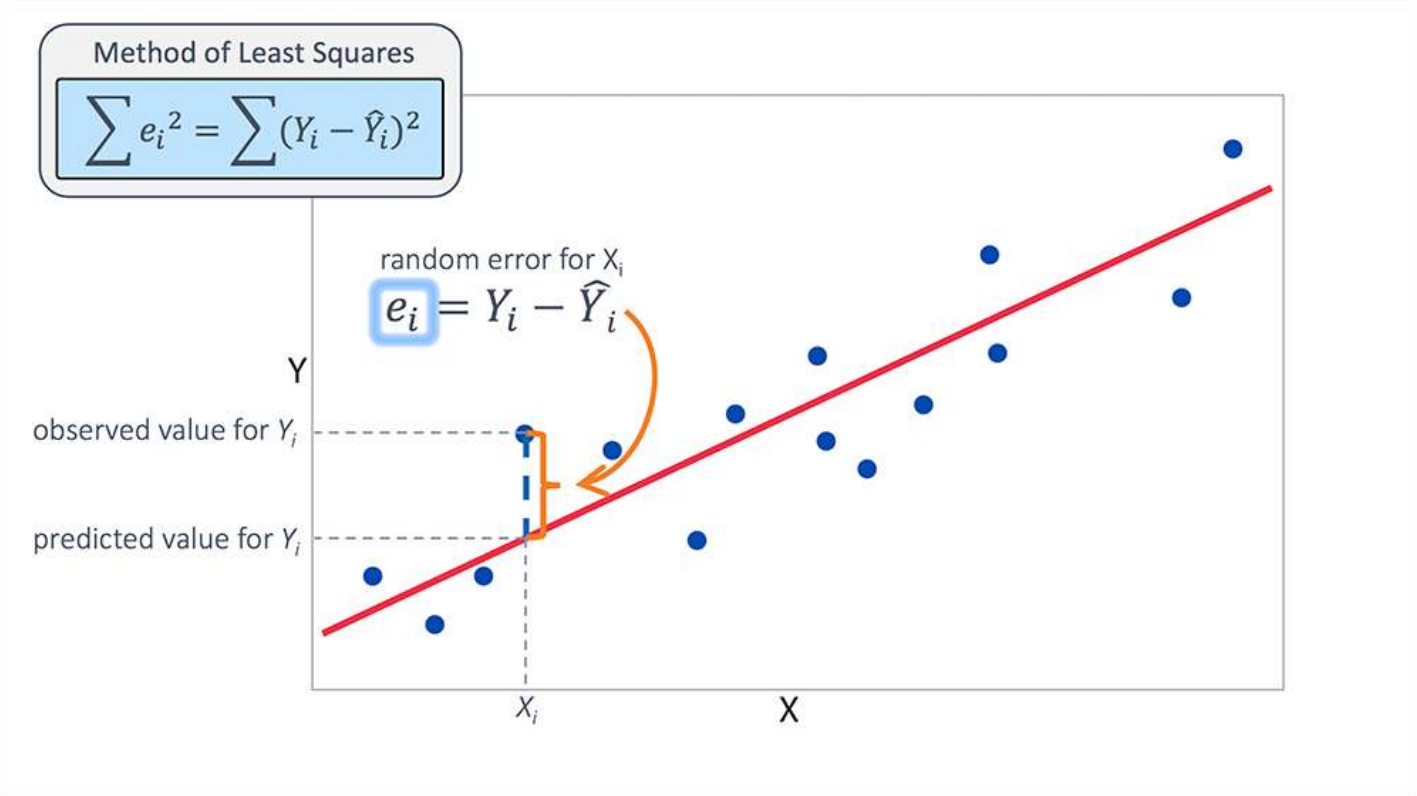
$$\frac{1}{n}\sum_{i=1}^{n}[f(x_i)-y_i]^2$$

其中，n 是样本数量，$f(x_i)$ 是模型对第 i 个样本的预测值，y_i 是对应的实际值。

- **均方误差**的大小反映了模型预测的准确性，误差越小，模型的预测性能越好。

线性回归模型的**策略**

- “均方误差最小化”有实际的几何意义：寻找一条直线，使得数据集中每个点在 y 轴方向上的垂直距离的平方和最小。



线性回归模型的**算法**

均方误差函数是 **凸函数**，因此通过求导得到的极小值点就是 **全局最小值点**。

线性回归模型：$Y = \alpha + \beta X + \varepsilon$ （模型的参数是？）

均方误差表达式为：$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$

目标是找出使得该误差最小的α和β

解析法

通过分别对 α 和 β 求导并设导数为零来求解模型参数，得到正规方程组：

$$\begin{cases} n\,\alpha + \sum_{i=1}^{n} x_i\,\beta = \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\,\alpha + \sum_{i=1}^{n} x_i^2\,\beta = \sum_{i=1}^{n} x_i y_i \end{cases}$$

线性回归**算法**

解析法

$$\begin{cases} n\,\alpha + \sum_{i=1}^{n} x_i\,\beta = \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\,\alpha + \sum_{i=1}^{n} x_i^2\,\beta = \sum_{i=1}^{n} x_i y_i \end{cases}$$

根据克莱姆法则，有

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \bar{y} - \bar{x}\hat{\beta}$$

多元线性回归**模型**

解析法

$$f(\boldsymbol{x}) = w_1 x_1 + w_2 x_2 + \ldots + w_d x_d + b \ ,$$

如果我们把数据集D写作

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1d} & 1 \\ x_{21} & x_{22} & \ldots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \ldots & x_{md} & 1 \end{pmatrix}$$

把w和b吸收入向量形式$\hat{w} = (w_1, w_2, \ldots, w_d, b)^T$

于是多元线性回归的
一般形式为:

$$f(\boldsymbol{x}) = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1d} & 1 \\ x_{21} & x_{22} & \ldots & x_{2d} & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \ldots & x_{md} & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \\ b \end{pmatrix} = \mathrm{X}\hat{w}$$

多元线性回归**策略**

解析法

真值　　预测值

$$\hat{\boldsymbol{w}}^* = \arg\min_{\hat{\boldsymbol{w}}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{w}})$$

残差向量

残差向量的内积（均方误差）

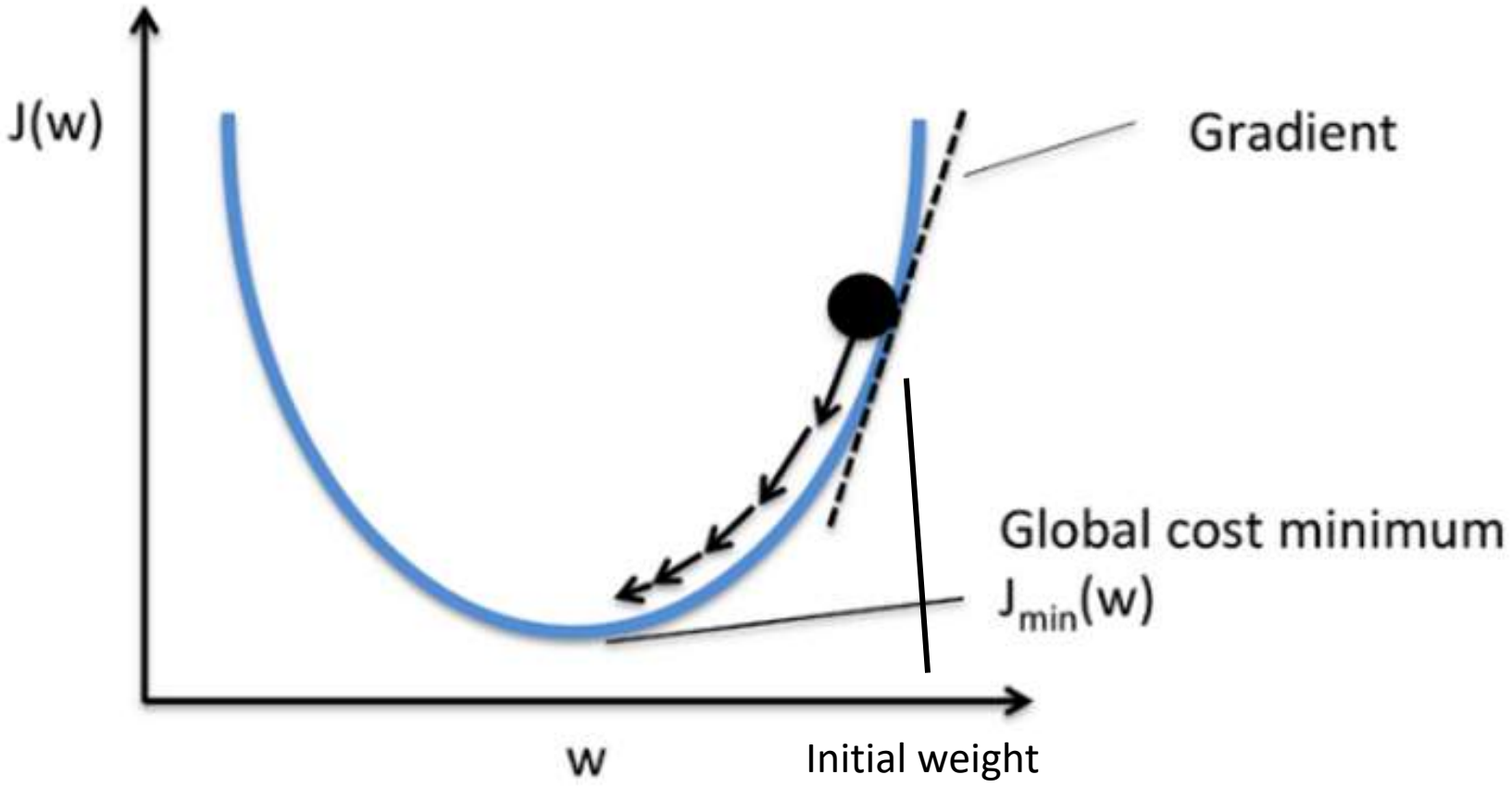
最小化均方误差

最后得到 $\hat{\boldsymbol{w}}^* = \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}$

线性回归**算法**

数值解法

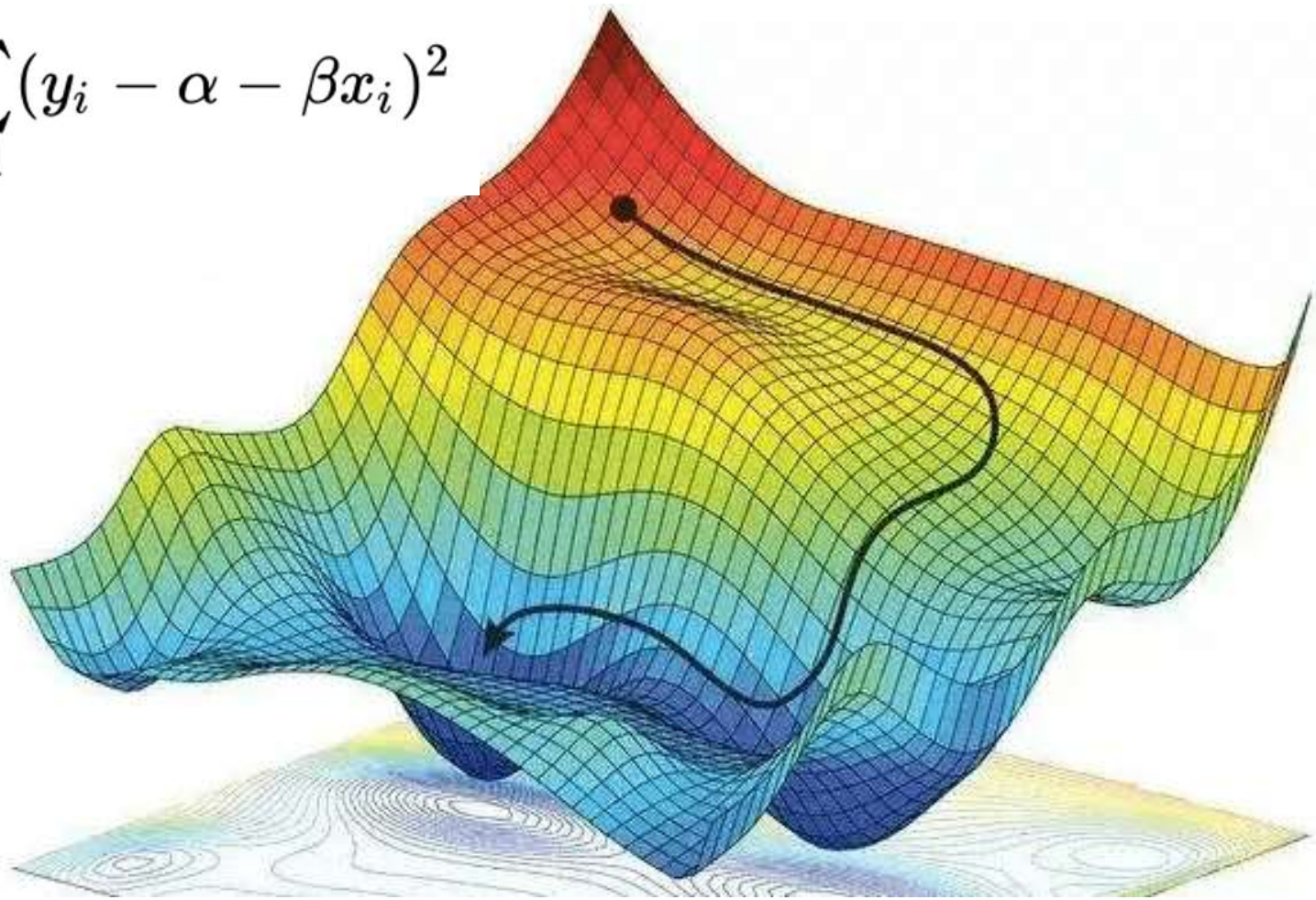
梯度下降



线性回归**算法**

数值解法　梯度下降

$$\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$$



Interactive Demo：https://blog.skz.dev/gradient-descent

线性回归的求解例

有5个数据点(1, 2), (2, 3), (3, 7), (4, 8), (5, 9)，请利用如下公式计算这些点的线性回归方程

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

提示：先计算如下表格

x	y	xy	x^2

线性回归的求解例

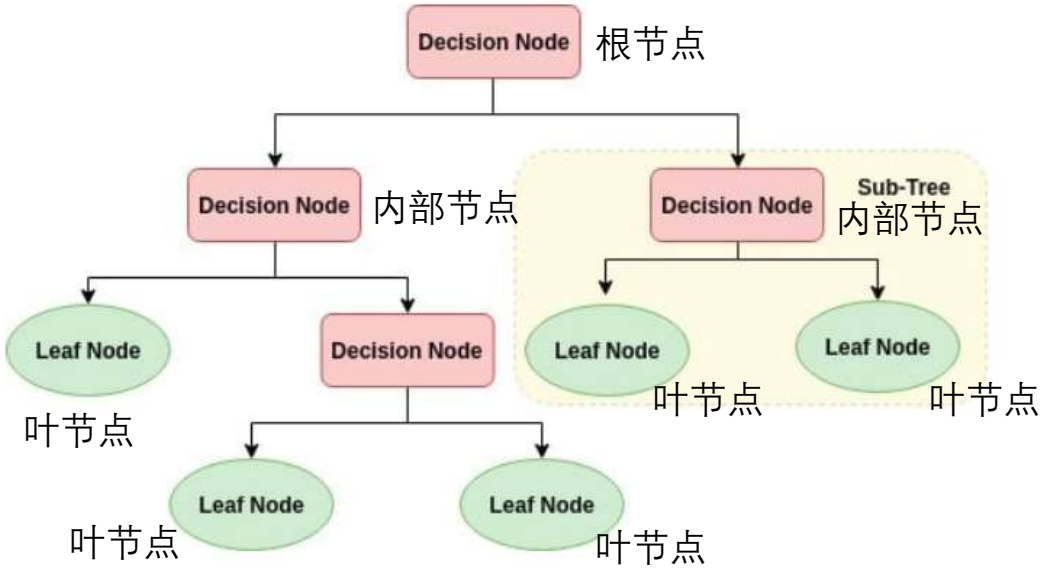
通过5个数据点(1, 2), (2, 3), (3, 7), (4, 8), (5, 9)，得到的线性回归方程是

$$y = 0.1 + 1.9\ x$$

决策树

监督学习：决策树

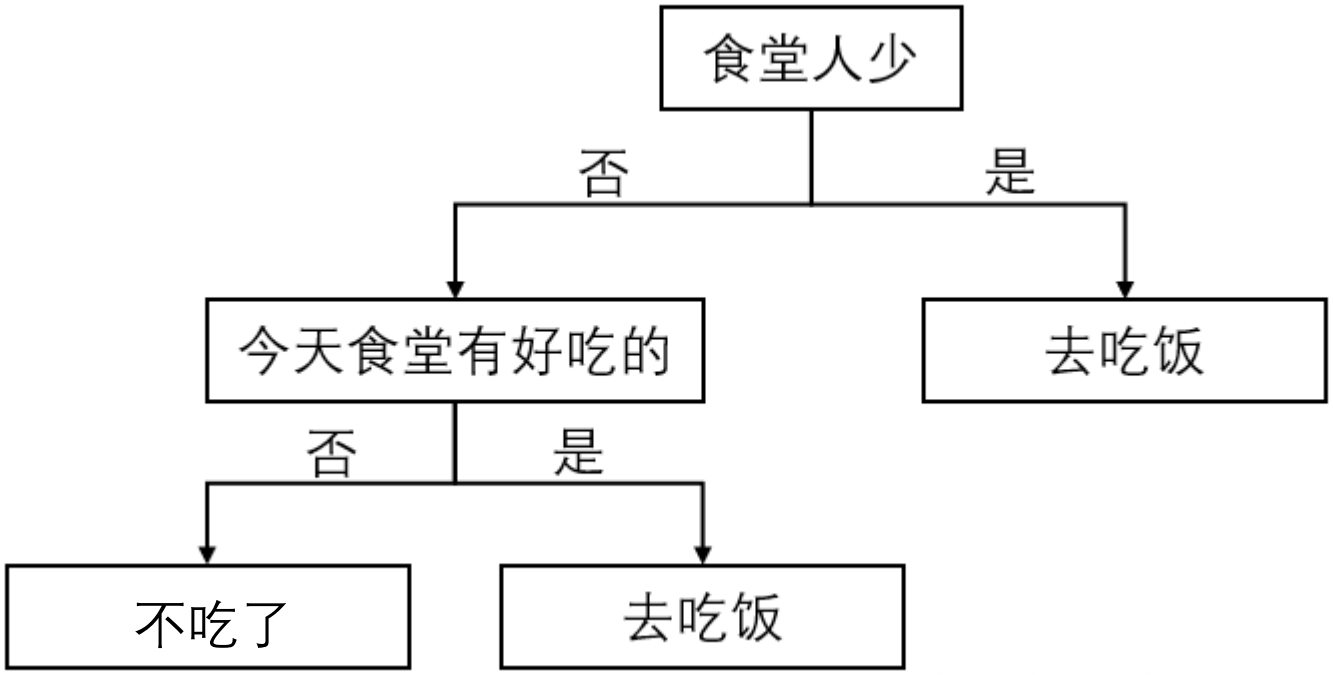
- 决策树是一种监督学习算法（1986年），既适用于分类，也适用于回归任务。
- 它呈现为一个树状结构，包括一个**根节点**、若干**内部节点**和多个**叶节点**。根节点一般代表整个数据集，内部节点代表数据的特征判断，叶节点则对应于最终的输出类别或数值结果。
- 在分类问题中，每个叶节点代表一个类别标签；在回归问题中，叶节点代表一个连续值。



- 常用的决策树有ID3，C4.5和CART（Classification And Regression Tree）

监督学习：决策树

- 决策树作为一种决策制定的策略，在日常生活中也被经常用到



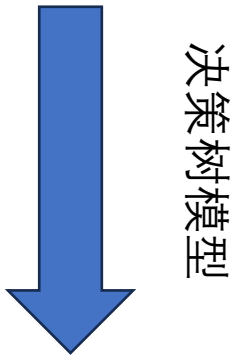
监督学习：决策树

例

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

根据天气状况（outlook）、湿度（humidity）和风力（wind）这三个因素来判断当天是否会打网球。

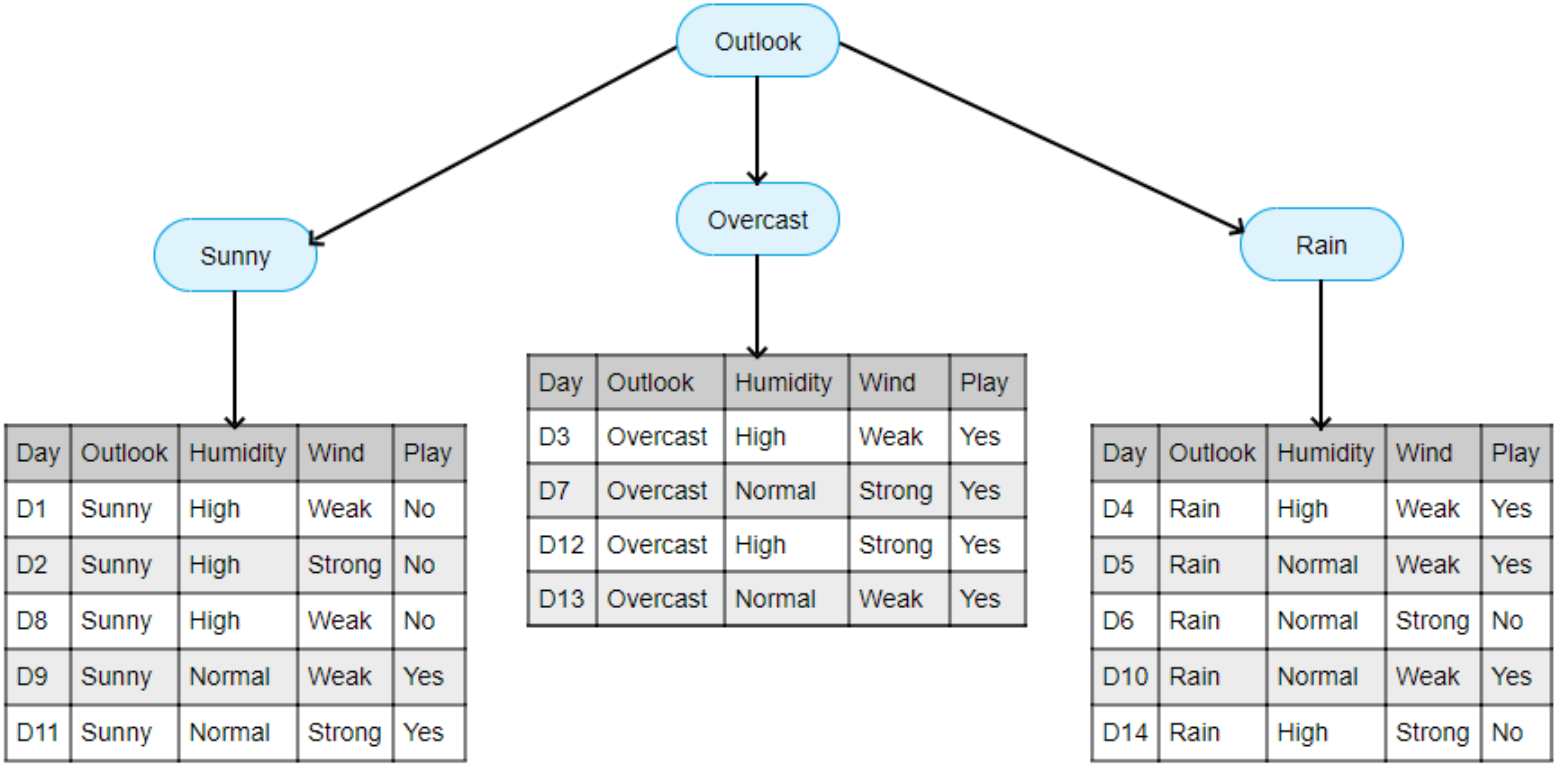
输入：天气状况（outlook）、湿度（humidity）和风力（wind）



输出：是否打网球

监督学习：决策树

在决策树中，“pure node”（纯净节点）指的是一个节点（通常是叶节点）当中，所有的数据点都属于同一个类别。



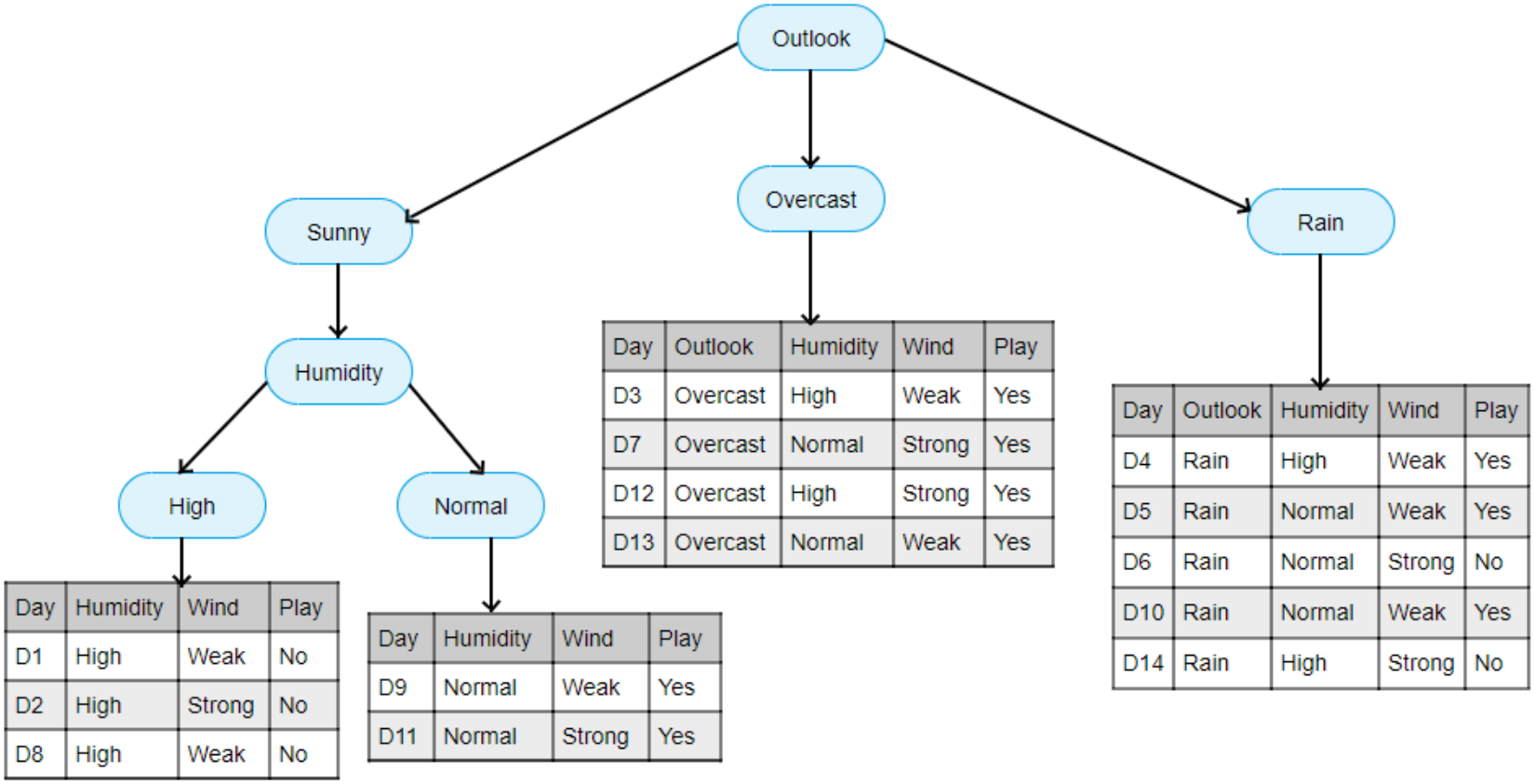
Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes

Day	Outlook	Humidity	Wind	Play
D3	Overcast	High	Weak	Yes
D7	Overcast	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes

Day	Outlook	Humidity	Wind	Play
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D10	Rain	Normal	Weak	Yes
D14	Rain	High	Strong	No

- 多云（Overcast）的情况下，小明肯定会打网球 （pure node）
- 晴天（Sunny）或者雨天(Rain)的情况下，不确定是否会打网球（impure node）

监督学习：决策树



Day	Humidity	Wind	Play
D1	High	Weak	No
D2	High	Strong	No
D8	High	Weak	No

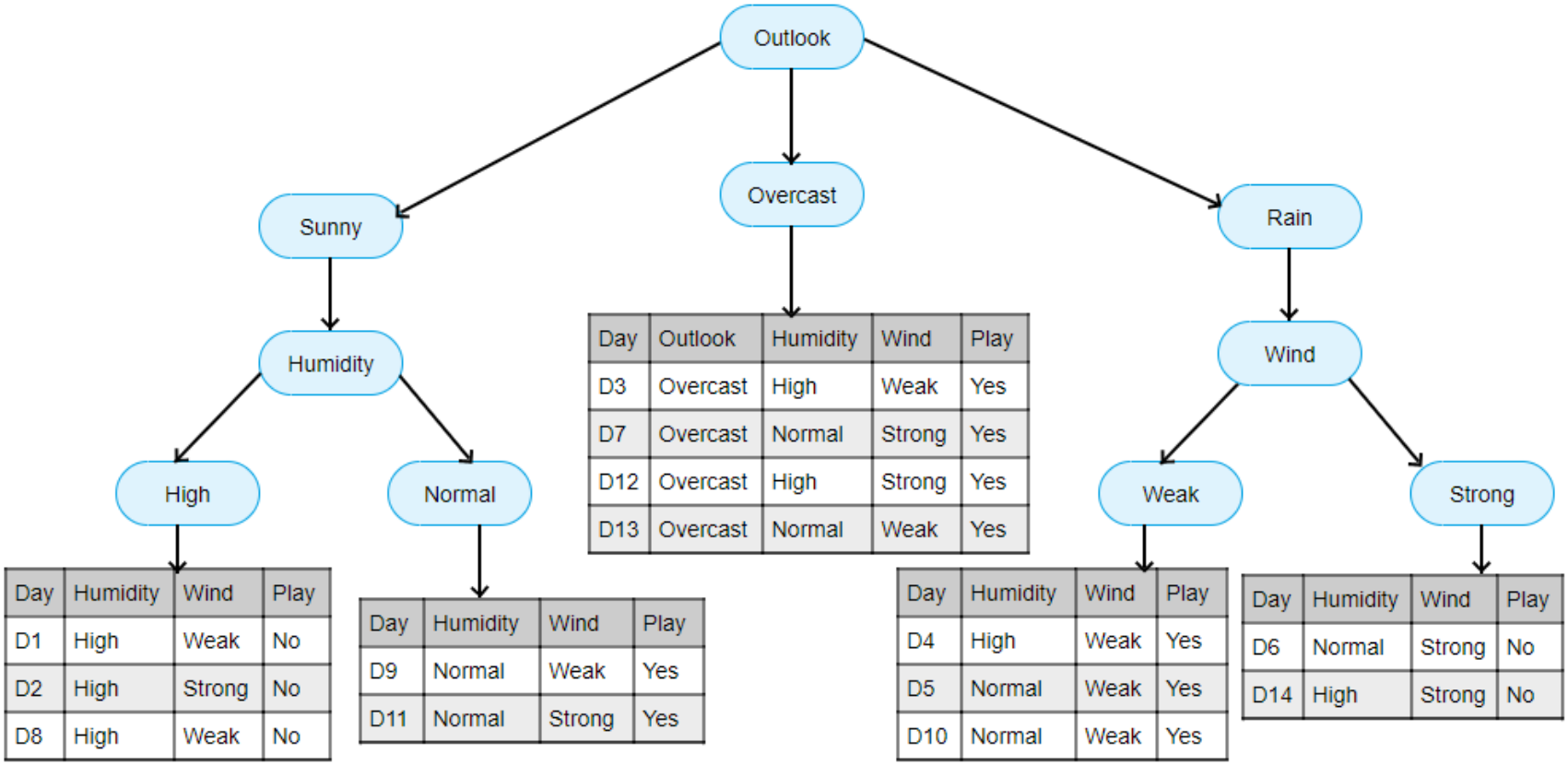
Day	Humidity	Wind	Play
D9	Normal	Weak	Yes
D11	Normal	Strong	Yes

Day	Outlook	Humidity	Wind	Play
D3	Overcast	High	Weak	Yes
D7	Overcast	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes

Day	Outlook	Humidity	Wind	Play
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D10	Rain	Normal	Weak	Yes
D14	Rain	High	Strong	No

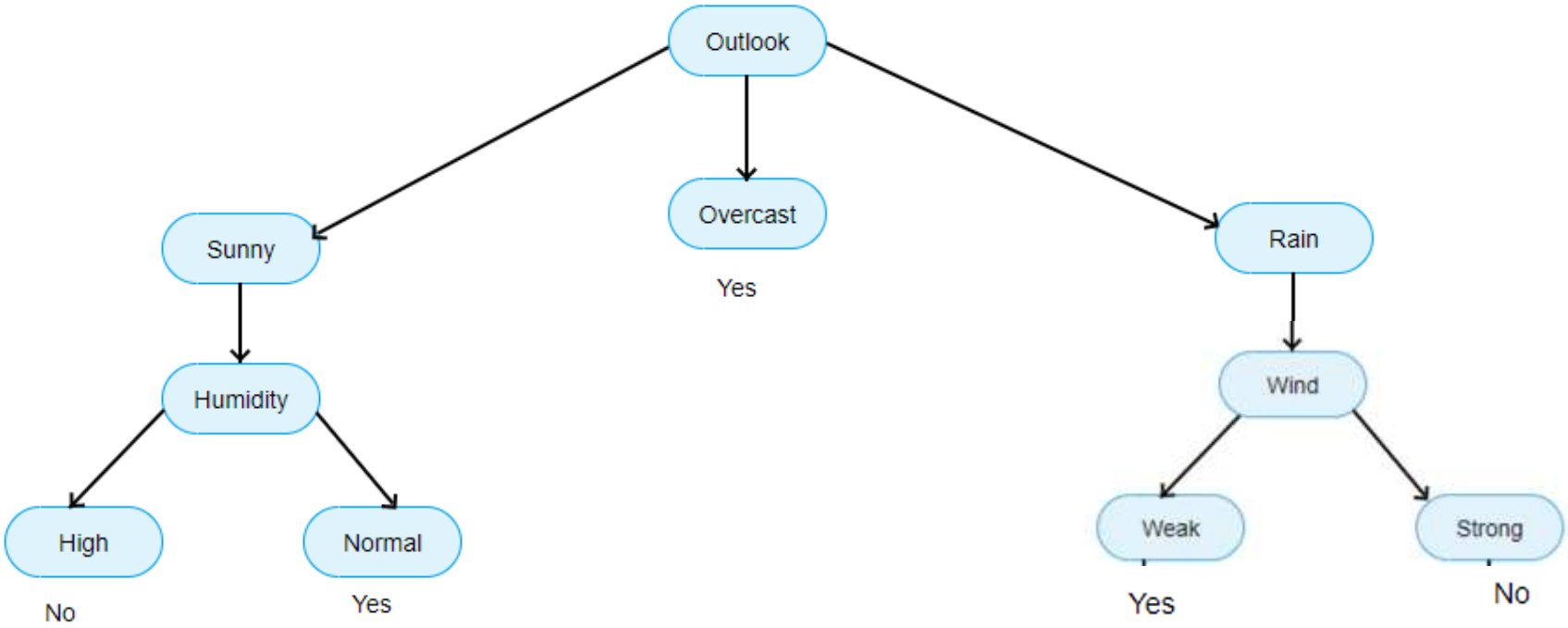
- 晴天（Sunny）这个类别下，对湿度（Humidity）进行分类可以得到两个pure node

监督学习：决策树



- 雨天情况下，对风的强度进行分类可以得到两个pure node

监督学习：决策树



- 最终决定小明是否会玩网球的模型

特征划分选择

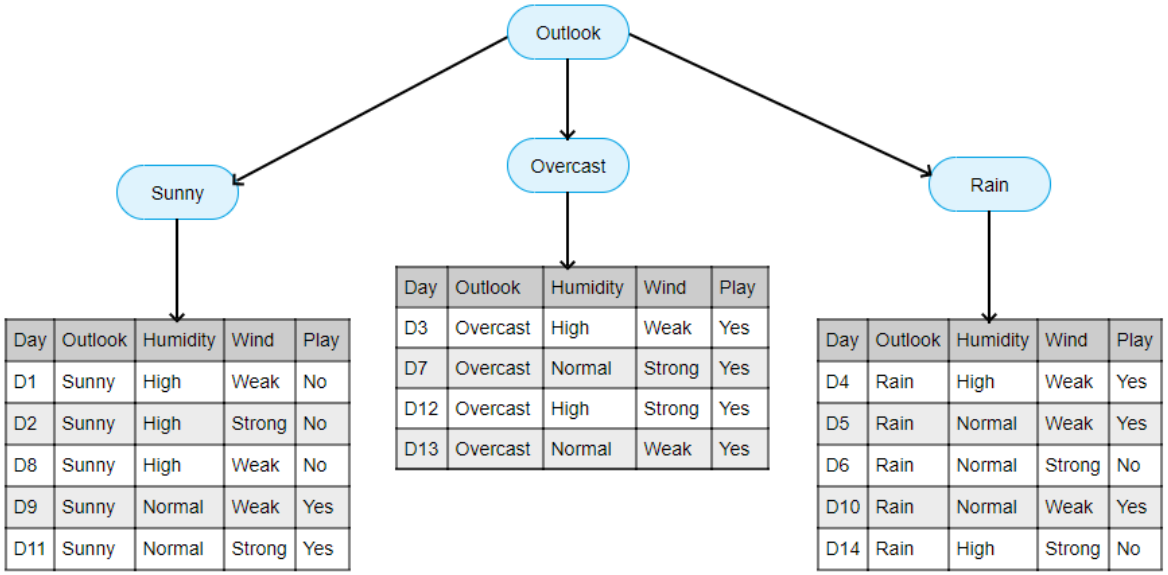
- 为了构建一个决策树，我们需要**确定哪个特征最适合作为根节点**。然后，我们需要为树的每个分支**选择适当的特征作为内部节点**，以便进一步细分数据集直至达到纯净节点（叶节点）。
- 我们可以用**信息熵**作为划分标准。**信息熵**是衡量数据无序程度的指标。初始时数据可能混杂，**熵值**较高。决策树通过其结构对数据进行分类，以达到更有序和一致的状态（pure node），这个时候数据的熵值就会很低。

计算某一数据子集的熵 $\mathrm{H}(T) = \mathrm{I}_E(p_1, p_2, \ldots, p_J) = -\sum_{i=1}^{J} p_i \log_2 p_i$

P_i代表着在一数据子集中，第i个数据类别出现的概率，它们的和等于1

特征划分选择

熵 $$\mathrm{H}(T) = \mathrm{I}_E(p_1, p_2, \ldots, p_J) = -\sum_{i=1}^{J} p_i \log_2 p_i$$



Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes

Day	Outlook	Humidity	Wind	Play
D3	Overcast	High	Weak	Yes
D7	Overcast	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes

Day	Outlook	Humidity	Wind	Play
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D10	Rain	Normal	Weak	Yes
D14	Rain	High	Strong	No

Sunny分支子集的熵: $$I_E([3,2]) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.9709508$$

Overcast分支子集的熵: $$I_E([0,4]) = -\frac{0}{4}log_2\frac{0}{4} - \frac{4}{4}log_2\frac{4}{4} = 0$$

Rain分支子集的熵: $$I_E([3,2]) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.9709508$$

特征划分选择

当Humidity作为根节点时：

High分支的熵：$I_E([3,4]) = -\frac{3}{7}log_2\frac{3}{7} - \frac{4}{7}log_2\frac{4}{7} = 0.98533$

Normal分支的熵：$I_E([1,6]) = -\frac{1}{7}log_2\frac{1}{7} - \frac{6}{7}log_2\frac{6}{7} = 0.59159$

当Wind作为根节点时：

Weak分支的熵：

Strong分支的熵：

特征划分选择

当Humidity作为根节点时：

High分支的熵：$I_E([3,4]) = -\frac{3}{7}log_2\frac{3}{7} - \frac{4}{7}log_2\frac{4}{7} = 0.98533$

Normal分支的熵：$I_E([1,6]) = -\frac{1}{7}log_2\frac{1}{7} - \frac{6}{7}log_2\frac{6}{7} = 0.59159$

当Wind作为根节点时：

Weak分支的熵：$I_E([6,2]) = -\frac{6}{8}log_2\frac{6}{8} - \frac{2}{8}log_2\frac{2}{8} = 0.81128$

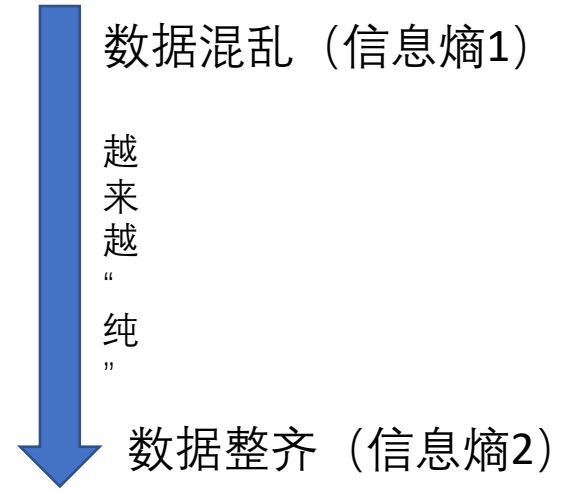
Strong分支的熵：$I_E([3,3]) = -\frac{3}{6}log_2\frac{3}{6} - \frac{3}{6}log_2\frac{3}{6} = 1.0$

特征划分选择

原数据集的信息熵（不论谁做根节点，根节点信息熵都是这个值）：

$$I_E([9,5]) = -\frac{9}{14}log_2\frac{9}{14} - \frac{5}{14}log_2\frac{5}{14} = 0.94029$$

信息增益（Information gain）



信息增益 = 信息熵1 - 信息熵2

特征划分选择

信息增益（Information gain）

用属性a对样本集D进行划分所获得的"信息增益" (information gain)

$$\mathrm{Gain}(D,a)=\mathrm{Ent}(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v)\ .$$

考虑到不同的分支结点所包含的样本数不同，给分支结点赋予权重。

当Humidity作为父节点时：

High分支的熵：$I_E([3,4])=-\frac{3}{7}log_2\frac{3}{7}-\frac{4}{7}log_2\frac{4}{7}=0.98533$

Normal分支的熵：$I_E([1,6])=-\frac{1}{7}log_2\frac{1}{7}-\frac{6}{7}log_2\frac{6}{7}=0.59159$

$$Gain(humidity)=0.94029-\left(0.98533\times\frac{7}{14}+0.59159\times\frac{7}{14}\right)=0.15183$$

特征划分选择

当Wind作为父节点时：

Weak分支的熵：$I_E([6,2]) = -\frac{6}{8}log_2\frac{6}{8} - \frac{2}{8}log_2\frac{2}{8} = 0.81128$

Strong分支的熵：$I_E([3,3]) = -\frac{3}{6}log_2\frac{3}{6} - \frac{3}{6}log_2\frac{3}{6} = 1.0$

$$Gain(wind) = 0.94029 - \left(0.81128 \times \frac{8}{14} + 1.0 \times \frac{6}{14}\right) = 0.04813$$

当Outlook作为父节点时：

Sunny分支的熵：$I_E([3,2]) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.9709508$

Overcast分支的熵：$I_E([0,4]) = -\frac{0}{4}log_2\frac{0}{4} - \frac{4}{4}log_2\frac{4}{4} = 0$

Rain分支的熵：$I_E([3,2]) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.9709508$

$Gain(Outlook) = ?$

特征划分选择

当Wind作为父节点时：

Weak分支的熵：$I_E([6,2]) = -\frac{6}{8}log_2\frac{6}{8} - \frac{2}{8}log_2\frac{2}{8} = 0.81128$

Strong分支的熵：$I_E([3,3]) = -\frac{3}{6}log_2\frac{3}{6} - \frac{3}{6}log_2\frac{3}{6} = 1.0$

$$Gain(wind) = 0.94029 - \left(0.81128 \times \frac{8}{14} + 1.0 \times \frac{6}{14}\right) = 0.04813$$

当Outlook作为父节点时：

Sunny分支的熵：$I_E([3,2]) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.9709508$

Overcast分支的熵：$I_E([0,4]) = -\frac{0}{4}log_2\frac{0}{4} - \frac{4}{4}log_2\frac{4}{4} = 0$

Rain分支的熵：$I_E([3,2]) = -\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5} = 0.9709508$

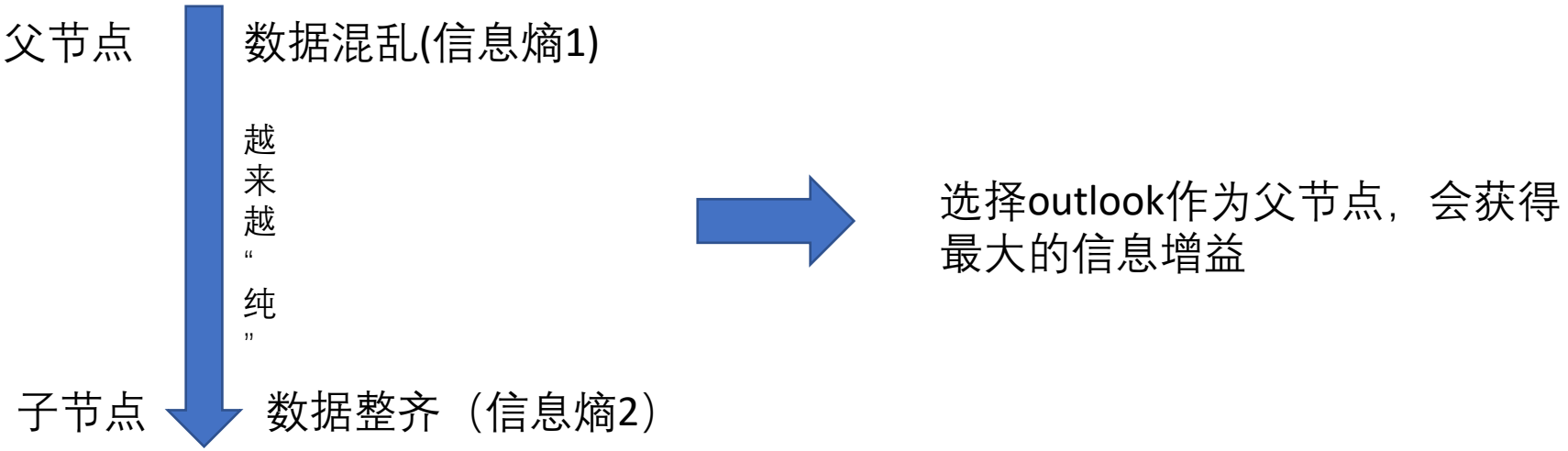
$$Gain(Outlook) = 0.94029 - \left(0.97095 \times \frac{5}{14} + 0 \times \frac{4}{14} + 0.97095 \times \frac{5}{14}\right) = 0.24675$$

特征划分选择

$$Gain(humidity) = 0.94029 - \left(0.98533 \times \frac{7}{14} + 0.59159 \times \frac{7}{14}\right) = 0.15183$$

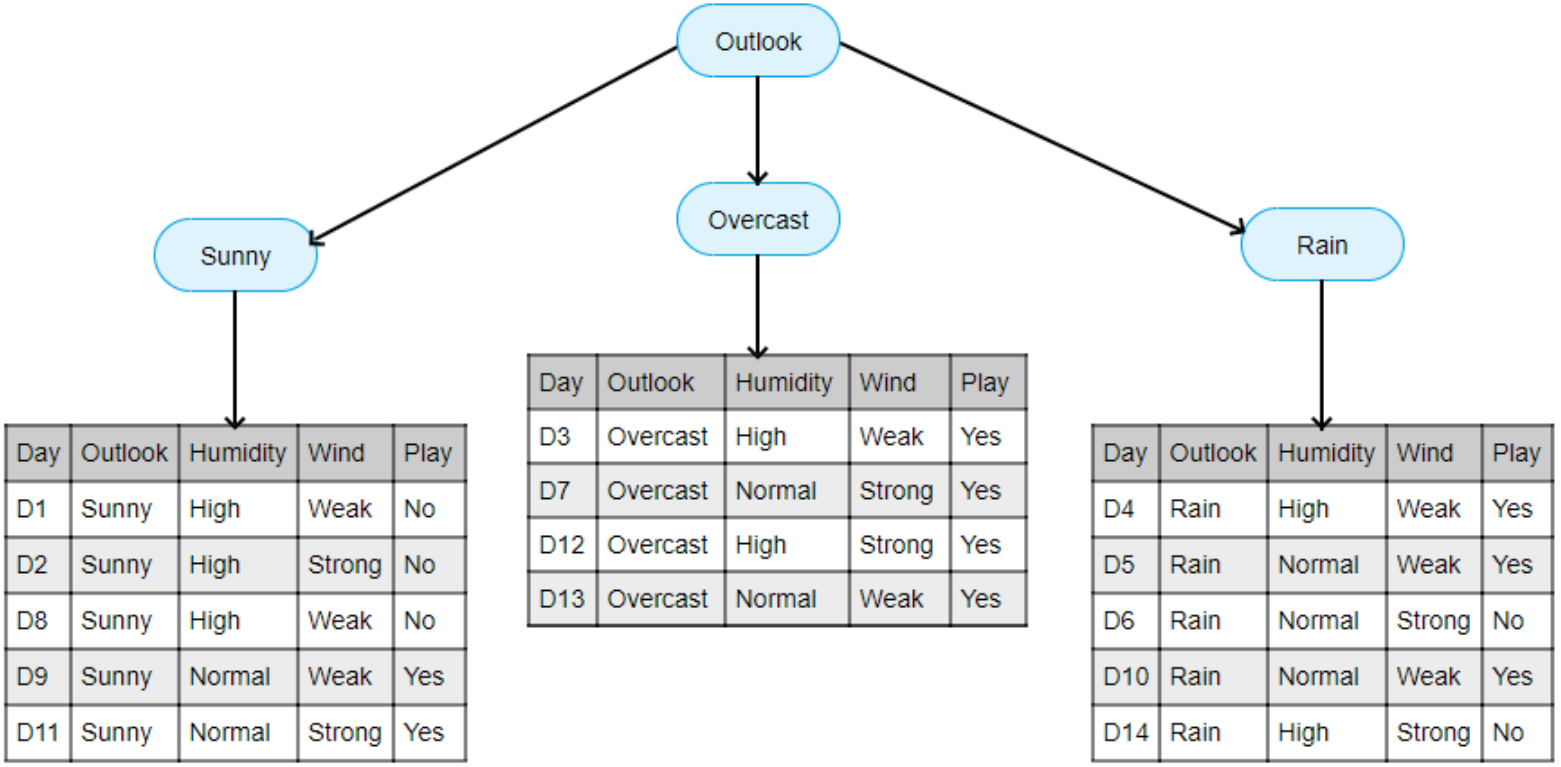
$$Gain(wind) = 0.94029 - \left(0.81128 \times \frac{8}{14} + 1.0 \times \frac{6}{14}\right) = 0.04813$$

$$Gain(Outlook) = 0.94029 - \left(0.97095 \times \frac{5}{14} + 0 \times \frac{4}{14} + 0.97095 \times \frac{5}{14}\right)$$
$$= 0.24675$$



特征划分选择

计算到这里，我们才确定了根节点为outlook



Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes

Day	Outlook	Humidity	Wind	Play
D3	Overcast	High	Weak	Yes
D7	Overcast	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes

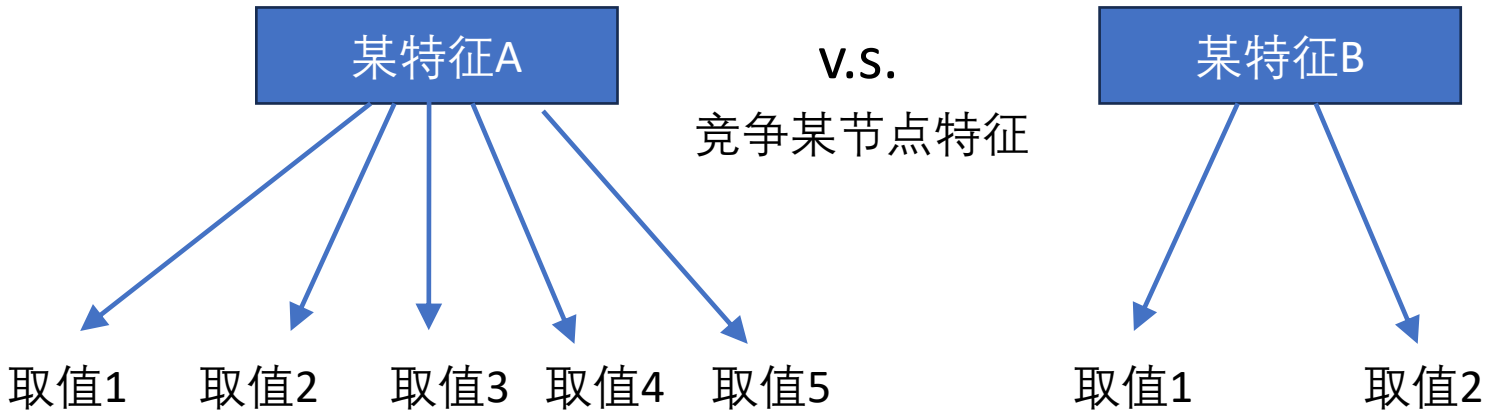
Day	Outlook	Humidity	Wind	Play
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D10	Rain	Normal	Weak	Yes
D14	Rain	High	Strong	No

Sunny的子节点是Humidity还是wind还要经过相似的计算。

这种用**信息增益**为准则来划分属性（特征）的方法就是**ID3决策树**学习算法

信息增益的局限性

- 信息增益对于拥有更多取值的特征拥有偏向性。



- 当一个特征（如特征A）有很多可能的取值时，会导致数据集被分割成许多较小的子数据集，这些小数据集往往因为样本量较少且同类别样本较高而具有较低的熵。因此，计算出的信息增益会相对较大，使得这个特征看起来像是一个很好的选择用于分割数据。

信息增益率

- 为了避免信息增益准则的偏向性，引入信息增益率。
- 信息增益率（Gain Ratio）是C4.5决策树算法中使用的特征选择方法，它是基于信息增益（Information Gain）算法ID3的一个改进。

$$\text{信息增益率} = \frac{\text{信息增益}}{\text{分割信息}}$$

$$SplitInfo(D,a) = -\sum_{v=1}^{V} \frac{|D^v|}{|D|} \log_2 \frac{|D^v|}{|D|}$$

当某特征拥有大量取值但每个取值的样本量很小时，它的分割信息就会相对比较大，导致其增益率不会过高，从而避免偏向性。

基尼不纯度

基尼不纯度（Gini Impurity）是CART（Classification and Regression Trees）决策树算法中用于度量数据子集的纯度的一个标准。它是一个衡量数据集中类别混杂程度的指标。基尼不纯度越低，表示数据集的纯度越高。

计算基尼不纯度

对于一个包含多个类别的数据集，基尼不纯度可以用以下公式计算：

$$Gini(p) = 1 - \sum_{i=1}^{n} p_i^2$$

其中：

- p_i表示数据集中第i个类别出现的概率。
- n是类别的总数。

基尼不纯度

$$Gini(p) = 1 - \sum_{i=1}^{n} p_i^2$$

- 如果数据集中的所有元素都属于同一个类别（即完全纯净），那么基尼不纯度为 0。如果数据集中的元素均匀地分布在各个类别中（即最不纯净），那么基尼不纯度达到最大值（$1 - \frac{1}{\mathrm{n}}$）。

- CART决策树通过计算**基尼不纯度**作为分裂节点特征的标准。

回归树（Regresssion Tree）

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
3	Overcast	Hot	High	Weak	46
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
7	Overcast	Cool	Normal	Strong	43
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
10	Rain	Mild	Normal	Weak	46
11	Sunny	Mild	Normal	Strong	48
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44
14	Rain	Mild	High	Strong	30

根据天气，气温，湿度，以及风速来判断打球的人数

回归树（Regresssion Tree）

- 在 **回归树**里，通常不用分类树里的 **信息增益（率）**，而是用 **方差/标准差** 作为节点划分的指标。

对于整个数据集，它的标准差计算如下：

Golf players = {25, 30, 46, 45, 52, 23, 43, 35, 38, 46, 48, 52, 44, 30}

Average of golf players = (25 + 30 + 46 + 45 + 52 + 23 + 43 + 35 + 38 + 46 + 48 + 52 + 44 + 30)/14 = 39.78

Standard deviation of golf players = √[((25 – 39.78)2 + (30 – 39.78)2 + (46 – 39.78)2 + … + (30 – 39.78)2)/14] = 9.32

回归树（Regresssion Tree）

Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

标准差：7.78

如果将outlook作为根节点

标准差：3.49

Outlook

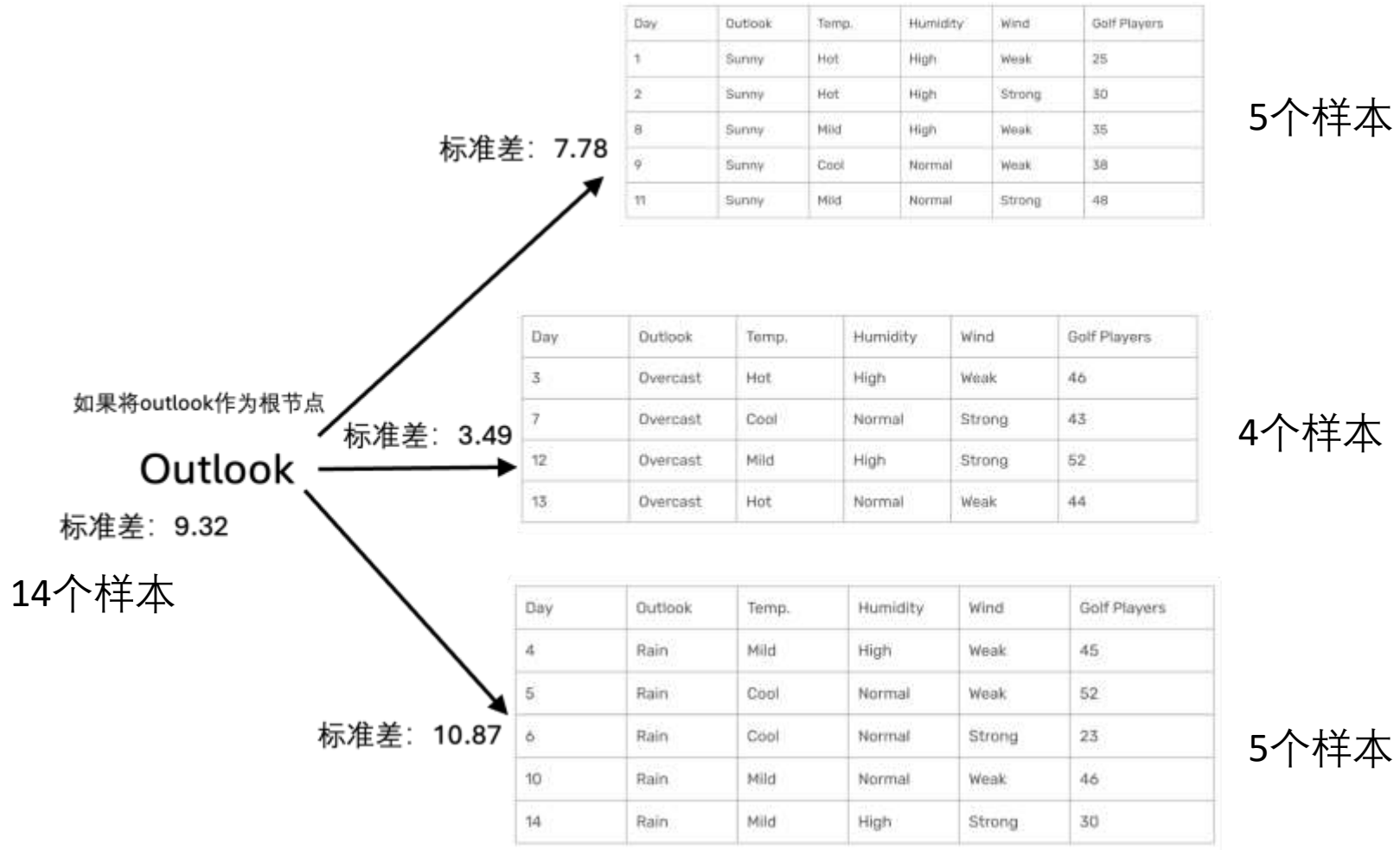
标准差：9.32

Day	Outlook	Temp.	Humidity	Wind	Golf Players
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

标准差：10.87

回归树（Regresssion Tree）



Day	Outlook	Temp.	Humidity	Wind	Golf Players
1	Sunny	Hot	High	Weak	25
2	Sunny	Hot	High	Strong	30
8	Sunny	Mild	High	Weak	35
9	Sunny	Cool	Normal	Weak	38
11	Sunny	Mild	Normal	Strong	48

Day	Outlook	Temp.	Humidity	Wind	Golf Players
3	Overcast	Hot	High	Weak	46
7	Overcast	Cool	Normal	Strong	43
12	Overcast	Mild	High	Strong	52
13	Overcast	Hot	Normal	Weak	44

Day	Outlook	Temp.	Humidity	Wind	Golf Players
4	Rain	Mild	High	Weak	45
5	Rain	Cool	Normal	Weak	52
6	Rain	Cool	Normal	Strong	23
10	Rain	Mild	Normal	Weak	46
14	Rain	Mild	High	Strong	30

如果将outlook作为根节点，那么标准差的损失是

9.32 - [(5/14)x7.78 + (4/14)x3.49 + (5/14)x10.87] = 1.66

回归树（Regresssion Tree）

同理

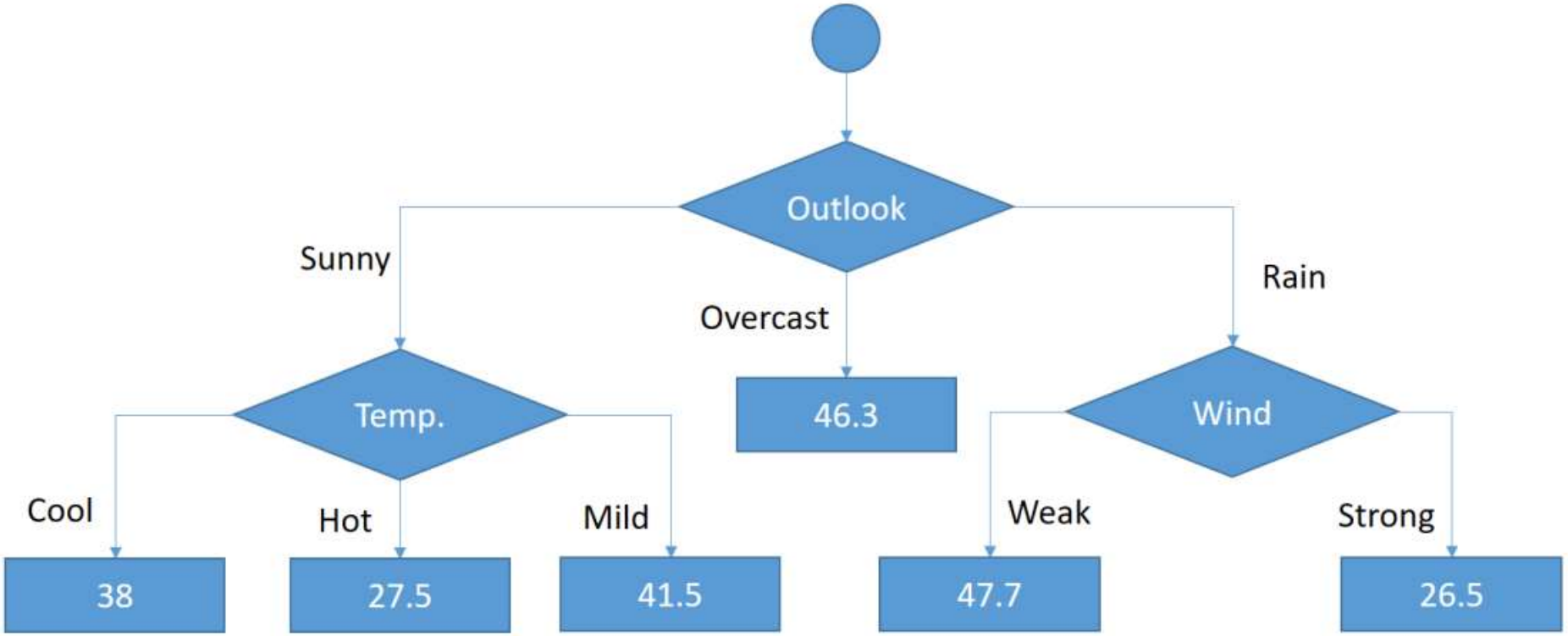
如果将Temperature作为根节点，那么标准差的损失是
9.32 - [(4/14)x8.95 + (4/14)x10.51 + (6/14)x7.65] = 0.47

如果将Humidiy作为根节点，那么标准差的损失是
9.32 - [(7/14)x9.36 + (7/14)x8.73] = 0.27

如果将Wind作为根节点，那么标准差的损失是
9.32 -[(6/14)x10.59 + (8/14)x7.87] = 0.29

Outlook能够使得标准差下降的最多，所以把Outlook作为根节点

回归树（Regresssion Tree）



在 回归树里，叶节点的值最后会有一个值，这个值是该叶节点数据子集的目标值的平均数（均值）。

决策树剪枝（Pruning）

- 剪枝，顾名思义，就是把树上“没必要的枝条”剪掉。主要是通过限制最大深度、最小样本数来减小模型的复杂程度，从而避免过拟合。

为什么要减枝?

决策树如果不加限制，会一直往下分裂，直到：

- 每个叶子里只有一个样本，或者
- 叶子里的数据完全纯净。

这种树虽然在训练集上误差很低（甚至为 0），但在测试集上往往会过拟合：

- 学到了太多训练数据里的“噪声”
- 预测新数据时效果不好

过拟合

所以我们要通过 **减枝** 来控制树的复杂度。

剪枝方法

预剪枝 (Pre-pruning)

在建树时就加限制，避免树长得太复杂。
例如：

- 限制树的最大深度 (`max_depth`)
- 限制叶子节点最少样本数 (`min_samples_leaf`)
- 限制分裂后信息增益或方差下降必须超过某个阈值

- 好处：快
- 坏处：可能剪掉了一些有用的分支（欠拟合风险）

后剪枝 (Post-pruning)

先把树尽量长大，再回头剪掉不必要的枝。

常见做法：

- 在验证集上评估，若某个分支并没有提高预测效果，就合并它。

剪枝方法

优点：

1. **可解释性强**：像 if-else 规则，人类容易理解。
2. **处理能力强**：能同时处理数值型和类别型特征，不需要标准化/归一化的预处理。
3. **能拟合非线性关系**：自动划分特征空间，适应复杂的决策边界。

缺点：

1. **容易过拟合**：不剪枝或不限制深度时，树会记住噪声。
2. **不稳定**：对数据的小变化非常敏感，树结构可能完全不同。
3. **单树性能有限**：预测效果往往不如集成方法（随机森林、XGBoost）。