人工智能通识教程
（农林院校版）

https://ai4ag.github.io

第四章

大数据分析：吹沙到金

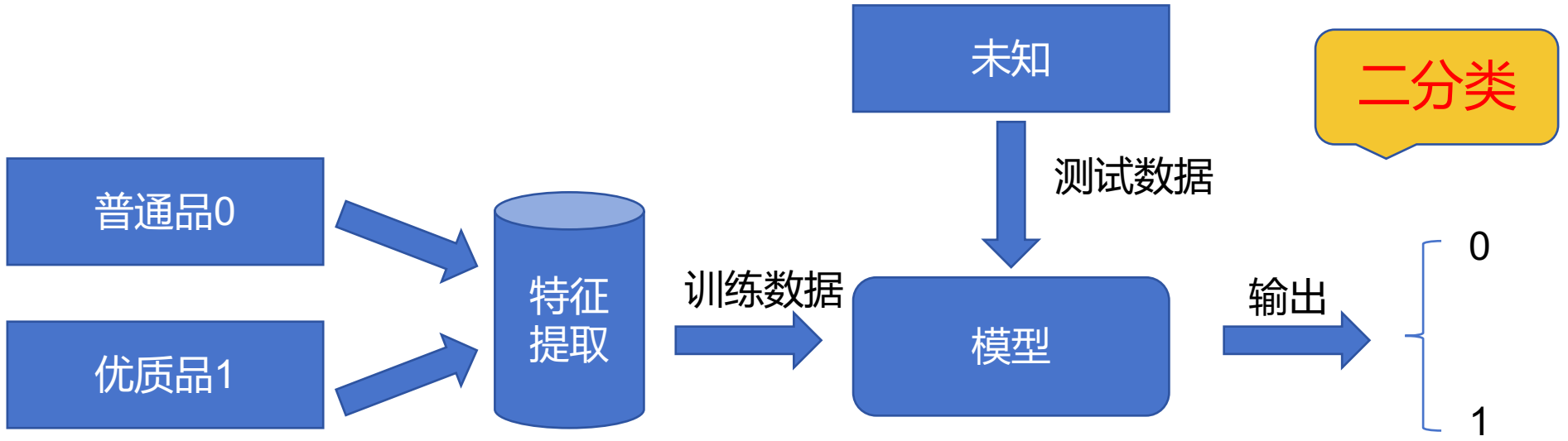
课程导读


大数据分析
知识背景
大数据分析的五个处理环节
数据采集
数据库与文件
传感器与物联网的应用
数据爬虫
定义
工作原理
预处理
数据清洗
特征缩放
归一化
标准化
特征编码
独热编码
标签编码
主成分分析
线性判别分析
PCA与LDA的区别
数据挖掘
数据建模
判别可视化

课程导读


大数据分析
基本概念剖析
分类
定义
分类的过程
常见的分类方法
逻辑回归
神经网络
支持向量机
决策树与随机森林
评价指标
过拟合处理
聚类
K-means聚类
谱聚类
层次聚类
密度聚类
多视图聚类
深度聚类
预测分析
回归分析
线性回归预测
分类分析
时间序列分析
应用场景
经典方法
知识图谱
知识图谱搭建
实体统一
指代消解
关系抽取
实体识别
数据获取
数据存储
应用
农业信息查询系统

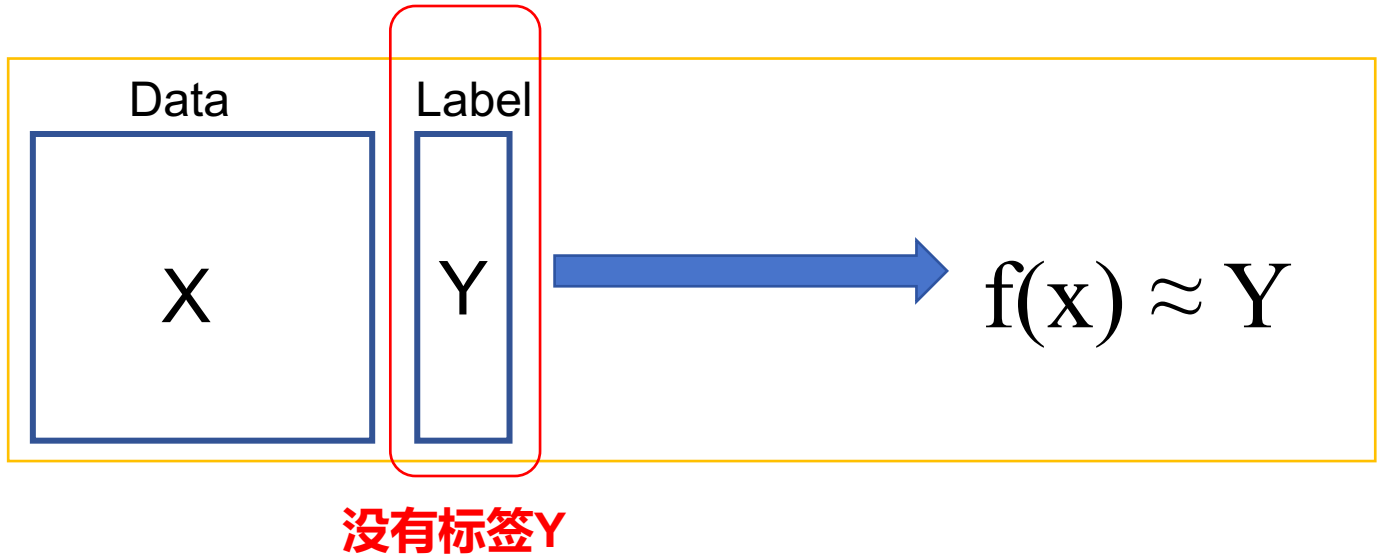
课程回顾

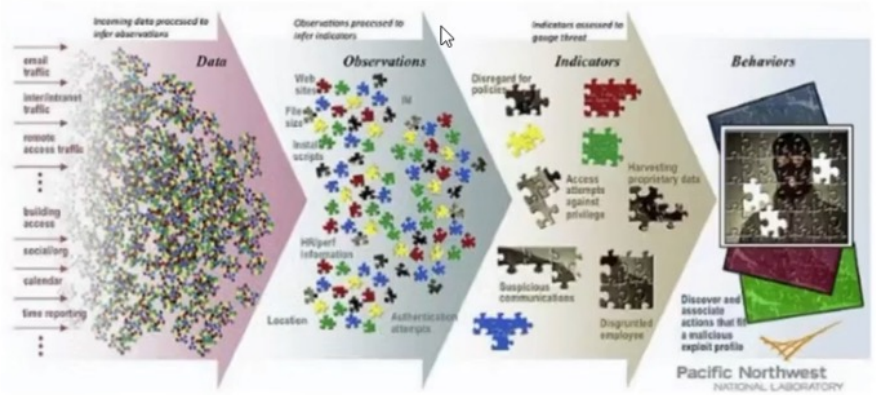


假设某农场在采摘番茄后，希望通过构建机器学习模型对每个番茄的品质进行判断，从而自动识别“优质品”（标签为1）与“普通品”（标签为0），以提高分拣环节的智能化水平，降低人工误判率。

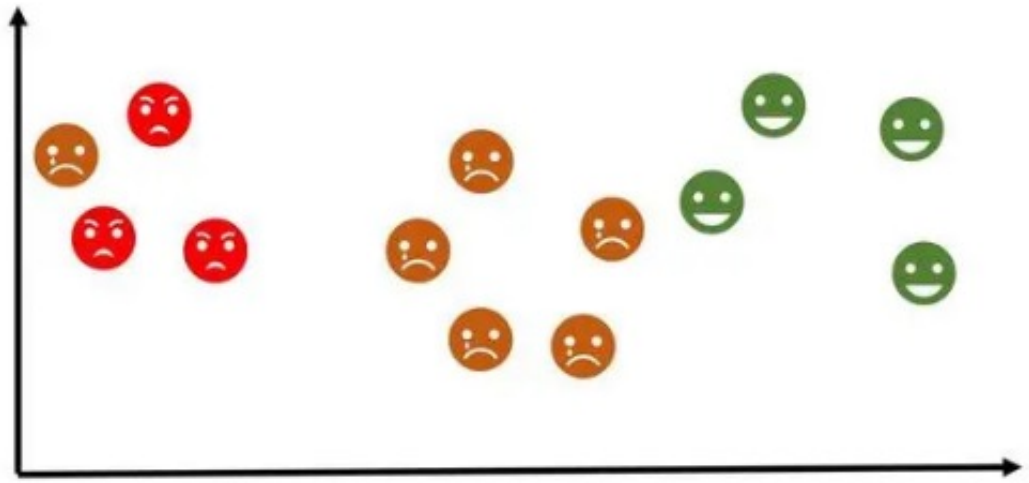
课程回顾

- 聚类分析是一种极具代表性的无监督机器学习
- 根据主观定义的相似性对未知分布的数据进行分类的过程
- 帮助人们去发现隐藏在数据中的类结构





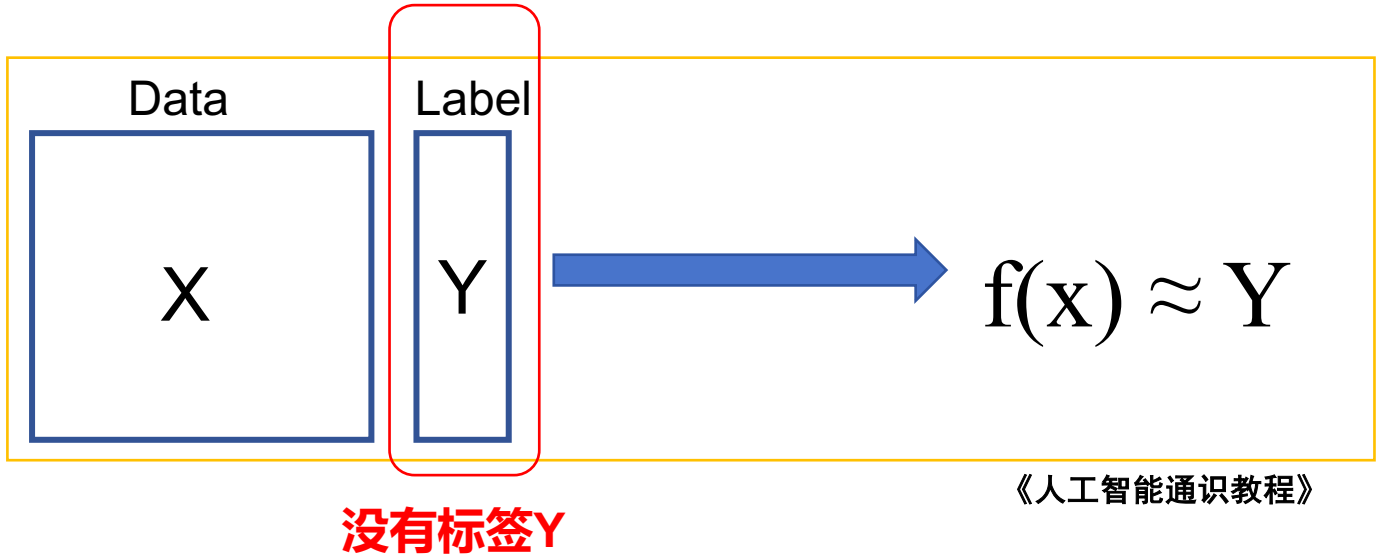
课程回顾

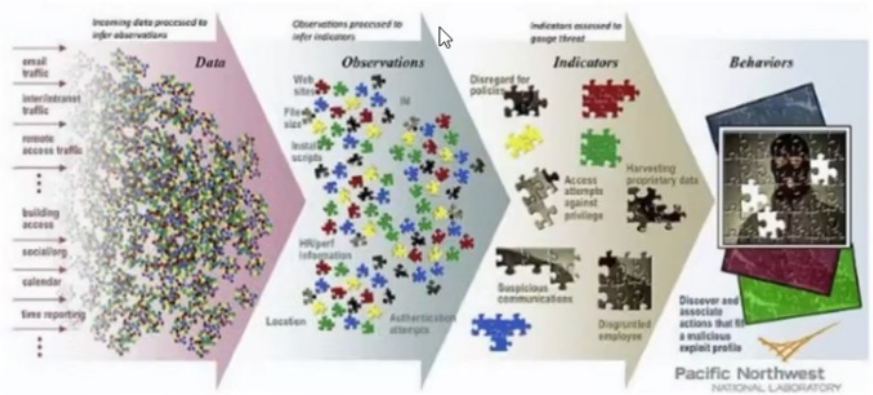


聚类就是对多个未知标注的数据，按数据的内在相似性将数据划分为多个类别，使类内的相似度较大而类间的相似度较小。把“相似的东西”归到同一组，让组内亲密，组间疏远。

课程回顾

- 聚类分析是一种极具代表性的无监督机器学习
- 根据主观定义的相似性对未知分布的数据进行分类的过程
- 帮助人们去发现隐藏在数据中的类结构





课程回顾

分类与聚类的区别

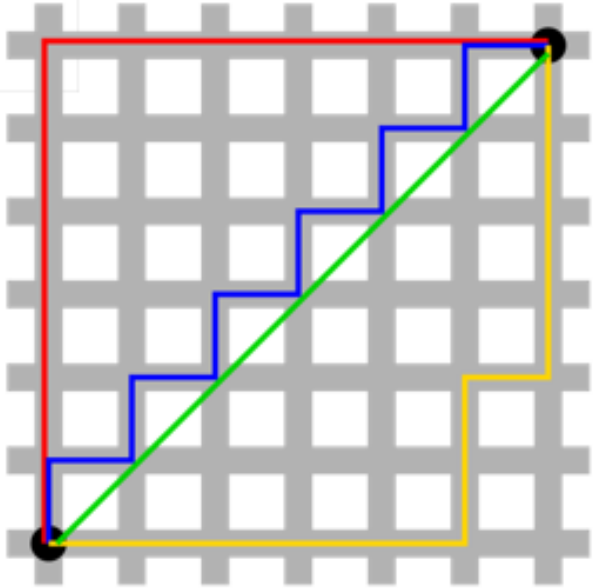
维度	分类（Classification）	聚类（Clustering）
学习方式	监督学习（依赖带标签数据）	无监督学习（无需标签）
目标	预测数据属于“已知类别”中的哪一类	发现数据中“未知的自然分组”
数据要求	需要大量标注数据（标签是关键）	仅需原始数据，无需标注
输出结果	数据所属的预定义类别（如“阳性 / 阴性”）	数据所属的簇编号（如簇 1、簇 2）
模型评估	可通过准确率、F1 值等指标量化评估	难以量化，依赖业务解释性（如轮廓系数）
典型算法	逻辑回归、决策树、SVM、神经网络等	K-Means、DBSCAN、谱聚类、层次聚类等

课堂小测

假定聚类中心个数K=3，拟采用曼哈顿距离，计算前4次迭代的聚类中心坐标值。

- 初始化

Samples	
A1	(2,10)
A2	(2,5)
A3	(8,4)
A4	(5,8)
A5	(7,5)
A6	(6,4)
A7	(1,2)
A8	(4,9)



曼哈顿距离

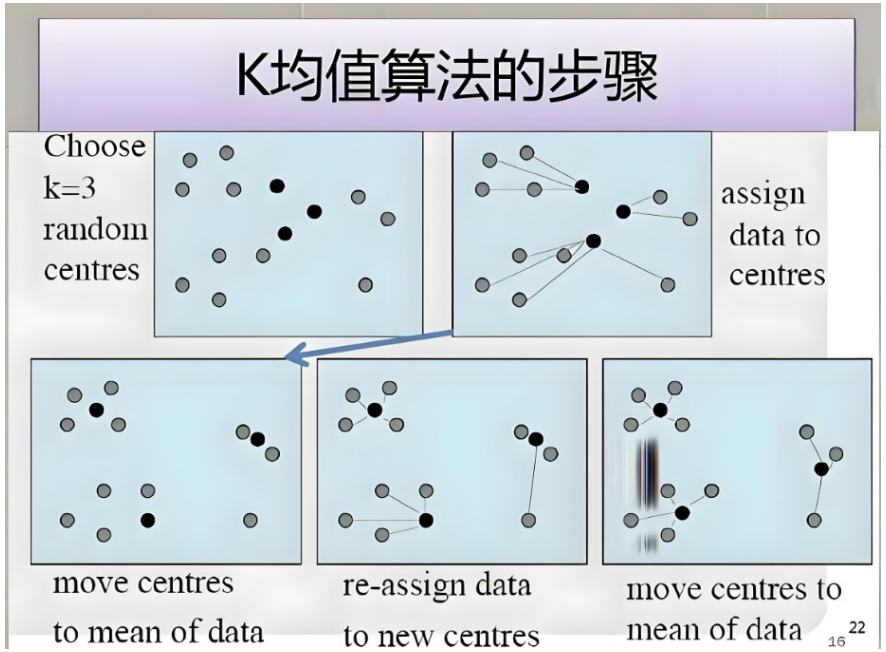
$$d_{u,v} = |u_1 - v_1| + |u_2 - v_2|$$

$$\begin{aligned} d(\mathrm{A1}, \mathrm{A2}) &= |5-2| + |8-10| \\ &= 3 + 2 \\ &= 5 \end{aligned}$$

课程回顾

K 均值聚类（K-means Clustering）目标是通过计算数据样本点与簇中心之间的欧氏距离，将数据划分为预定义的K 个簇。基本实现步骤如下：

- **初始化：**随机选择K 个样本点作为初始簇中心。
- **分配：**计算每个样本点到所有簇中心的距离，并将其分配到最近的簇。
- **更新：**重新计算每个簇的中心点为该簇内所有样本点的均值。
- **迭代：**重复分配和更新步骤，直到簇中心不再变化或达到预定的迭代次数。

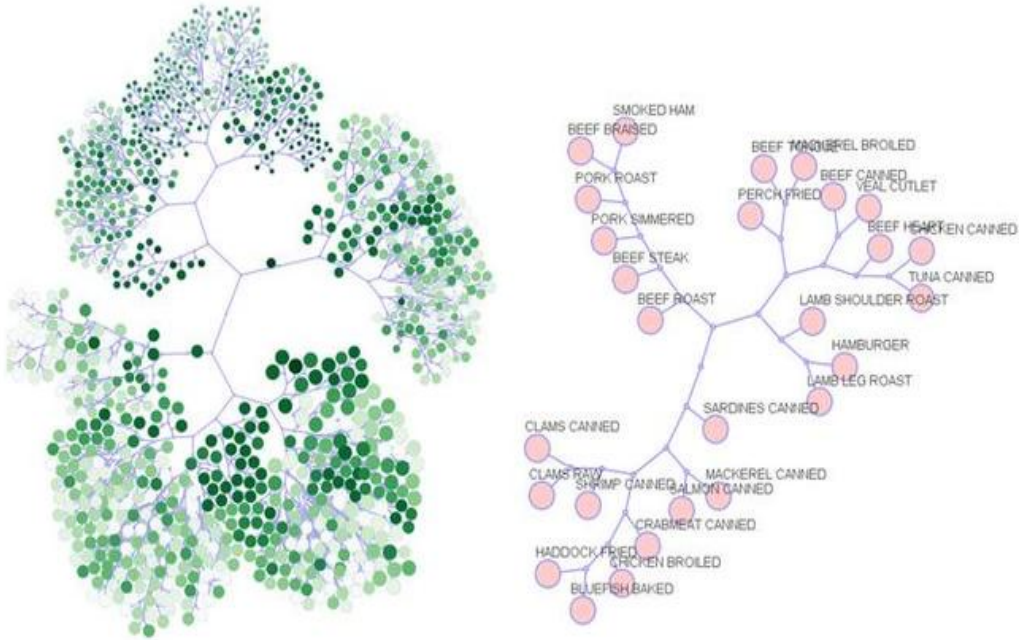


课程回顾

谱聚类（Spectral Clustering）是一种基于图论和特征向量分解的聚类技术，特别适用于处理非凸形状的簇。

谱聚类的基本实现步骤如下：

- **构建相似性矩阵**：计算样本点之间的相似性，通常使用高斯核函数或欧氏距离。
- **构建拉普拉斯矩阵**：根据相似性矩阵计算图的拉普拉斯矩阵。
- **特征向量分解**：对拉普拉斯矩阵进行特征分解，选择前K个最小的特征值对应的特征向量。
- **低维空间聚类**：将特征向量作为新的特征空间，使用K均值算法进行聚类。



本节目录

- 数据可视化的核心方法
- 大数据技术在相关领域中的应用

数据可视化

数据可视化

- **在海量的数据集中，我们从表格中难以直观获取有价值信息，应该怎么办?**

某农业公司A地鸡肉销售情况表

月份	销量	金额	重量	单价	天龄
1	87	2311.94	327.45	7.06	79.14
2	135	3960.00	550.00	7.20	78.00
3	90	2394.29	341.67	7.01	79.67
4	114	3226.90	455.00	7.09	80.47
5	54	1469.23	205.42	7.15	79.50
6	64	1849.94	231.24	8.00	77.49
7	160	4392.79	615.86	7.13	80.35
8	89	2429.37	340.14	7.14	80.04
9	165	4405.68	621.40	7.09	76.93
10	297	7753.20	1092.00	7.10	75.00
11	289	7928.21	1129.40	7.02	78.26
12	487	13455.43	1855.75	7.25	86.44

某农业公司B地鸡肉销售情况表

月份	销量	金额	重量	单价	天龄
1	87	2311.94	327.45	7.06	79.14
2	135	3960.00	550.00	7.20	78.00
3	90	2394.29	341.67	7.01	79.67
4	114	3226.90	455.00	7.09	80.47
5	54	1469.23	205.42	7.15	79.50
6	64	1849.94	231.24	8.00	77.49
7	160	4392.79	615.86	7.13	80.35
8	89	2429.37	340.14	7.14	80.04
9	165	4405.68	621.40	7.09	76.93
10	297	7753.20	1092.00	7.10	75.00
11	289	7928.21	1129.40	7.02	78.26
12	487	13455.43	1855.75	7.25	86.44

数据可视化

数据可视化

- 如果我们能够将这些抽象的、复杂的数据通过图形、图表、地图、信息图、仪表盘等视觉化形式呈现，将数据中隐藏的规律、趋势、关联或异常转化为直观易懂的图像，帮助人们快速理解数据含义、发现问题、分析逻辑并做出决策。基于此，我们称之为数据可视化。
- 常见的可视化图有：折线图、直方图、泡泡图、箱图、小提琴图，柱状图、散点图、饼图等

数据可视化

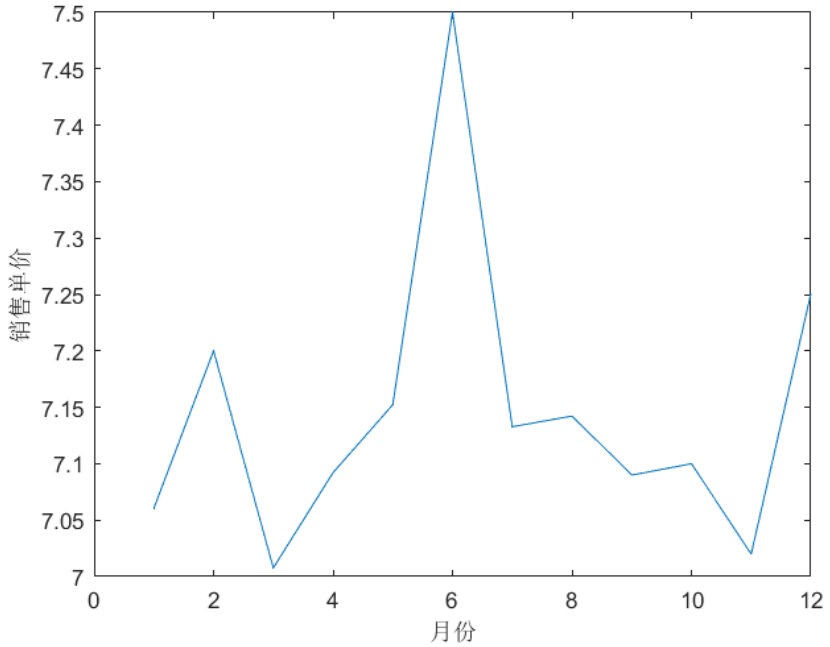
Matlab的使用

- Matlab是一款功能强大的编程软件，广泛应用于物理、数学、金融学、生物学等领域，内置了丰富的可视化绘图工具，能够快速实现数据可视化操作。MATLAB 的绘图功能非常灵活，可通过组合函数实现复杂可视化，也可通过交互界面（如工具栏）手动调整图形细节。
- 下载链接：https://ww2.mathworks.cn/

Python的使用

- Python 的可视化能力依赖于各类可视化库，这些库封装了大量绘图函数，能快速将数据转化为图表。最常用的库包括 Matplotlib（基础库）、Seaborn（基于 Matplotlib 的高级库）、Plotly（交互式可视化库）等。
- 下载链接：https://www.python.org/

数据可视化



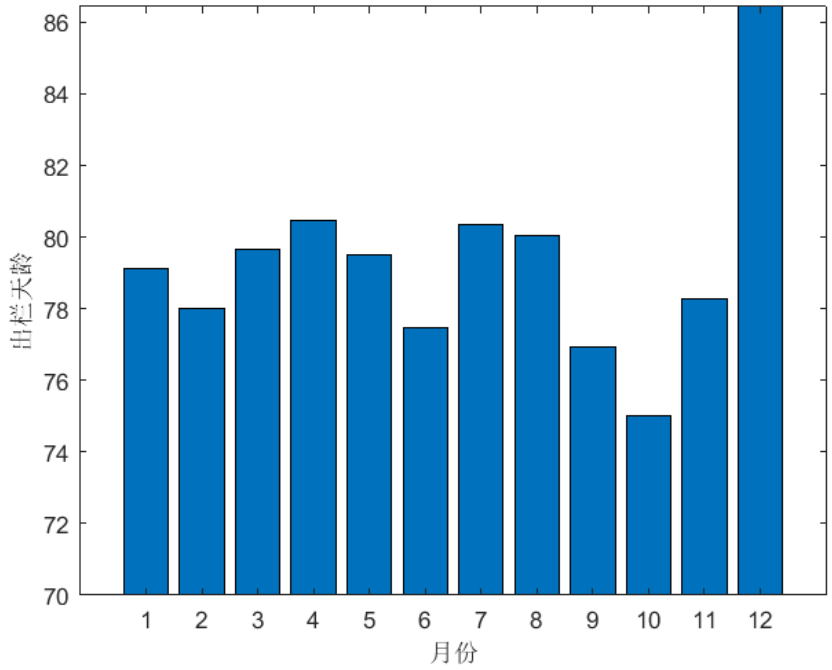
Matlab 代码

```
Table_1 = xlsread('./demo.xlsx','sheet1'); % 读取Excel表格
Month=Table_1(:,1); % 表格中第一列为月份
Prices = Table_1(:,5); % 表格中第五列为销售单价
plot(Month,Prices); % 使用Plot函数绘制折线图
xlabel('月份'); % 设置横坐标名称
ylabel('销售单价'); % 设置纵坐标名称
```

绘制折线图：折线图能够展示数据随时间变化的趋势。

数据可视化

直方图



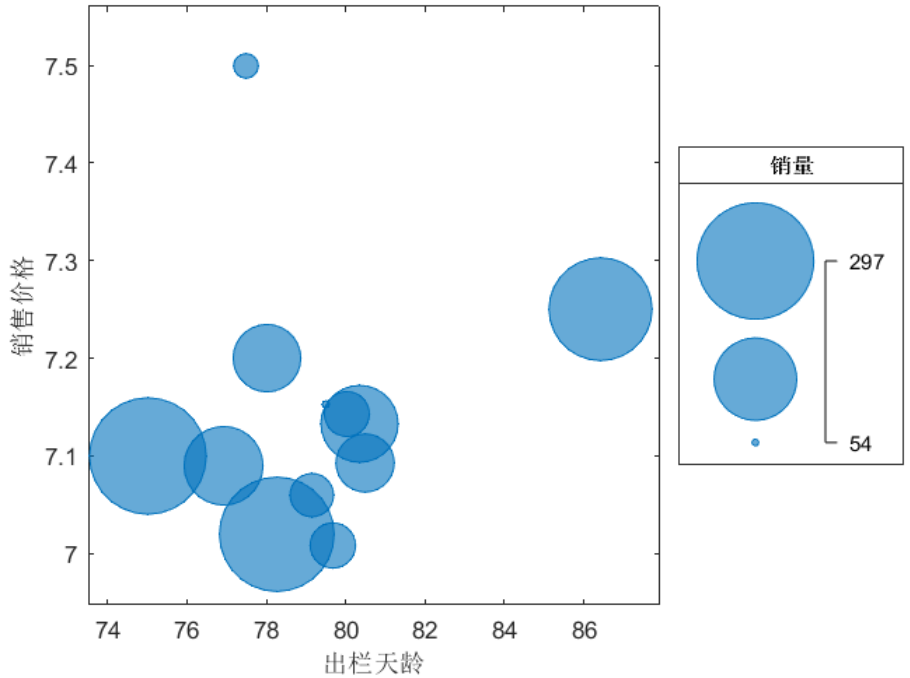
Matlab 代码

```
Month=Table_1(:,1); % 表格中第一列为月份
Times = Table_1(:,6); % 表格中第六列为出栏天龄
bar(Month,Times); % 使用bar函数绘制直方图
ylim([70 inf]) %设置直方图纵坐标最小值
xlabel('月份’ );
ylabel('出栏天龄');
```

绘制直方图：直方图常用于展示数据的分布情况，可将数据划分为不同的区间，计算每个区间数据的频次或频率，以柱状的形式进行可视化呈现。

数据可视化

泡泡图



Matlab 代码

```
Sales = Table_1(:,2); % 表格中第二列为销量

Times = Table_1(:,6); % 表格中第六列为出栏天龄

Prices = Table_1(:,5); % 表格中第五列为销售单价

bubblechart(Times,Prices,Sales); % 绘制泡泡图

xlabel('出栏天龄');

ylabel('销售价格  ');

bubblelegend('销量','Location','eastoutside'); % 添加注释框
```

绘制泡泡图：泡泡图是一种三维数据可视化图表，在二维平面上使用圆形泡泡的位置和大小来展示三维数据的相互关系。

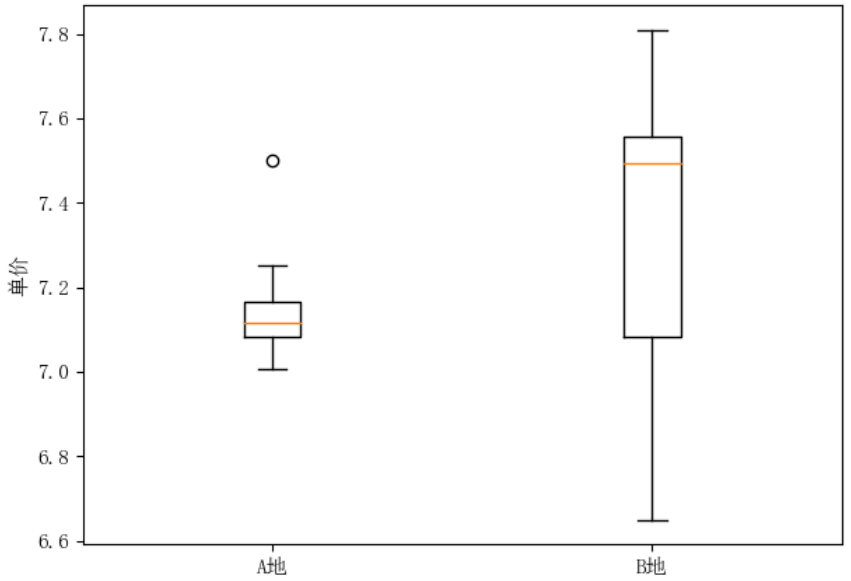
课间休息


CCTV 13
新 闻
CCTV.com
星期二 09:04
出口量较去年下降约25%，产量减少12%。
CCTV新闻 巴西东南部发生暴雨

数据可视化

数据可视化

箱图



Python 代码

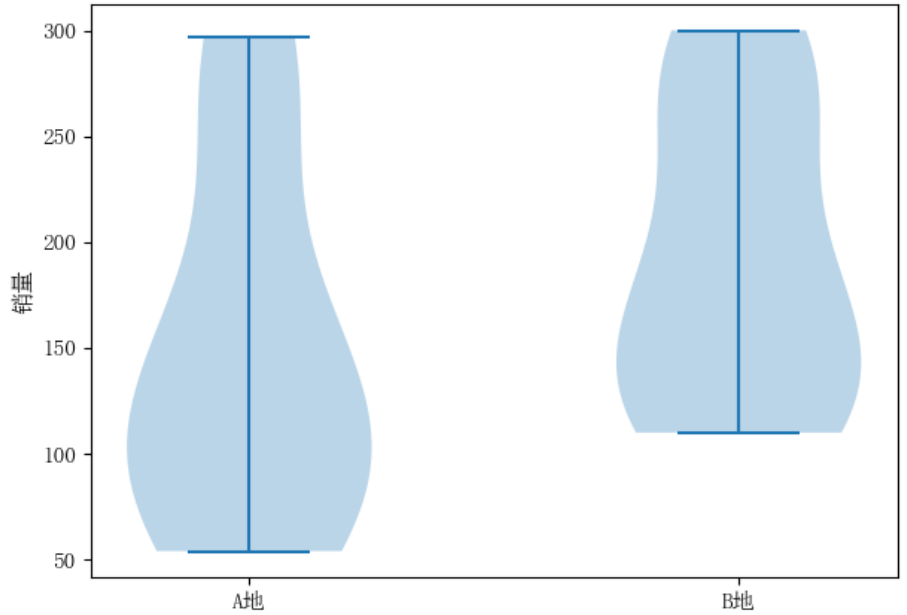
```
Prices_A = data_A[`单价`] # 读取表格中A地的销售单价
Prices_B = data_B[`单价`] # 读取表格中B地的销售单价
labels = 'A地', 'B地' # 设置图例名称
plt.ylabel("单价")
plt.boxplot([Prices_A, Prices_B], labels=labels) # 绘制箱图
plt.show() # 显示图像
```

绘制箱图：箱图用于展示数据的整体分布情况，包括中位数、四分位数和异常值。

数据可视化

数据可视化

小提琴图



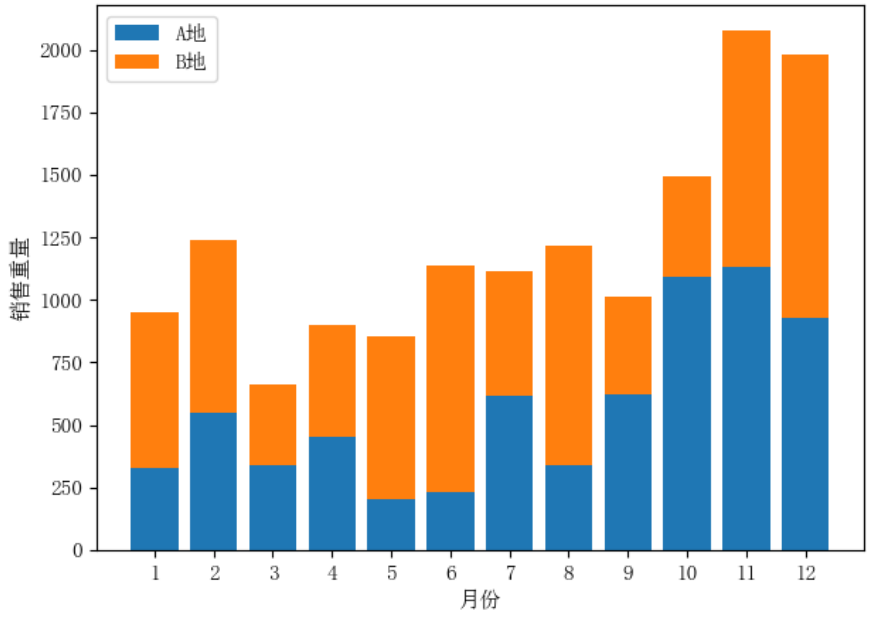
Python 代码

```
Sales_A = data_A['销量']
Sales_B = data_B['销量']
labels = 'A地', 'B地' # 设置图例名称
plt.xticks([1, 2], labels)
plt.ylabel("销量") # 设置y轴坐标名称
plt.violinplot([Sales_A, Sales_B]) # 绘制小提琴图
plt.show() # 显示图像
```

绘制小提琴图：小提琴图结合了箱图和密度图的特点，能够展示数据的分布和概率密度。

数据可视化

堆叠柱状图



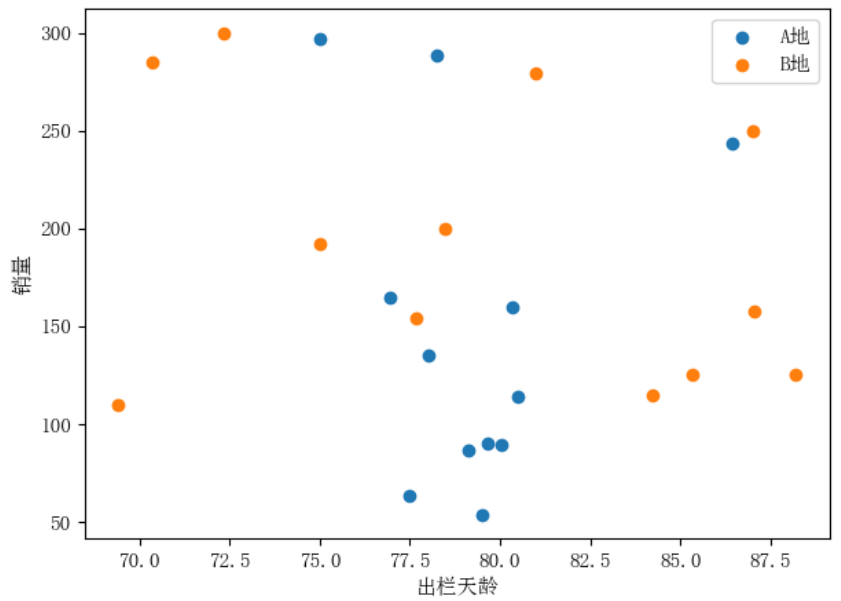
Python 代码

```
Weights_A = data_A['重量']
Weights_B = data_B['重量']
labels = ['1','2','3','4','5','6','7','8','9','10','11','12'  ]
#设置x轴坐标
plt.bar(labels, Weights_A,label='A地')  # 绘制A地柱状图
plt.bar(labels, Weights_B, bottom=Weights_A,label = 'B地')
# 绘制B地柱状图
plt.xlabel("月份")
plt.ylabel("销售重量")
```

绘制堆叠柱状图：堆叠柱状图可用于展示多个类别的相对比例，通过比较不同柱子的相对高度，直观地了解各个类别的贡献度。

数据可视化

散点图



Python 代码

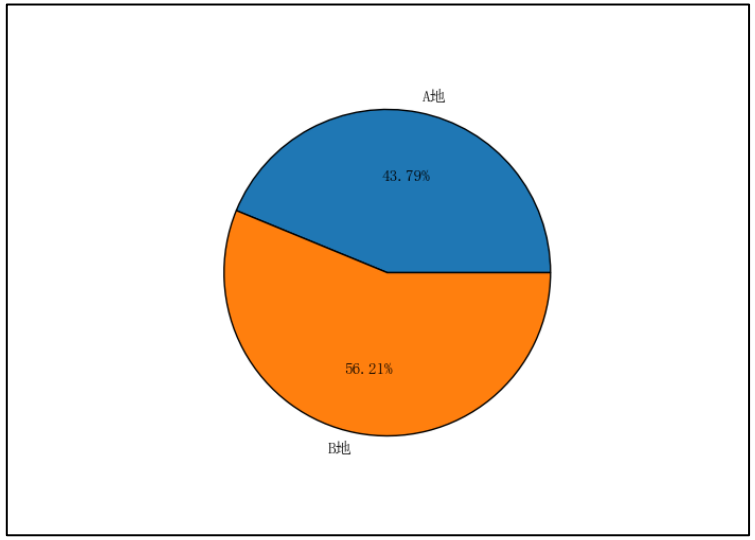
```
Days_A, Sales_A = data_A['天龄'], data_A['销量']
Days_B, Sales_B = data_B['天龄'], data_B['销量']
plt.scatter(Days_A, Sales_A, label='A地') # 绘制A地散点图
plt.scatter(Days_B, Sales_B, label='B地') # 绘制B地散点图
plt.xlabel("出栏天龄")
plt.ylabel("销量")
plt.legend(loc=1) # 将表格说明放置于图像第一象限
plt.show()
```

绘制散点图：散点图可用于展示两个变量之间的关系。

数据可视化

数据可视化

饼状图



Python 代码

```
Sales_A = data_A['销量']
Sales_B = data_B['销量']
plt.pie([sum(Sales_A), sum(Sales_B)],
labels=['A地', 'B地'],
autopct='%.2f%%', # 设置百分比
wedgeprops={'linewidth': 1, 'edgecolor': "black"}) # 设置边缘
plt.show()
```

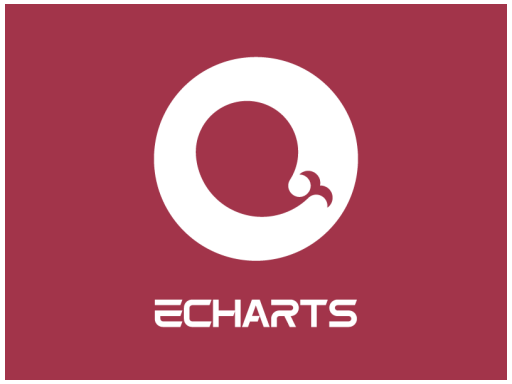
绘制饼状图：饼状图用于展示数据中各类别的比例关系。

数据可视化

数据可视化

- **使用其他高级工具实现数据可视化**

Wolfram|Alpha集成了数学、科学、技术等领域的数据处理功能，支持复杂计算与数据分析，满足多场景需求。



ECharts 是一款基于JavaScript的开源交互式可视化库，提供丰富图表类型与灵活定制能力，支持大规模数据渲染与响应式布局。

Plotly是一款跨语言的开源绘图库，支持Python、R、JavaScript等多种环境，可生成交互式、高质量且可导出或嵌入的动态图表。

大数据分析技术应用

应用概述

行业现状背景

现代农业的显著特点是规模庞大、供应链复杂，每天都会产生海量的农业数据。这些数据涵盖了供给侧与需求侧、客户与公司、农户与产品等多方面的复杂关系。在传统的农业生产模式中，许多工作依赖人工完成，例如养殖生产规划、农产品定价、精准营销等。但是，面对如此庞大的数据量，仅靠人工处理几乎是不可能的，导致大量历史数据被浪费，经验无法有效沉淀。

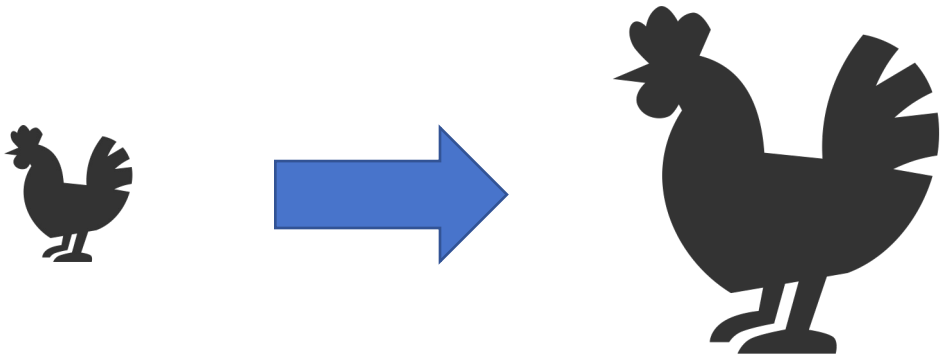
本节主要内容

随着大数据技术的快速发展，人工智能、数据挖掘和数据分析等技术为现代农业注入了新的活力。这些技术不仅能够高效地处理海量数据，还能从中提取有价值的信息，为农业生产和管理提供科学决策支持。在本节中，我们将通过四个实际应用案例，讲解大数据分析技术如何赋能现代农业。

案例4-1 基于LSTM的毛鸡品种毛利预测

- **应用背景**

毛鸡品种的培育周期约三个月，生产商需要根据历史数据预测未来三个月的毛利，以便合理规划养禽品种布局

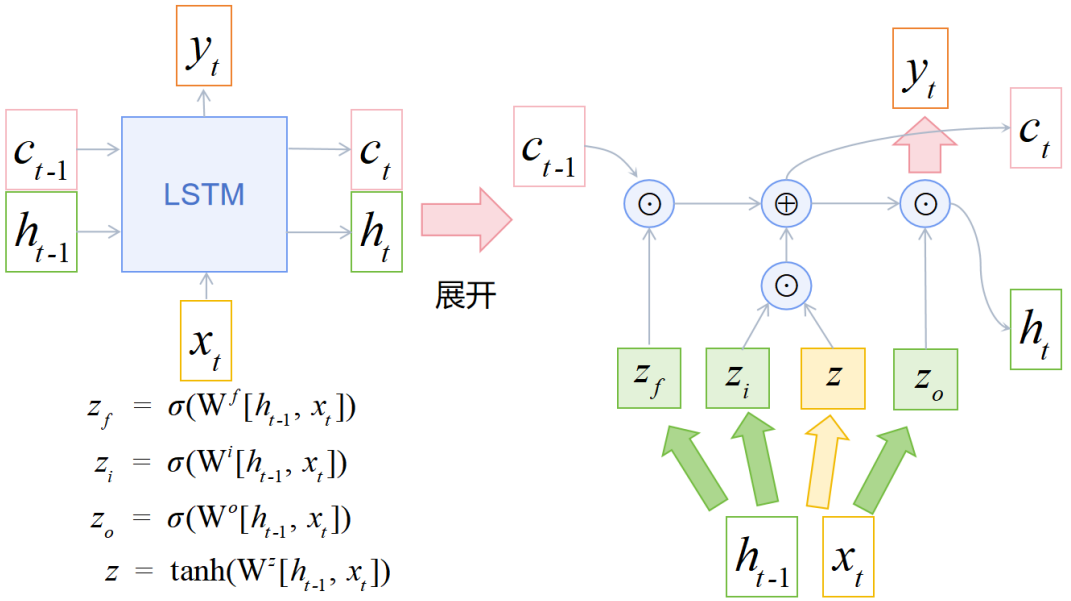


传统方法依赖人工经验，难以应对复杂的市场变化

案例4-1 基于LSTM的毛鸡品种毛利预测

技术实现

基于 LSTM（长短期记忆网络）的时间序列预测技术，能够有效捕捉历史数据中的趋势和规律，为应对毛利预测问题提供模型工具



LSTM

案例4-1 基于LSTM的毛鸡品种毛利预测

技术实现

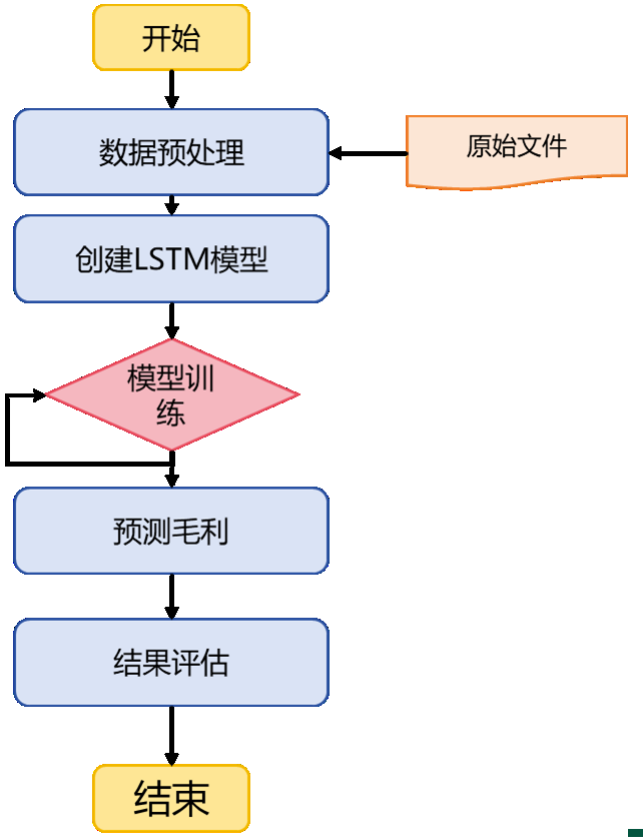
- 数据预处理

将每个毛鸡品种历史销量视作一个独立的时间序列，在此基础上，对历史毛鸡品种的毛利数据进行清洗和归一化处理，确保数据在同一量级并处理缺失值和空缺值

- 模型构建与训练

使用 LSTM 模型进行时间序列预测（如图 4-28 所示），以捕捉时间序列中的长期依赖关系。模型输入为某一品种过去六个月的毛利数据，输出为该品种未来三个月的毛利预测值

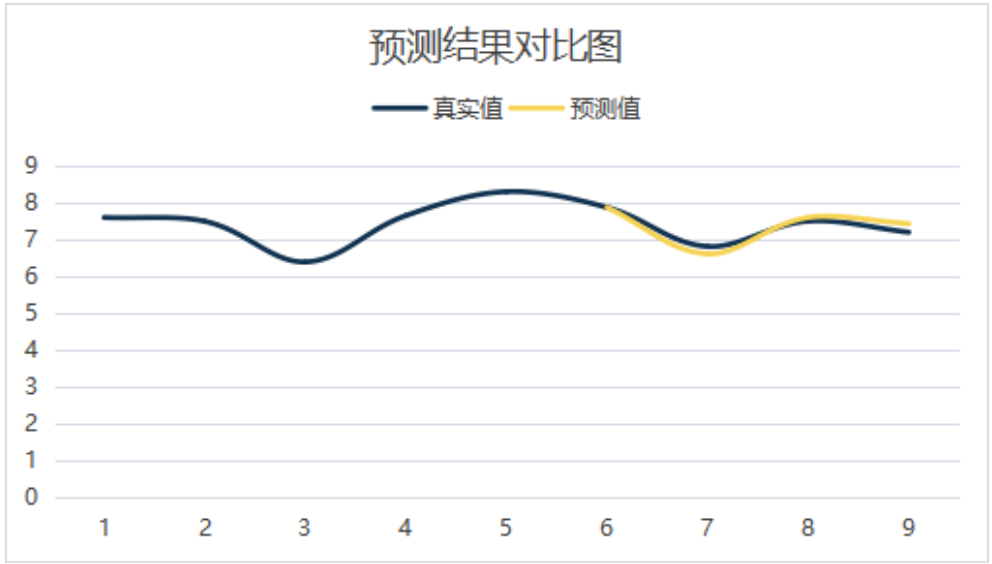
- 预测结果与评估



案例4-1 基于LSTM的毛鸡品种毛利预测

- **预期结果**

模型在测试集上的结果表明模型具有较高的预测准确性

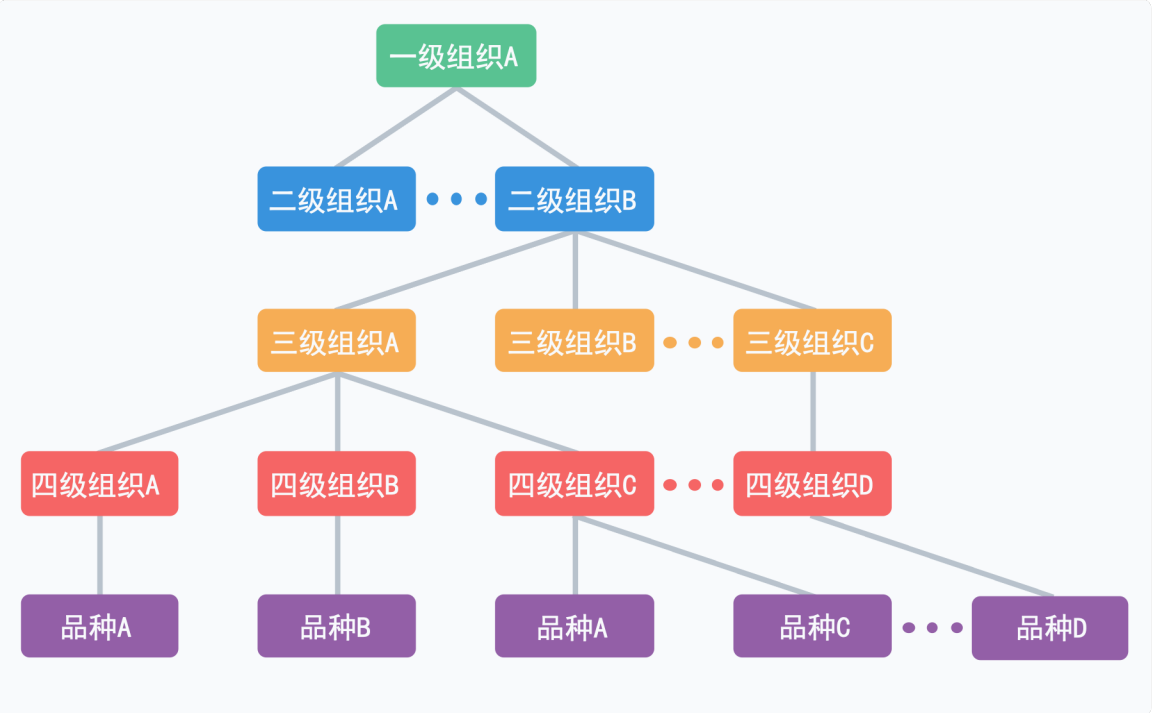


对于“大土 2 项”品种，模型预测的 7-9 月毛利与实际毛利的误差仅为 5%左右。

案例4-2 基于时空神经网络的毛鸡规划布局

应用背景

某养禽企业的毛鸡品种多样，且在全国范围内有多家子公司销售。由于产能有限，如何合理规划各子公司的毛鸡生产和销售布局，成为提升利润的关键。

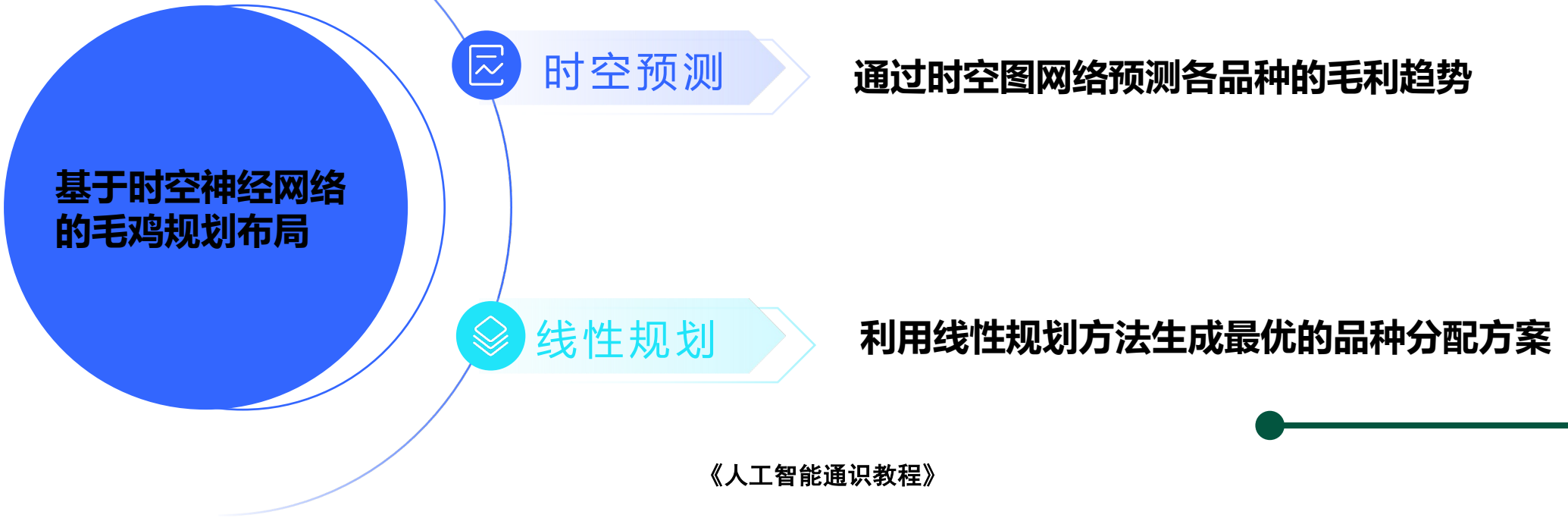


传统方法依赖人工经验， 容易出现区域效益不平衡、品种分配不合理等问题

案例4-2 基于时空神经网络的毛鸡规划布局

- **技术实现**

本案例旨在通过数据驱动的方法，优化该养禽企业的毛鸡生产和销售布局。在本案例中结合时空神经网络和线性规划，构建了一个两阶段的规划模型



案例4-2 基于时空神经网络的毛鸡规划布局

技术实现

- 数据预处理

依托该养禽企业的毛鸡销售数据进行建模，以该企业提供的各个子公司销售数据为订单记录，对历史销售数据进行清洗和汇总，以“销售品种”为粒度进行销量求和。

- 空间信息建模

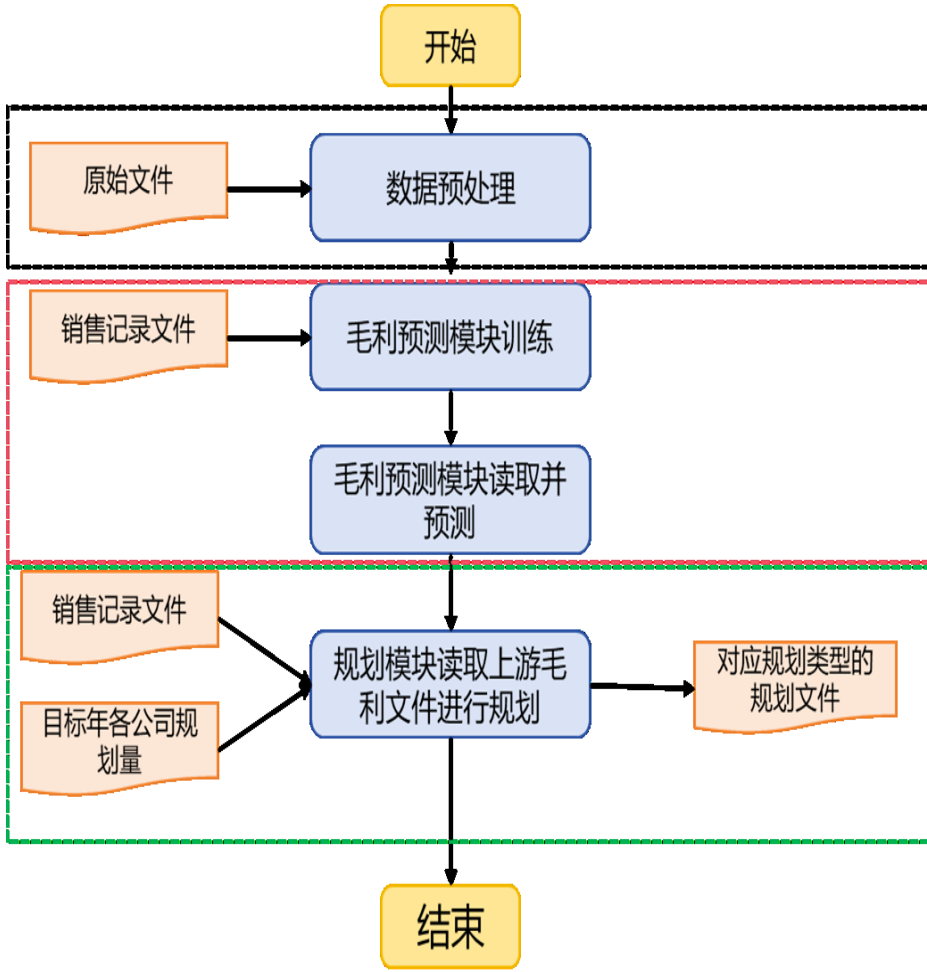
在案例 4-1 的基础上，引入空间信息，构建时空图网络模型。每个毛鸡品种对应一个节点，通过GAT捕捉品种之间的空间依赖关系。

- 时间信息建模

由图注意力网络得到的带有空间信息的高阶特征向量作为 LSTM 的输入，以捕捉时间序列中的模式，对其时间依赖性进行建模。

- 基于历史数据的规划布局

通过提取历史销售数据中的性别比例、月度销量比例等信息，构建约束变量上下界。



案例4-3 基于数据-知识双驱动的农产品辅助定价

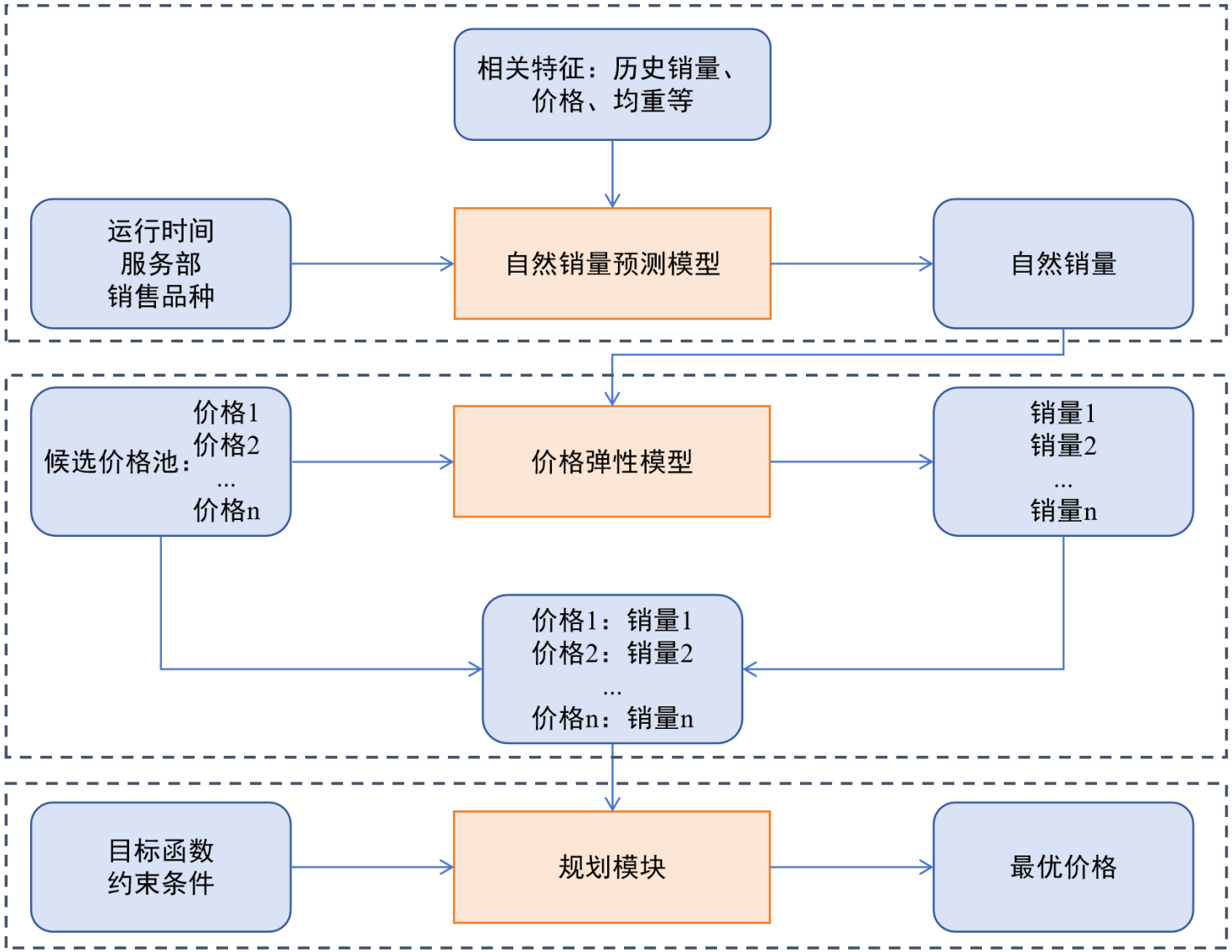
应用背景

- 人工定价：存在主观性强、效率低以及价格调整滞后等问题。
- 智能辅助定价：通过融合大数据分析与机器学习技术，实现价格动态优化，同步提升运营效率与经营收益。

技术实现

- 自然销量预测：首先，通过多种特征预测自然销量（即，不受人工干预的正常销量）。
- 价格-销量建模：引入价格弹性系数，建模价格对销量的影响。
- 建议价格生成：综合考虑销售带来的收益与积压带来的成本，给出最优的建议价格。

流程图


相关特征：历史销量、
价格、均重等
运行时间
服务部
销售品种
自然销量预测模型
自然销量
候选价格池：
价格1
价格2
…
价格n
价格弹性模型
销量1
销量2
…
销量n
价格1：销量1
价格2：销量2
…
价格n：销量n
目标函数
约束条件
规划模块
最优价格

案例4-4 基于多尺度信息提取的农产品精准营销

应用背景

本案例聚焦于to-B (Business)的农产品销售场景：

- 某个大型农产品公司从养户中回收可上市的农产品，并作为中间商将农产品销售给各个客户。这些客户基本上都是一些公司。
- 同时，每天滞销的商品会进入库存。该公司也会尽可能地通过降价促销等手段清空积压。

业务痛点

基于人工决策的传统农产品营销具有效率低和主观性强等局限性，具体来说表现为以下四点：

1. 难以通过客户的购买历史来确定客户的偏好是否变化，因此也难以拓展客户的业务；
2. 当某个品种积压时，难以精确定位促销目标客户；
3. 难以制定合理的降价策略，在保证收益的同时有效促销清库存；
4. 销售活动很大程度依靠销售员的经验水平，新员工难以快速上手，培养周期长。

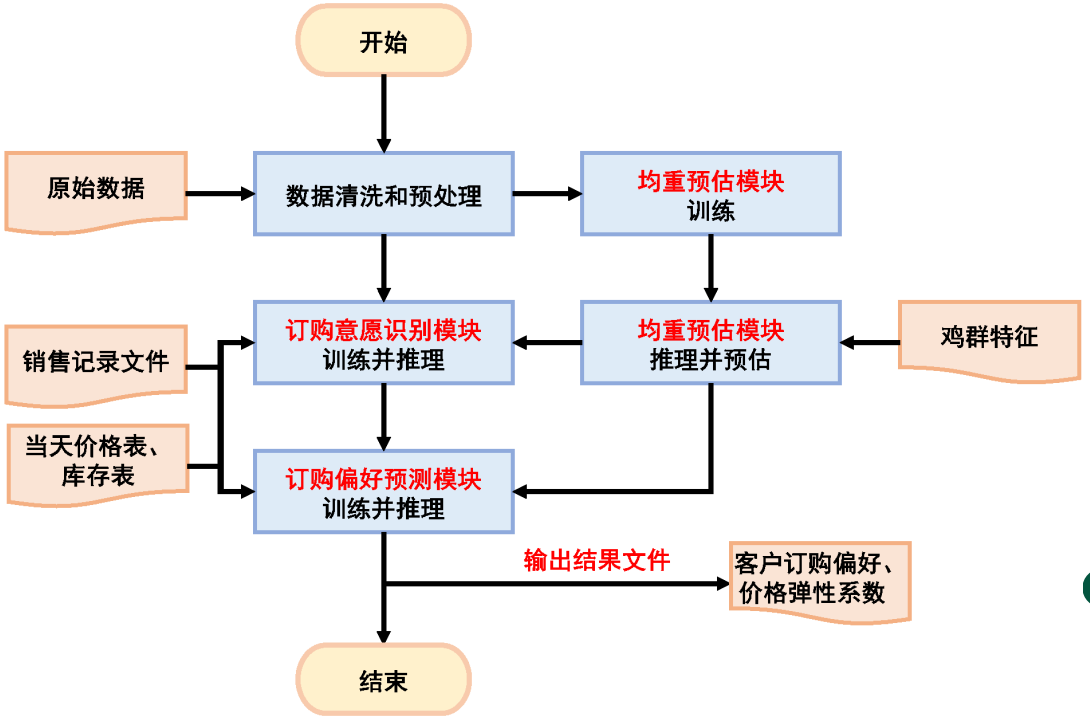
本案例基于多尺度信息提取的精准营销系统。通过分析客户历史行为、库存信息等多维度数据，实现客户采购倾向和偏好的精准预测，为相关农牧企业提供高效的营销解决方案。

案例4-4 基于多尺度信息提取的农产品精准营销

系统框架

本案例的核心任务是构建一个农产品精准营销系统，系统主要包含三大功能模块：

1. 销售品种均重预估：预测农产品在某一天龄的均重；
2. 客户订购意愿识别：预测客户在某一天前来订购的概率；
3. 客户订购偏好预测：预测客户在某一天订购某种农产品的概率。



案例4-4 基于多尺度信息提取的农产品精准营销

销售品种均重

本案例使用多层感知机(Multi-layer Perceptron, MLP)模型，预测特定农产品在给定天龄下的均重。

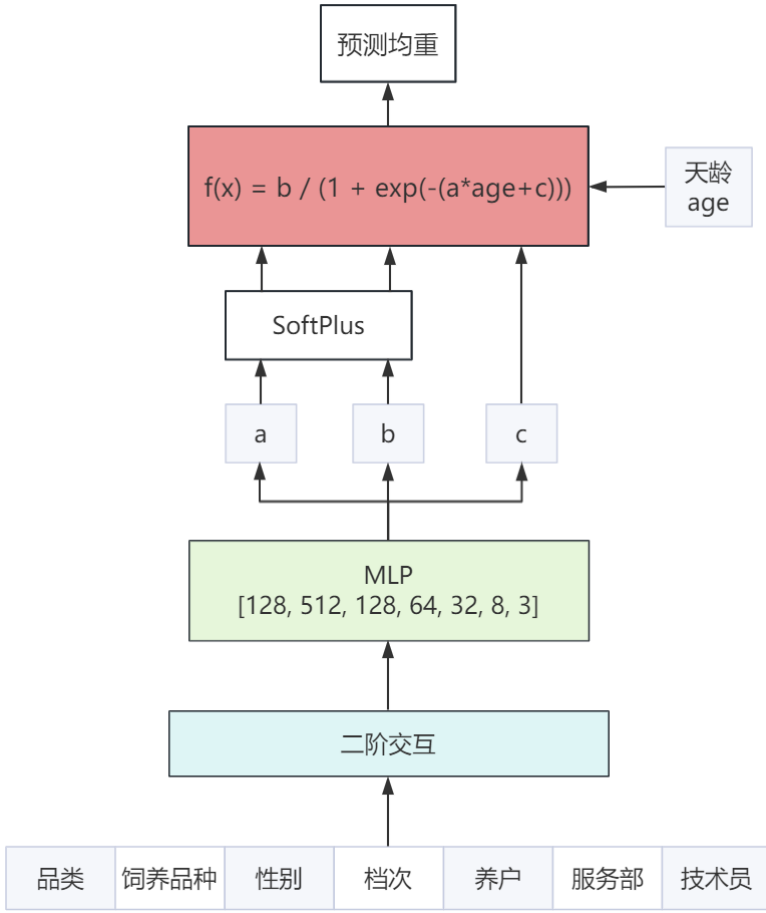
模型会考虑农产品的品种、饲养环境等因素，拟合S型生长曲线，预测未来某一天的均重。

技术实现

- 通过农产品的特征信息（如品类、性别、档次、养户等）拟合以下S型生长曲线：

$$y = \frac{b}{1 + e^{-(ax+c)}}$$

- 输入待预测的天龄x，得到预估的均重y。



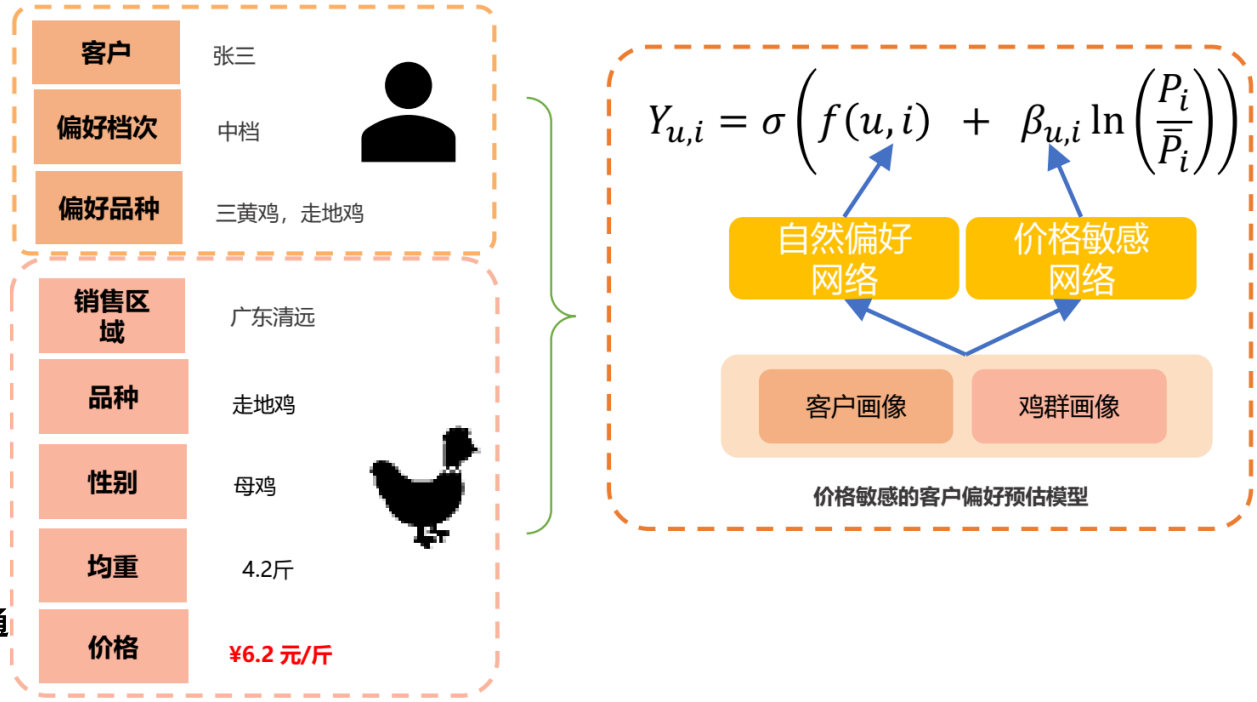
案例4-4 基于多尺度信息提取的农产品精准营销

➢ 客户订购偏好预测

本案例通过建模客户历史兴趣表征、客户显性表征以及客户与农产品之间的复购模式，来生成客户的表征，结合价格弹性系数，预测客户对特定价格、天龄、均重的销售品种的订购概率。

➢ 技术实现

- 把客户画像和农产品画像输入价格敏感的偏好预估模型中；
- 分别拟合客户对产品的自然偏好和价格敏感偏好；
- 整合后输出客户的总偏好。



《人工智能通

开放讨论

1．数据可视化的目的是什么？你觉得一张“好图”需要具备哪些特点？

答：数据可视化是将抽象数据转化为直观图形的过程，其核心价值在于降低信息理解成本、挖掘数据规律并辅助决策。数据可视化的目的是“让数据说话”，而一张好图需同时满足准确、清晰、相关、可读、有洞察力—— 既不扭曲数据，又能高效传递信息，最终帮助用户理解数据背后的意义。

开放讨论

2. 为什么要进行数据可视化？数据‘看得见’有什么意义？能不能用图替代所有分析？

答：数据可视化的核心价值在于解决“数据复杂性”与“人类认知局限性”之间的矛盾，让抽象信息变得可感知、可理解。数据可视化的核心是“让数据更易被理解和使用”，但它是分析过程中的“桥梁”而非“终点”。实际应用中，需结合“可视化发现规律→量化分析验证规律→文字 / 公式解释规律”的完整流程，才能实现从“看见数据”到“理解数据”再到“用数据决策”的闭环。

开放讨论

3. 你认为在农业中，数据分析最大的价值体现在哪个环节？（如种植、预测、销售等）为什么？

答：在农业领域，数据分析的价值贯穿种植、预测、销售等全链条，但在“种植环节”的价值尤为核心——这是由农业“生产端决定价值基底”的特性决定的。种植环节是农业价值创造的“起点”，数据分析在此环节的作用是将“靠天吃饭”转化为“数据驱动的可控生产”，从根本上提升农产品的产量、品质与资源利用效率。而预测、销售等环节的数据分析，本质是对种植环节产出的“价值放大”——若没有生产端的精准化基础，后续环节的优化将失去意义。因此，种植环节是农业数据分析价值的核心载体。

开放讨论

4. 大数据分析过程中，模型重要还是数据重要？你怎么看“两者的关系”？

答：在大数据分析中，数据与模型并非对立关系，而是 “地基与建筑” 的依存关系—— 数据是分析的 “原材料”，决定了分析的深度与边界；模型是 “加工工具”，决定了数据价值的挖掘效率与转化能力。两者的优先级会随场景动态变化，但本质上缺一不可。在实际分析中，先确保数据的 “完整性、准确性、相关性”，再选择适配的模型—— 这是更高效的路径。但最终目标是让两者形成闭环：用模型发现数据的缺陷，用更优的数据反哺模型迭代，如此才能持续释放大数据的价值。

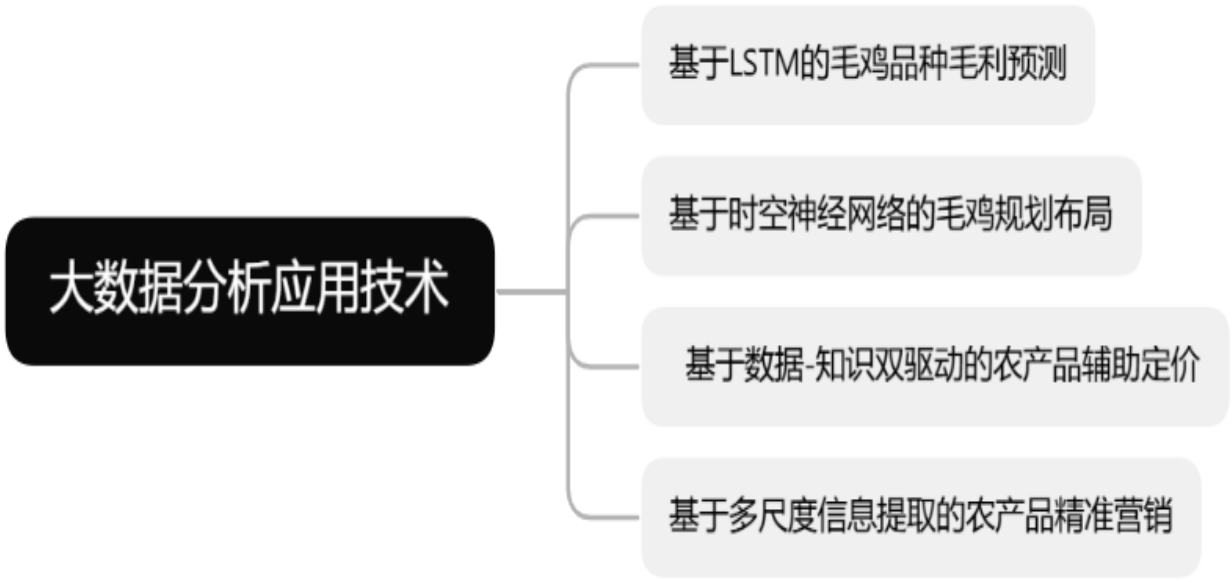
开放讨论

5. 除了课堂讲到的农业领域，你还能想到哪些行业或任务，适合用大数据分析方法来处理？为什么？

答：大数据分析方法的核心价值在于从海量、多维度、动态的数据中挖掘规律、预测趋势、优化决策，因此几乎所有存在 “数据积累” 和 “复杂决策场景” 的行业，都能通过其提升效率或创造新价值。例如：医疗健康行业，医疗数据具有高维度（生理指标、基因、行为等）、强关联性（某一症状可能关联多种疾病）、高价值（早发现 1 例癌症可节省数十万元治疗成本） 的特点，传统经验决策（如医生个人判断）难以覆盖复杂关联，而大数据模型能通过统计规律提升准确性。

课堂总结


数据可视化
折线图
直方图
泡泡图
箱图
小提琴图
柱状图
散点图
饼图


大数据分析应用技术
基于LSTM的毛鸡品种毛利预测
基于时空神经网络的毛鸡规划布局
基于数据-知识双驱动的农产品辅助定价
基于多尺度信息提取的农产品精准营销