数科导期末复习

Section 01 基础概论

1. 数据

- 计算机的数据：是表达信息的某种符号，是信息的一种量化表示
- 数据是信息的表现形式与载体，可以是符号、数字、文本、音频/语音、图像/视频
- 数字的产生：0 1分别代表低电位与高电位

1.1 数据类型

1.1.1 数据种类

- 数字数据
- 文本数据
- 音频数据
- 图像数据

1.1.2 数据形态

- 结构化的数据：
  - 数据集/数据表/工作表
  - 栏位/字段/属性/特征
  - 自变量/预测变量/解释变量
  - 因变量/被预测变量/被解释变量
  - 统计的数据
- 半结构的数据：有识别模式的文本数据文件，支持结构分析，例如HTML、XML文件
- 准结构的数据：这类文本数据带有不规则的数据格式，但是可以通过工具规则化（不一致的数据值和格式的网页点击流量数据）
- 非结构的数据：处理非结构化的数据是数据科学的核心工作，用于进行全数据的分析

1.1.3 数据的存在形式

- 全量数据
- 增量数据

1.2 数据同步

- 数据如果保留多份，就会存在一致性问题，就需要同步
- 数据同步分为两大类：全量和增量
- 全量同步，就是把数据从一个地方拷贝到另一个地方
- 增量的基础是全量，就是使用某种方式拷贝全量数据

1.3 数据来源

- 人工：填写、问卷、访谈
- 系统
  - 数据库：数据交换、日志
  - 信息系统：网页、EKP
- 爬取：网页、文本（专利、文献）
- 设备：机械设备（传感器）通信设备（信号记录）专用设备（眼动仪）

	数据分析	数据挖掘	数据科学
人工	很多	较少	一般
系统	很多	很多	很多
爬取	较少	较少	很多
设备	较少	一般	很多

2. 数据科学/分析/挖掘的目的

2.0 系统思维

- 数据是事后的结果
- 分析是一个科学推论的过程
  - 结果->假设规则->情况->确定规则
- 推论任务=找规则=建模

2.1 从“业务”的角度

- 类型上包括预测型分析、描述型分析、诊断型分析、指示型分析
  - 描述型分析：发生了什么?
  - 诊断型分析：为什么会发生?
  - 预测型分析：未来会不会发生?
  - 指标型分析：如何让它发生?
- 情景上包括事前分析、事中分析、事后分析
  - 事前分析：预测、建立指标、优化决策
  - 事中分析：监控、情况分析、建立指标、实时决策
  - 事后分析：情况分析、建立指标、优化决策

2.2 从“输出”的角度

- 部署系统包括仪表盘、子模块（代码嵌入系统）、研究报告（纸面可视化）
- 模型包括分类、聚类、回归（拟合）、关联等
  - 分类：类别概率统计，包括KNN、逻辑回归、朴素贝叶斯、决策树、支持向量机等
  - 回归（预测）：即值的预估，包括SVR、广义线性模型（Lasso、岭回归、贝叶斯回归等）
  - 聚类：将群体中的个体以相似度群组起来，包括K-means、AGNIS、DBSCAN、BIRCH等
  - 关联：频繁项目集挖掘，购物篮分析，包括Apriori，FP-growth

    - 数据简化：以一组较小的数据取代原本较大组的数据，包括因子分析、主成分分析
- 报告包括业务选题、数据描述、模型构建、结果展示

3. 数据科学的流程

- 问题界定->数据准备->数据建模->结果展示->知识应用
- 问题界定
    - 业务理解：理解业务、评估需求
    - 数据探索：数据收集、处理、检视
- 数据准备：数据检视、处理、分割
- 数据建模
    - 数据建模：选择方法/工具、建立模型
    - 模型评估：建模过程评估、模型结果评估
- 结果展示（应用）：结果应用、系统构建

Section 02 问题界定

0. 问题界定流程

- 业务理解：描述业务、构造问题、明确数据源
- 数据理解：数据检视、数据处理
- 定义指标：将数据通过指标与问题（目标）关联起来
- 项目评估：可行性评估（数据、时间、人力资源）
- 项目确认：问题界定报告
- 问题界定的核心：明确问题的本质、找出问题的关键因素、寻找因素可量化的指标

1. 业务理解

1.1 决策流程

- 了解问题背景
    - 调研决策情境
    - 探查利害关系人
    - 发现决策局限项
- 定义决策目标
    - 理解决策的目的
    - 决定决策的输出
- 因素分析
    - 理解业务的结构
    - 分解业务问题

- 属性分析
  - 明确因素的属性
  - 赋予可衡量的数值
- 综合评估
  - 提供解决方案

1.2 业务理解的重点

- 找出真正的问题
- 找出问题的关键因素
- 寻找现存因素的指标

1.3 分析方法

- SWOT分析法
- 5W2H分析法
- 鱼骨图分析法
- 逻辑树分析法
- 金字塔分析法
- 麦肯锡7步分析法
- 思维导图

1.4 决策任务的种类

- 特定任务：一次性任务
- 通用任务
  - 标准化任务
  - 信息化任务
  - 智能化任务

1.5 决策的输出

- 标准化任务的输出是为特定业务或产品制定固定工作流程或指标要求
- 信息化任务的输出是依据标准化的流程或指标制作信息系统
- 智能化是利用人工智能实现决策支持系统

1.6 因素及属性

- 因素：决定事物成败的原因或条件
- 属性：一个事物的性质或关系都叫做事物的属性
- 业务结构
- 可衡量的指标：定量变量&定性变量

2. 数据理解

- 明确数据源
  - 方法：数据库调研

    - 输出：图表绘制、报告
- 对象属性分析
    - 方法：关联对象属性
    - 输出：图表绘制、报告
- 数据集的分析
    - 方法：统计分析方法
    - 输出：图表绘制、报告

3. 定义指标

- 明确因素及属性的指标
    - 方法：鱼骨图分析法
    - 输出：鱼骨图绘制、报告

3.1 属性及指标尺度

- 因素的属性
    - 通过数据检视去理解数据
    - 探索每个属性的分布、范围、特殊点、趋势等（单变量分析）
    - 探索属性间的关系（多变量分析）
- 属性值：制定数据清理、拆合、转换、归约及生成的方案
- 制定指标：风险、效率、规模、速度、效益

3.2 指标的评估

- 验证标准
    - 专家打分法
    - 焦点小组法
    - A/B test
- 未来的灵活性

4. 项目评估

- 分析报告（书面）
    - 业务问题说明
    - 业务指标 结果展示
    - 新指标（挖掘）
- 仪表盘（系统）
    - 业务指标 结果展示
    - 新指标（挖掘）
    - 仪表系统（部署）
- 模型参数
    - 代码
    - 效能

    - 准确度
- 可行性分析

Section 03 数据准备

- 数据准备是指从原始数据到形成最终数据集的所有操作，确保数据能反映业务问题的情形下满足数据建模的要求
- 目的
    - 仔细理解数据
    - 去除无用的数据
    - 为数据建模阶段生成数据集
- 执行
    - 数据检视（宏观了解数据集、微观检查字段）
        - 利用统计方法深度理解数据
    - 数据处理（清理、拆合、转换、归约）
        - 形成更能反映业务问题的数据
        - 形成符合数据建模需求的数据
    - 数据生成（切割数据，形成训练集与测试集）
        - 产生建模需求的数据集
- 数据处理
    - 为了业务目的处理数据
        - 从业务的角度去理解数据，用业务的领域知识处理数据问题（缺失、异常、不一致等），视为“业务”数据处理
        - 包含清理、拆合、转换、归约数据
    - 为了建模需要处理数据
        - 从建模的角度去监视数据，依据优化模型的需要产生适用于建模的数据集，且不过分地遗失数据特征，视为“建模”数据处理
        - 包含清理、拆合、转换、归约数据

1. 数据基础

1.1 数据获取

- ETL(Extract 抽取 Transform 转换 Load 加载)是将业务系统的数据经过抽取、清洗转换之后加载到数据仓库的过程或工具，目的是将企业中分散、零乱、标准不统一的数据整合到一起

1.2 数据基本知识

1.2.1 数据的测量尺度

- 定类变量：值的表现是数字仅作为代码，数字大小不具有任何意义，也不能做数学运算，仅包含特定意义，表现的文字也算
    - 已婚、未婚
- 定序变量：值的表现是数字，且表现方案之间的大小顺序关系，但不能运算，表现的文字也算
    - 高、中、低

- 定距变量：值的表现是衡量的数字可有意义地描述并比较数字之间的差距大小，无固定原点。也可调整分隔的间距大小
  - 温度
- 定比变量：衡量的数字可做比率倍数的比较
  - 成绩、身高

1.2.2 定性变量与定量变量

- 定量变量，也是数值型变量
  - 是连续量
  - 由测量或计数、统计得到的量
  - 这些变量具有数值特征
- 定性变量，也是类别变量，并非真有数量上的变化，而只有性质上的差异
  - 有序变量：没有数量关系，只有次序关系
  - 名义变量：无等级关系，也无数量关系

1.2.3 连续变量与离散变量

- 离散变量
  - 离散变量的数值用计数的方法取得
  - 离散变量是通常以整数位取值的变量
- 连续变量：可取值能充满一整个区间的变量
- 为了方便统计或运算时，有些性质上属于连续的变量也按整数取值，即可以把它们当作离散变量来看待，例如年龄。

1.2.4 时间序列数据

- 时间序列数据：依据规律时间间距下连续观察的量测值
- 连续型时间序列
- 离散型时间序列
- 单变量时间序列
- 多变量时间序列

1.2.5 数据质量

- 数据质量是指数据符合使用目的，能满足业务场景具体需求的程度
- 在符合业务情境下适合建模的数据集

1.2.5.1 数据的“量”

- 样本个数
- 字段（定量）个数
- 属性值类型个数

1.2.5.2 数据的“质”

- 统计趋势、变异程度与特征覆盖业务的情况

1.2.6 数据精度

1.2.7 数据分布

1.2.8 数据维度

2. 数据检视

- 数据检视是探究原始数据的各类情况，判断能否达到业务问题与建模的可能要求，形成可行的数据处理方案
- 任务：
  - 厘清数据的质量
    - 记录、字段、精度、分布、维度
    - 准确性、完整性、一致性、时效性、可信性、可解释性
  - 判别系统要求
    - 硬件要求、软件要求、计算资源
  - 判别业务要求（可行性）
  - 判别建模要求（可用性）

2.0 数据检视的步骤

- 检视总体情况
  - 数据质量的分析
    - 记录与字段的统计描述
    - 中心趋势分析
    - 离散趋势分析
  - 实体、关联、属性的关系描述
- 探查字段情况
  - 描述
  - 类型（定类、定序、定距、定比、文本……）
  - 数值意义与格式
  - 统计分布
    - 单变量分析
    - 多变量分析
  - 存在问题
- 构建数据改造方案
  - 清理方案（缺失、空白、错误值等）
  - 转换方案（编码、过滤）
  - 折合方案（两个字段合并或相反）
  - 归约方案（降低维度）

2.1 检视总体情况

2.1.1 结构的分析：对象属性

- 基于物理意义的实体、属性关系分析
- 问题结构与数据结构的对应、解释程度、关联情况

2.1.2 描述性统计

- 记录分析
  *有效、无效、不确定的数据?
  - 数据够吗?
  - 是否有偏重?
- 字段分析
  - 业务对象有哪些字段
  - 对象属性有哪些字段

2.1.3 中心趋势分析

- 以”因变量“为主
- 均值、中位数、众数

2.1.4 离散趋势分析

- 以“因变量”为主
- 衡量指标
  - 极差=极大值-极小值
  - 四分位数
  - 观测数（总个数）
  - 异众比率=非众数/总个数
  - 方差、标准差
  - 变异系数（离散系数）=标准差/平均值*100%
- 分布的形状
  - 左偏分布：均值<中位数<众数
  - 对称分布：均值=中位数=众数
  - 右偏分布：众数<中位数<均值
- 对于偏态的分布，应使用中位值作为集中趋势，只有单峰和基本对称的图形，用均值作为集中趋势才是合理的

2.1.5 维度情况

2.2 检视数据情况

- 工作表
  - 表的数量、意义
  - 表内的记录数、字段数、有效数
  - 表内的关联
- 全数据集
  - 记录数、有效数
  - 字段描述、数值类型、意义
  - 基本统计描述
- 分析对象
  - 记录数、对象属性及其包含的字段
  - 基本统计描述

2.3 检视字段情况

- 单变量分析

    - 意义分析：依据对象属性来判断数值的情况
    - 统计分析：基本统计、类型、精度、分布
- 多变量分析
    - 意义分析：依据对象属性来判断多字段数值情况的合理性
    - 统计分析：多变量的相关分析
- 相似性判别：多变量的相关分析来探索问题结构与数据结构的对应、解释程度、关联情况

因变量	自变量	统计方法	目的
定量	定量	回归分析/相关性分析	自变量与因变量的因果依存关系或相关关系
定性	定量	方差分析	类自变量连续因变量之间的关系
	定量	协方差分析	连续自变量影响下类自变量与连续因变量的关系
定性	定性	列联分析	定类变量与定性变量之间的依赖关系
	定性	逻辑回归/聚类分析	定性或定量变量与分类变量之间的关系

2.4 检视异常数据

- 异常点检测（离群点检测）是找出 其数据值不如预期的检测过程，这些对象被称为异常的或离群点
    - 局部离群点、全局离群点、集体离群点
- 异常点检测方法
    - 标准差法：如果观测值与平均值的差值超过3倍标准差，那么可以将其视为异常点
    - 箱型图法
        - IQR：四分位点的内距，是离散程度的度量，通过用箱型图来观察
        - 设Q1和Q3为两个四分位点，通常“<Q1-1.5IQR”或者“>Q3+1.5IQR”的数据点视为异常点

2.5 数据改造

- 探索发现
    - 字段间相似性
    - 异常判别
        - 质量问题：缺失/空白、异常/噪声、不一致、离群点、冗余、精度过低、分布偏重、差异不显著
        - 编码问题：字段解释、建模需要
- 改造方案
    - 记录改造：样本（采样）、精度、分布偏重、维度
    - 字段改造：缺失/空白、异常/噪声、不一致、离群点、冗余、分布、编码、精度、建模需要

3. 数据处理

- 数据处理主要依据数据检视后进行的数据清理、数据拆合、数据转换、数据归约的工作，其目的在于将原始数据集改造成更符合业务需求、满足建模要求的数据集
- 从两个角度理解并处理数据
    - 为了业务目的去处理数据
    - 为了建模需要去处理数据
- 目的
    - 提升数据品质

- 提升数据萃取的价值
- 缩短数据建模时间
- 减少系统I/O代价和数据库容量

3.1 数据清理

- 数据清理的目的：发现、纠正数据中可识别的错误，包括检查数据一致性，处理无效值和缺失值等
- 发生的问题：缺失/空白、乱码/错位/异常/离群点、不一致、噪声/离群点、冗余/重复

3.1.1 缺失与空白的处理

- 忽略属性值或该记录（删除）
- 使用一个全局常数填充缺失值（填补）
- 使用属性平均值（填补）
- 使用与给定变量值属于同一类别的所有样本之平均值（填补）
- 使用最可能的值去填充缺失值（填补）
  - 回归
  - 最近邻补差（最近邻估计法）
  - 插值法（拉格朗日插值法、牛顿插值法、Hermite插值、分段插值、样条插值）

3.1.2 异常/不一致/冗余/重复的处理

- 异常：平滑处理
- 不一致：无法从数据角度解决（无法判断正确性）
- 冗余/重复：仅留一项记录数据，其余删除或利用两个字段的相关系数，来确认是否字段间存在冗余

3.1.3 噪声的处理

- 平滑处理技术：将数据平缓化，以降低其对结果的影响
- 分箱法：利用“近邻”值来局部平滑储存在同一箱子的数据值
  - 等深法：每箱数量相同
  - 等宽法：每箱同区间范围
  - 数据平滑：利用箱中的平均值、中位数、边界值等三种值进行
- 适配：回归、最近邻估计

3.1.4 离群点的处理

- 通过箱型图/标准差法/聚类/观察法找出离群点
- 做法：删除或用其他值替换
- 离群点在本阶段处理基本以数据处理方便为主

3.2 数据转换

- 将数据转换成适合的数据格式或为丰富化信息的内容
  - 重新编码（尺度不合适）
  - 主要执行字段内的数值处理
- 字段数值的转换
  - 归一化
  - 标准化

- 字段类别的转换
  - 离散性数据转成连续性数据
  - 连续性数据转成数值处理

3.2.1 字段数值的转换

- 对字段内的值进行规范化处理
  - 把数据映射到0~1或其他小范围之内处理，使数据处理更加便捷迅速
  - 把有量纲（物理单位）表达式变成无量纲表达式，便于不同单位或量级的指标能够进行比较和加权
- 如果不进行转换的话，要不维数过多增加了计算成本，或者数据过于集中很难找到数据之间的特征

3.2.1.1 归一化

- 将数据按比例缩放到一个特定的区间
- 可防止较大初始值域与较小初始值域属性间互相比较的情况，以及权重过大的问题，这会导致学习过程中，收敛过慢或者是不收敛的情况
- 极小值-极大值归一化：对原始数据进行线性转换，假设min和max分别为属性A的最小值与最大值，给定属性A新的范围newmax和newmin
  - $y = \frac{x-min_a}{max_a-min_a}(newmax_a - newmax_b) + newmin_a$
- 对数归一化：$y = log_{10}(x)$
- 反余切归一化：$y = \frac{arctan(x)}{\pi/2}$

3.2.1.2 标准化

- 条件：原数据应符合正态（高斯）分布
- 当属性A的最大值和最小值未知，或孤立点左右极小值-极大值正规化时，可改用标准化方法
- 量化后的特征将服从标准正态分布
- 基于属性A的平均值和属性A的标准差将数据标准化。A的值x标准化后为y，经由下式计算而得：$y = \frac{x-\overline{x}_A}{S_A}$（其中$\overline{x}_A$与$S_A$分别为属性A的平均值与标准差

3.2.2 字符类别的转换

- 字段内值的数据分辨度太低或太高

3.2.2.1 离散转离散

3.2.2.2 离散转连续

- 为每个字段确认适合建模的离散值的距离或相似程度
- 以类似的矩阵定义出数值与数值之间的距离或相似程度，再利用此距离或是相似程度把离散的数据转换成连续型的数据型态
  - Eg. 学生成绩A对应至90分，B+对应至85分
- 通常需要结合专家意见

3.2.2.3 连续转离散

- 为每个字段确认适合建模的离散值或相似程度
- 将连续数据分配到数个小区间（做法与量化相似），以类别尺度取代原有连续数据的尺度
- 数据离散化可以同时进行数据拆合与数据归约

3.2.2.3.1 直方图分析法

- 等宽法：在一个等宽的直方图中，每个箱的宽度（范围）是相同的

- 等高法：在一个等高的直方图中，每个箱中的数据个数是相同的
- MaxDiff方法：以相邻数值（对）之差为基础，一个箱的边界则是由包含s-1个最大差距的数值对所确定，其中s为用户指定的阈值
- 使用分箱方法对数据进行近似，每个箱代表一个区域范围内的值，箱的宽度代表值域范围，箱的高度代表这个范围内的值的个数

3.2.2.3.2 自然划分区段

- 将数值区域划分为相对一致的、更直观或更自然的区间
- 3-4-5规则：根据最重要数据的数值区域，递归地、逐层地将给定的数值区域划分为3，4或5个等宽的区间
  - 分成3块：一个区间最高有效位上包含3，6，7或9个不同的值，就将该区间划分为3个等宽的子区间（7分为2，3，2）
  - 分成4块：一个区间最高有效位上包含2，4或8个不同的值，就将该区间划分为4个等宽子区间
  - 分成5块：一个区间最高有效位上包含1，5或10个不同的值，就将该区间划分为5个等宽子区间
  - 将该规则递归地应用于每个子区间的概念分层

3.2.2.3.3 基于聚类的方法

- 聚类算法可以将数据集划分为若干个类或组
- 每个类构成了概念层次树的一个节点
- 每个类还可以进一步分解为若干子类，从而构成更低水平的的层次
- 当然类也可以合并起来构成更高层次的概念水平

3.2.2.3.4 基于熵的处理方式

3.2.2.3.5 连续型数据的转换

3.2.2.3.6 离散性数据的转换

3.3 数据拆合

- 不同字段或单一字段的处理，为了将字段的意义单一化，拆合的处理多是针对某个物理意义的数据拆合
- 拆，新增数据字段（原始数据字段隐含信息量太大，计算机不易得到有效特征），例如，购买的时间字段为“年-月-日”，拆分为三个字段。
- 合，符合业务需求的字段合并，减少维度，方便建模，例如，购买时间“年-月-日-时-分”，合并为平均多久购买一次
- 处理方式：
  - 依据物理量进行数据合并或数据拆解
  - 生成新字段后，处理方式与数据转换相同

3.3.1 实体识别

- 对象实体：数据集合中发现同一实体的不同描述
  - 数据集合中发现实体的描述，是根据业务问题决定的，需要领域知识
  - 解决方式：利用业务理解选择合适的分析实体
- 实体属性：在数据质量管理中可以用于错误检测、不一致数据发现等。或多个属性关系同时存在在一个字段内
  - 多个属性同时存在在一个（些）字段内，是需要从“业务”角度去理解发现

3.3.2 数据拆解

- 某些字段的意义隐含有更多的信息，对问题界定的角度可能都不同，有时候需要分解成不同维度的表示（多维度）

- 有些字段是可转换成数个易理解字段，方便建模（结构化）
- 分析过程中，有些必要构建时间序列的字段来新增信息（流程化）
- 数据拆解的处理方式包括字段拆解与字段赋值

3.3.2.1 字段拆解（字段间）

- 通过对象实体的物理属性自然分解：结构分析
- 依据业务流程切分出时间序列数据：流程分析

3.3.2.2 字段赋值（字段内）

- 对物理属性自然分解，并重新编码，通常以定类、定序为主
- 聚类法、直方图法、决策树等

3.3.3 数据合并

- 若一个字段可以从其他字段中推演出来，那这个字段就是冗余，则数据合并可减少建模的负担
- 字段间：对数据字段进行合并操作（数个字段意义相似）
  - 两个或以上实体不需要区分
  - 多属性的描述精度过高
- 字段内：
  - 名称不同，意义相似
  - 序数中，某些阶段不需要分开分析
  - 区分区间的精度不需过高

3.4 数据归约

- 原因：均匀数据分辨率
  - 数据太多
  - 精度太高
  - 分布偏重
  - 维度过高
- 目的：
  - 提升数据品质
  - 缩短数据建模时间
  - 提升数据价值、知识价值的获取
  - 减少数据库容量限制

3.4.1 减量

3.4.1.1 缩小可能的记录

3.4.1.2 采样技术

- 非概率采样：由人为主观选择样本，无法估计样本被抽出的概率，所以较难对样本做出推论
  - 方法：方便采样、配额采样、判断采样、滚雪球采样
- 概率采样：可客观地选取样本与估计样本结果的可信度，所以较可经由数理的推导，做出母体推论
  - 方法：简单随机采样、分层随机采样、群采样
- 随机采样：是最常用的方法并且在很多场合中出现，初始化时计算数据的随机样本，依据随机性采样数据

- 分层随机采样：利用已知的数据特征，将数据区分为数个不重叠的分层，使层与层之间的差异大，而层内的资料差异小，样本大小与各层数据成比例，对各分层随机抽取数据记录，样本具有代表性，较不易失去过多的数据讯息
- 分群采样：数据母体集合中依照已知标准或特征所排列的集群作为抽样单位，然后再依据要抽取的集群数量，选取抽样集群中所有数据作为样本，**当群间差异越小，群内差异越大时，抽出的样本越准确**
- 数据不平衡问题：在实践中，经常会遇到实际数据中正负样本比例不平衡的情况，也叫数据倾斜，导致了训练效率低下和简单的负面样本会引发整个模型表现下降的问题。
  - 解决思想：获取更多的少量样本数据
  - 解决做法：过采样、欠采样
- 随机过采样：过采样是为少量样本增加样本数，随机过采样，采取简单复制少样本的策略来增加少样本的数量
  - 但这样容易产生模型过拟合的问题，即使得模型学习到的信息不够泛化
- 随机欠采样：欠采样则是丢弃大样本一方大量的数据，随机欠采样，则是从多样本的群体中，随机采样出与少样本差不多数量的样本作为策略
  - 和过采样一样会存在过拟合的问题

3.4.2 缩值

- 以符合业务需求为主
- 做法
  - 分箱与平滑处理
  - 数据聚类
  - 量化编码

3.4.3 降维

- 主要针对建模角度进行
- 特征选择法：特征子集选择或属性选择
  - 以因变量作为比较基准，删除与因变量不相关的变量
  - 特征选择实质是从原始数据集中选取最优子集的过程
- 特征提取法：主成分分析法PCA，将变量作线性变换，只留下提供较多信息的几个主成分，藉以缩小变数维度，此法不需要目标变量作为比较基准，目的在于找出最能解释数据变异的线性组合

4. 数据生成（数据分割）

- 训练组：建立模型
- 测试组：评估模型
- 验证组：衡量模型的好坏，如分类错误率、均方误差
- 数据分割的比例有不同的定义，均应代表原来的数据
  - 80%用于训练模型，20%用于测试模型
  - k-折叠交叉验证（k-fold交互验证）

5. 数据准备最后的输出

5.1 报告

5.1.1 数据现状描述

- 记录与字段的统计描述（字段的意义叙述、类型、值分布情况）
- 业务问题与数据的对应（即最后数据表说明）
  - 可以用E-R图、对象图与业务问题对应说明

5.1.2 过程记录

- 清理、拆合、归约、转换记录

5.2 数据集

- 训练数据
- 测试数据
- 验证（通常会省）

6. 数据准备之可视化

6.1 探索性数据分析

- 探索性数据分析 vs. 验证性数据分析
  - 验证性数据分析偏重模型与假设
  - 探索性数据分析没有假设没有模型，对待解问题的理解会随着烟酒的深入不断变化

6.2 一个变量的分析

6.2.1 描述性统计

- 最大值、最小值
- 众数
- 平均值
- 中位数和其他百分位数
- 箱型图、小提琴图

6.2.2 分布

- 左偏分布、对称分布、右偏分布
- 表格、柱状图
- 直方图
- 密度分布
- 累积分布

6.3 两个变量分析

- 观察散点图
- 相关系数
  - Pearson 相关系数：连续数据、正态分布、线性关系
  - Spearman相关系数

6.4 三个或三个以上变量的分析

- 散点图进化为珍珠图
- 多个图两两之间分别看
- 可视化：蜂巢图、等高线图、热力图

6.4.1 回归分析

- 在做回归之前：
  - 每个变量的分布特点
  - 每组变量之间的关系
  - 有没有特殊点
  - 想建的模型、研究问题和数据的现状吻合吗
- 需要注意的问题
  - 变量的多重共线性：使用VIF检验等
  - 内生性问题：遗漏变量、选择性偏差、双向因果关系
- 考察：因变量服从什么分布
  - 正态分布：OLS回归
  - 计数变量：泊松（方差近似等于均值）或负二项回归（方差远大于均值）或对应的零膨胀模型（“0”很多）

Section 04 数据建模

1. 基础概念

1.1 数据建模

- 找到合适的模型
- 经常使用缺省（预设）参数来运行多个模型，然后再对这些参数进行微调或回到数据准备阶段以便执行所选模型所需的操作
- 建模时通常会执行多次迭代
- 建模过程就是调参过程，是经验的累积

1.2 文本数据的结构化

1.3 图像数据的结构化

1.4 数据与模型

2. 数据建模的流程

2.1 选择模型：挑选符合建模目标的全部模型

- 确定建模目标
  - 条列所有可用的算法，检查是否具有足够的数据为给定的模型生成可靠的结果。
- 清点建模要求

    - 条列所有可用的算法，调研给定的模型所需要特定的数据质量级别，检查当前数据是否能达到这一级别
- 选择可能的模型
    - 筛选出所有符合目标与要求的模型
    - 开始缩小建模工具选择范围时，记录决策制定过程

2.2 评估标准：定义模型评估标准

- 依据准确率（模型效能）
    - 估计特定模型的错误率
    - 检全率、检准率、错误率、MSE等
- 依据运算效率（计算机效能）
    - 包括易于部署或所需处理时间等标准
    - 在特定系统硬件环境要求下的运算效率
- 依据解释性（领域知识）
    - 易于解释、能解释的程度（偏主观）

2.3 训练模型：参数选择、训练与测试模型

- 参数设置
    - 大多数建模技术都具有大量参数或设置，对这些参数和设置进行调整即可控制建模过程
    - 记录所有参数的制定过程
- 运行算法
    - 执行代码、运行训练数据与测试数据、返回模型结果
    - 记录当时参数（模型）、效能（运行时间）、结果（数据）

2.4 决定模型：挑选最合适的模型

- 基于评估标准检查所有模型结果
- 是否达到在问题界定阶段建立的目标
- 考虑某个模型的结果是否易于部署

3. 模型基础

3.1 任务数据（？）

3.2 建模三要素

- 对数据样本、特征间关系进行一定的约束假设，方便构建有限范围内的模型并求解，常见的假设有如下
    - 独立同分布假设：样本之间相互独立而且来自于同样的分布
    - 线性可分假设：多条样本可以在高维空间中找到超平面完美分割
    - 特征条件独立假设：特征之间不存在相互条件相互关系
    - 正态分布假设：特征与因变量之间服从标准正态分布
- 建模方法=模型+策略+算法
- 模型：x和y之间关系的函数设计
- 策略：在函数空间中，“寻找”最适合当前数据的模型的准则
- 算法：寻找函数空间中最优模型的计算方法

3.3 模型

- x和y之间的数学假设为$y = f(x; w)$，其中w为未知的函数参数模型，f构建了x和y之间的关联，有的函数包括超参数h，需要人工指定。
- 函数假设空间：
  - 取不同的w，函数将有不同的形式，而所有w取值范围对应的f的空间叫做函数假设空间
- 机器学习则是需要通过算法寻找对应某一套w能够满足我们对策略的要求
- 参数模型 vs. 非参数模型
  - 参数模型：假设了x和y之间的关系是一个带有参数集w的函数f，参数模型通过计算通用参数获得预测能力
  - 非参数模型：通过已有x之间的相互关联进而推断新的x对应的y，典型的方法包括KNN和Kernel核方法，非参数模型依赖于记忆以后x数据获得预测能力
- 生成模型 vs. 判别模型
  - 生成模型：假设x与y关系式认为y为因，x是如何由某些未知的y生成得到的，典型的生成模型即朴素贝叶斯
  - 判别模型：假设x与y关系的时候认为x为因，y通过对x进行一系列计算获得，典型的判别模型是逻辑回归

3.4 策略

- 策略是评估某一套参数设定下的模型对当前数据集优劣程度，是优化模型的准则
- 损失函数与风险函数
  - y与模型结果可能不一致，那么度量错误的程度需要损失函数
  - 损失函数，度量一次模型预测的好坏
  - 风险函数，度量平均意义下的模型预测的好坏
- 常见损失函数
  - 0-1损失函数
  - 平方损失函数
  - 绝对值损失函数
  - 对数损失函数
  - 指数损失函数
  - 铰链损失函数
- 风险函数
  - 经验风险函数
  - 结构风险函数

3.5 算法

- 算法是求解模型的具体方法，根据策略我们已经将原问题转化为模型在假设空间中优化问题，算法则是指此优化所需要的方法
- 算法应用场景
  - 模型拟合和预测过程中包括一系列计算，命名为“学习” “推断” “预测”
  - 算法则是应用于这些环节的具体方法
- 常见算法
  - 最大期望算法
  - 前向后向算法、HMM等序列模型的推断算法
  - 梯度下降、拉格朗日乘子法等优化算法

3.6 模型评估

3.6.1 有监督任务

- 有监督任务一般拆分数据集为训练集、验证集、测试集
- 验证集用于训练时检验模型效果，测试集则是模型拟合结束之后进行说明效果
- 分类常用指标：准确率、精准率、召回率、F-score、混淆矩阵、AUC
- 回归常用指标：均方误差、平均绝对误差、R^2

3.6.2 无监督任务

- 无监督任务需要看是否存在标注，有的聚类任务存在已有的人工标注
- 一般来讲无标注信息的无监督任务难以评估，应结合具体问题具体分析
- 聚类问题的评估常用方法：Rand指数、互信息、轮廓系数

3.6.3 混淆矩阵&分类评估

- 准确率
- 精准率
- 召回率
- F-score
- ROC
- 多元分类问题
- 指标选择问题

3.6.4 回归评估

- 均方误差
- 平均绝对误差
- R^2
- 指标选择问题

3.7 注意事项

- 任务确定：对任务类型的确定，分类、聚类、拟合、回归等
- 模型选择：多种模型具有不用特性和假设，根据任务和实际对数据理解判断选择模型
- 数据处理：实际数据分布与模型假设数据分布存在差异，或数据类型本身比较复杂，需要进行数据处理（回到数据准备）
- 数据准备的缺失
  - 不平衡数据集
  - 极端值敏感性
  - 仍须重新数据准备

3.8 决策树

3.8.1 思想

- 树模型：利用一系列特征值分割进行分类回归的模型
- 完成分类任务：树中的每一个节点代表一个类别
- 完成回归任务：树中的每一个节点对应一个预测值，多个节点之间的值可能不同。对比线性模型等，树模型的结果更像是分段定值函数

3.8.2 划分准则

- 常用划分准则
  - 信息增益
  - 信息增益率
  - Gini系数
- 使用划分准则
  - 输入：数据集，某一个属性
  - 计算：对当前数据集所有属性值依据划分准则计算
  - 输出：基于此属性分割当前数据是否能得到较合适的子结点

3.8.3 信息增益

- 信息熵
- 条件熵
- 信息增益

3.9 K-means聚类

- 思想：给定K，算法是将数据构建成K个互不相交的类，并通过反复迭代以改变分组的重定位技术，使得每一次改进之后的聚类方案都能够比前一次好
- 类型：原型聚类、密度聚类、层次聚类
- 目的：使各节数据点到所属类中心的总距离差异平方和最小

3.9.1 K-means聚类流程

1. 决定K，在数据集中随机选取K个数据点作为K个起始类的中心
2. 计算剩下的数据点与各“类中心”之间的距离，选距离最近的类，并根据类中的数据点，重新计算各类中的均值（或距离差异平方和）
3. 若发现总距离差异平方和与上次的不同（下移），则表示类中心有所改变，需重新计算类中心
4. 直到总距离差异不再下降或达到所设定的计算次数为止

3.9.2 K-means聚类特点

- 数据分布
  - 仅适用于数据分布是呈圆形（或者高维球形）分布的，但在实际应用中，很少能有这种情况
  - 仅适用于“数值型”的数据类型，不能直接处理“类别型”数据
- 优点
  - 原理简单（靠近中心点），实现容易
  - 效果中上（根据K的选择），容易解释
- 缺点
  - 须事先决定类数目K
  - 结果不稳定（受输入顺序影响）
  - 易受到离群值或噪声影响

Section 05 结果展示

1. 数据可视化

- 可视化：利用计算机图形学和图像处理技术，将数据转换成图形或图像来显示并进行交互处理的理论、方法与技术
- 数据可视化的目的：
  - 获取摘要信息
  - 提升理解程度
  - 容易阅读

1.1 可视化编码

- 数据可视化的核心内容使可视化编码
- 可视化编码：将“信息”映射成可视化元素，包括几何标记（图形元素）和视觉通道
- 几何标记：通常是一些几何图形元素，例如，点、线、面、体
- 视觉通道：用于控制几何标记的展示特性，包括标记的位置、大小、形状、方向、色调、饱和度、亮度等
  - 定性（分类）的视觉通道：如形状、颜色的色调、空间位置
  - 定量（连续、有序）的视觉通道：如直线的长度、区域的面积、空间的体积、斜度、角度、颜色的饱和度和亮度等

1.2 数据可视化流程

- 定义业务问题（子议题）
  - 要说明的问题（原因、过程、结果）
- 明确展示目标
  - 定义现在任务所能提供的信息：关联、顺序、构成（比例）、分布、位置、对比、重点
  - 重点体现出支持的结论
- 选取数据
  - 数据范围（个体、时间、位置）与字段属性
  - 有可能依据图表设计的需求进行数据转换
- 决定图表设计（视觉通道）
  - 编码、类型、维度、色彩、形状、位置

1.3 常用数据可视化的图表

- 联系：散点图、气泡图、网格图
- 分布：直方图、地图、散点图
- 构成：面积图、饼图
- 比较：条形图、雷达图、折线图、单词云

1.4 数据可视化的优点

- 帮助信息降维
- 引导左右脑共同解决问题
- 容易理解数据
- 有信息的客观决策
- 提升沟通效率
- 让人更愿意阅读

2. 结果展示

- 展示的重点
  - 数据描述
  - 业务问题
  - 问题解答
  - 重点发现
  - 模型能力

2.1 数据描述

- 点（特殊点）
  - 极值、平均值、离群点、特殊意义点
  - 有意义的数值或记录的单位
- 线（趋势）
  - 趋势线、数据分布
  - 两类数据（同属性）的趋势比较
- 面（关系）
  - 趋势面、趋势线、数据分布
  - 多维关系的表示

2.2 仪表盘

- 目的
  - 检测
  - 控制
  - 决策
- 应用场景
  - 整体状况（概览）
  - 关键指标的运行情况（关键细节）
  - 哪些指标出现异常（报警）

2.3 结果展示的任务

- 依据业务问题提供数据支持的关键论点
- 用易于理解的可视化方式描绘数据结果
- 依据项目需要实现可视化

2.4 结果展示的目的

- 给读者产生报告
  - 述说业务问题的来龙去脉，以议题树的结构去组织论点，并使用数据的可视化去支撑论点
  - 重点在于说服而不在于模型
- 产生仪表盘给决策者
  - 在特定的空间范围内，让决策者能快速了解整体情况，并发现问题。找出关键事件、关键节点（指标）、关键数值，以最易于察觉的可视化去呈现
  - 重点在于关键数值而不在于美观

2.5 结果展示的工作流程

- 确定展示目的：根据“问题界定”的业务问题与目标作为展示目的，决定展示关键点（故事点）
- 结果展示方案的制定与实现：构建故事线，突出关键、重点叙述、数据支撑，数据可视化的视觉效果，制作图表
- 系统开发（项目决定）：开发软件程序、部署嵌入系统

Section 06 数据的发展趋势与培养方案