统计.

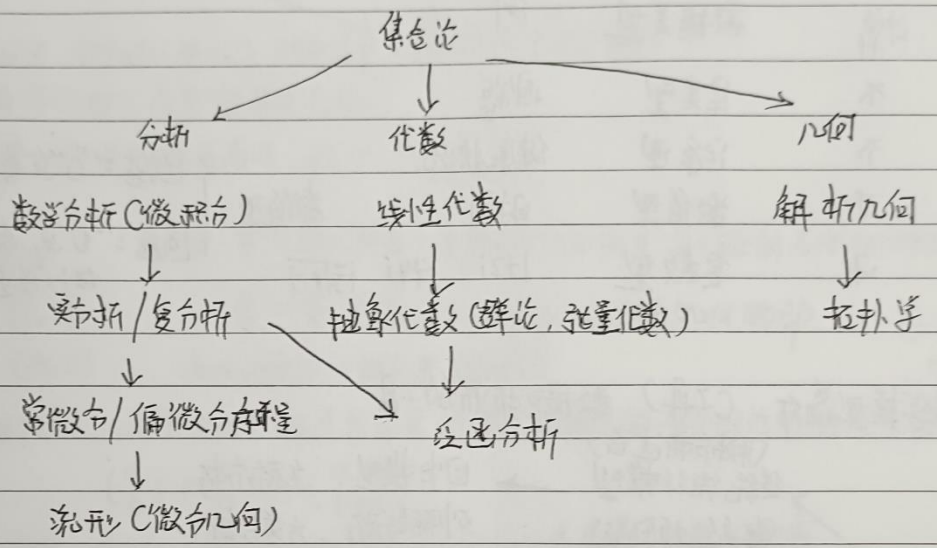
第一天：预备知识，线性函数

第二天：函数，微积分

第三天：数据度量，抽样分布，参数估计，假设检验

第四天：列联表分析，相关分析，回归分析

第一节：数学概况



第二节：数据类型

离散型数据.

离散随机变量是指一个可取有限个或可数无限个数值的随机变量。

通常用泊松型来描述.

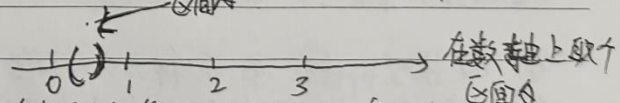
连续型数据.

连续随机变量是指一个取任何实数的概率都为零的变量。

通常用几何概型来描述.

$\frac{1}{\infty}=0$ 是没办法通过数频数计算的

例：数轴



离散 N(全体自然数) 如果区间A有可能不包含这个数集里的任意一个元素，那么此数集为离散数集

连续
- 稠密 Q(全体有理数) 无论区间取在哪，无论区间大小，此区间都至少含有这个数集里的元素，那么这个数集为稠密数集
- 连续 R(全体实数) 无论区间取在哪，无论区间大小，此区间都至少含有这个数集里的元素，且此数集是连续的，那么这个数集是连续的

(可以无限细分)

离散型数据直接数频数就可以计算出来，

数据 — 但带上时间属性就必然携带了某种相关性（过去与未来的相关性）

数据类型

横截面数据：在一个时间点或者一个时间段内取到的所有数据叫做横截面数据

时间序列数据：随着时间会发生变化的数据，默认只分析一个序列

面板数据：既有时间属性、又有空间属性的数据。{时间属性 / 空间属性}

例：去年企业在各个省份的销量

排序	计算	数据类型	例
不	不	定类型	国籍
可	不	定序型	健康状况
可	可	数值型	时间
不	可	复数型	$1+2i$　$3+4i$　$i=\sqrt{-1}$

数值型 { 定距："0"点有意义 / 定比："0"点没有意义 例："体重""身高" }

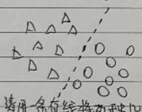
第二节 数学模型简介（工具）数据分析师的工具

数据分析模型

- （用来分析过去）传统统计模型（解决结构问题）→ 回归模型、主成分分析、列联表分析、方差分析，等等等等……
- （用来预测未来）数据挖掘模型（解决预测问题）→ 决策树、神经网络，支持向量机、贝叶斯网络，关联规则，等等等等……

支持向量机

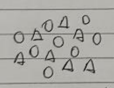
例



支持向量机：本质上为二分类问题模型

多个二分类模型拼接在一起，可以解决多分类问题　请用一条直线将两种几何类型分开

例：人类识别：把中国人根据脸部特征平均分成14亿类。



把低维数据投射到高维空间

但是非线性分类怎么办。

现代数据分析

把数据放到一个数据空间中，通过对该空间的各种形变与分析，从而挖掘出数据最底层所携带的信息。

线性代数

1. 向量

物理：有方向的量　统计学：空间中的点　计算机：数表

数学里，我们可以将以上三种向量的概念合并到一起。

通常我们默认向量的起点在原点。

向量一般用小写字母表示，如 α, β, μ, ν。

向量的坐标。数表中，第一个数（标量）告诉我们从原点沿X轴的方向如何移动。第二个数告诉我们从原点沿Y轴如何移动。

$[-2,3]$　$[1,2]$　$[2,-1]$　$[4,0]$　此为2维空间的向量的坐标　$[-2,3,4]$ x Y Z

向量的加法：2维空间内，就是相给定2个向量所围成的平行四边形的对角线。

$$\begin{pmatrix}a\\b\end{pmatrix}+\begin{pmatrix}c\\d\end{pmatrix}=\begin{pmatrix}a+c\\b+d\end{pmatrix}$$

向量的数乘：将给定向量按比例缩放（拉伸），负数表示反向拉伸。

$$\alpha\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}\alpha x\\\alpha y\end{pmatrix}$$

补充：向量的减法

线性组合与向量空间

线性组合

将一个向量组中的向量做数乘后相加，即得到该向量组的一个所谓的线性组合。

定义如下：空间V中的一组向量 $v_1, \cdots v_n$ 的线性组合是指形如 $a_1v_1+\cdots+a_nv_n$ 的向量，其中 $a_1, \cdots, a_n$ 为常数。

例：在 R^2 空间中，$(6,1)$ 是 $(2,3)$，$(2,-1)$ 的线性组合，因为 $\begin{pmatrix}2\\3\end{pmatrix}+2\begin{pmatrix}2\\-1\end{pmatrix}=\begin{pmatrix}6\\1\end{pmatrix}$

生成空间。以二维空间为例：在线性组合中，给定两个非零向量。一个向量固定，另一个向量自由变化，其线性组合可得到一条直线。

但是，当两个向量都自由变化，可以得到一个平面吗？

大多数情况下确实如此，除非两向量共线。

统计学分析数据的方法

- 描述性分析。研究数据收集、处理和描述的
  总体规模、对比关系、集中趋势、离散程度、偏态、峰态、……
- 推断性分析。研究如何利用样本数据来推断总体特征的统计学方法
  估计、假设检验、列联分析、方差分析、相关分析、回归分析……

统计学的基本概念 — 数据

结构化数据：指所有可以用数字来表示

非结构化数据：不可以直接通过数字来表示

非结构化 转成 结构化 → 再进行分析

1. 统计学的对象是数据

2. 数据的形式 {
   - 数字（结构化数据）可以进行比较、加减乘除等运算，严格的数据符号需有一定的数字表示
   - 文字（非结构化数据）不可运算，如男、女等

   正则表达式（专门用来处理文字）→ 词模型 }

只要能携带信息的东西都是数据。注：没有信息也是信息

例：缺失的东西告诉我们某东西是缺失的

数据的分类

1. 按照计量尺度分类 {
   - 分类型数据：对事物进行分类的结果，如人的性别分为：男女。
   - 顺序型数据：对事物类别顺序的测度，如产品分为：一等品、二等品、三等品
   - 数值型数据：对事物的精确测度，如身高为：175cm，180cm }

2. 特点：

- 分类型数据：不可排序，不可计算
- 顺序型数据：可排序，不可计算

→ 定性数据　低级

- 数值型数据：可排序，可计算　定量数据　↓ 高级

PS：复数型数据　不可排序，可计算　课外知识；

例：$1+2i$　$2+3i$　$i=\sqrt{-1}$

虚数单位 i

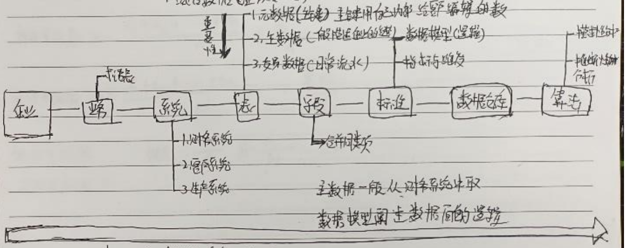
虚复数的极限收敛于一个平面内

顺序型数据：例：消费者行为偏好，

领研、paperweekly　arxiv.org（前沿文章）　中国国家数据中心

数据的其他分类

1. 按来源分 { 直接来源（一手数据、原始资料） / 间接来源（二手数据、次级资料） }
2. 按收集方式不同 { 观测的数据 / 实验的数据 }
3. 按与时间的关系不同 { 截面数据：在一个时间点或一个时间段内分析数据 / 时间序列数据： / 混合数据（面板数据）： }
4. 按类型分 { 离散型数据 / 连续型数据 }
5. 一种特殊的数据：虚拟变量数据



构建数据仓库流程。

构建数据库 { 业务能力 / 数据库能力 { SQL / 自动化 python（爬虫） } }

PS. 时间序列数据分析有争议

时间序列统计方法 {
- ARIMA（传统）— 宏观的趋势
- 人工智能分析方法（LSTM）— 微观 例：人工智能 }

实际问题中 宏观用 ARIMA

微观用 LSTM 例：自动问答机器人

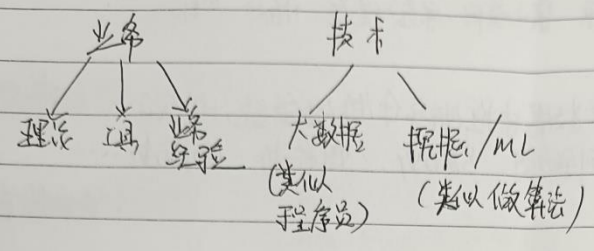
LSTM（长短期记忆网络）

统计学的基本概念 — 总体和样本

SQL熟练度

统计的基本概念

描述，推断，回归，实践



统计学定义

统计学是一门收集，处理，分析，解释数据并从数据中得出结论的科学

数据是怎么来的　　不但出　　数据接入

数据分析步骤　　核心：数据

收集数据 → 处理数据 → 分析数据 → 解释数据

描述性统计分析　　推断性统计分析

收集数据阶段

数据埋点：获取用户在网站上的行为数据　　(根据业务理解进行埋点)

取出流水线上关键点的数据

爬虫：(爬虫当有法律风险)

数据接入：从各个不同的系统中去取数，同时可以接入外部系统 (数据统计口径和中台)

最后接出的左边是一张表，如果是三张表，就没有什么意义

↓

倒：~~填缺失值~~

处理数据

↓

分析数据

↓

如果是对行数据的处理，一般称为数据清洗

如果是对列数据的处理，一般称为特征工程

一列为一字段　　一列为一变量　　一列为一特征

一列为一"维度"(数学中)

特征工程用的对可以优化算法

解释数据：某些数据是有规定的语法