Data analysis

1. 看懂数据集
2. 清洗数据（去掉 NAN，去掉重复行列）
3. 数据准备，多加一列来分类，（例：小于 18 岁为青少年，18 岁到 60 岁，成年人，大于 60 岁老年人）
4. 题目给出要求分析

概念：

相关程度：变量之间的相关性有多高

Regression：建立一个数学方程来预测

分数的度量：

R 平方：拟合度，越高越好（回归模型的 score）

```
from sklearn.metrics import accuracy_score,f1_score
```

准确率分数：分类准确率

F1 分数：二分类的指标，兼顾精确率（正确的分类÷提取的样本总数）与召回率（正确的分类÷样本总数）

假设验证：

零假设：假设 A 与 B 没有统计学差异（xxx 相等）

对立假设：有统计学差异（xxx 不等）

做 t-test，算出 t 值与 p 值，t 为正，决定 A 的大于 B，t 为负，决定 A 小于 B

P 小于 0.05 拒绝零假设

卡方分析：看两类别变量是否有关系

```
from scipy.stats import chi2_contingency
from scipy.stats import chi2
table = pd.crosstab(df['xxx'],df['xxx'],margins=True, margins_name='xxx')
print(table)
stat,p,dof,expected = chi2_contingency(table) # stat 卡方统计值，p: P_value，dof 自由度，expected 理论频率分布
print('dof=%d'%dof)
print(expected)

prob = 0.95 # 选取 95%置信度
critical = chi2.ppf(prob,dof)  # 计算临界阀值
print('probality=%.3f,critical=%.3f,stat=%.3f '%(prob,critical,stat))
if abs(stat)>=critical:
    print('reject H0:Dependent')
else:
    print('fail to reject H0:Independent')
```

线性回归：

y=Ax+b,

y 可以是一个数，也可是个向量，

x 可以是一个数，也可是个向量，

b 可以是一个数，也可是个向量，

x 一般是一个向量，多元线性回归，x 是 predictor，y 是 response

多项式拟合 y=a0x^n+a1x^(n-1)+…+an，n 是自己看图定的
套用 y=ax+b,将 x 变为 x^n,再用这组数进行线性回归

Machine Learning

监督学习（supervised）：有标签，机器知道他做得对不对 （regression，classification）
无监督学习（unsupervised）：没标签，不知道对不对 （clustering）
半监督学习（semi-supervised）：有些有标签有些没有
强化学习（reinforcement）：有奖励
KNN 分类算法
K 指的是最近的 k 个点，能归到一类
输入 trainset 训练，然后用 testset 评价模型，让这个模型分类。
Flat clustering vs Hierarchical clustering
平坦型聚类算法的一个共同点，也是缺陷，就是类别数目难以确定。层次聚类从某种意义上说解决了这个问题，不是它能给出类别数目，而是它在 Clustering 的时候不需要知道类别数。其得到的结果是一棵树，聚类完成之后，可在任意层次横切一刀，得到指定数目的 cluster。
Hard clustering vs soft clustering
硬聚类，强行分成是或不是
软聚类，属于所有类，但概率不同
K-means 聚类算法，肘方法确定 k
用处：归类，然后分析不同类的一些统计学数据上的差异
Decision Tree 分类算法
类别变量，非连续变量，变量属性少时好用

```
#KNN
knn_model=KNeighborsClassifier(n_neighbors=10)
knn_model.fit(train.iloc[:,:-1],train.iloc[:,-1])
pred=knn_model.predict(test.iloc[:,:-1])
print("KNN")
print("Accuracy score: ",accuracy_score(test.iloc[:,-1],pred))
print("F1 score: ",f1_score(test.iloc[:,-1],pred))
print()

#DecisionTree
dt_model = DecisionTreeClassifier()
dt_model.fit(train.iloc[:,:-1],train.iloc[:,-1])
pred=dt_model.predict(test.iloc[:,:-1])
print("DecisionTree")
print("Accuracy score: ",accuracy_score(test.iloc[:,-1],pred))
print("F1 score: ",f1_score(test.iloc[:,-1],pred))
print()
```

Data Base & SQL

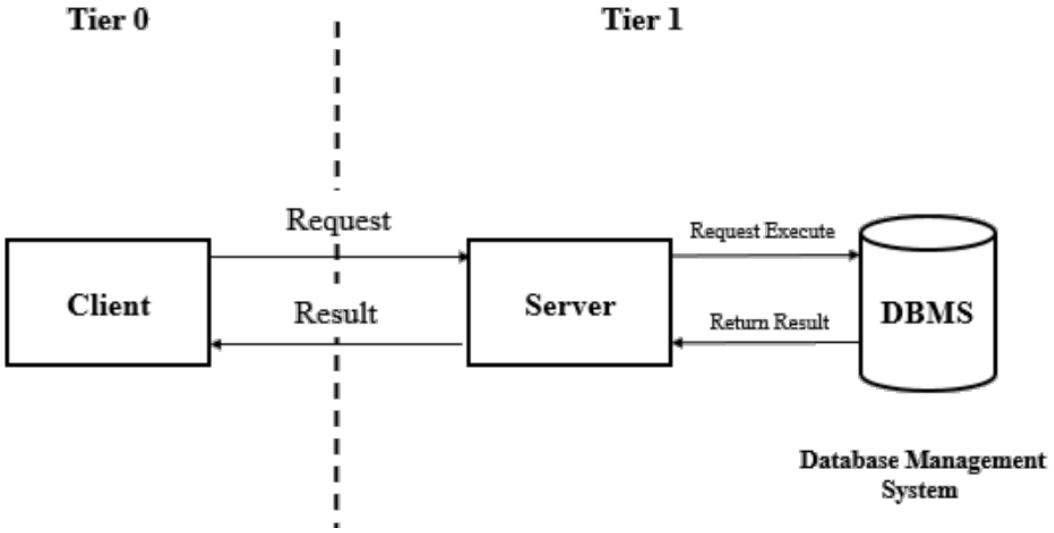
数据库模型：

1. Flat file model: 一张表
2. hierarchical model: 将数据组织成树状模型，实际例子：IBM Information Management System (IMS)
3. network model：有向图表示数据模型
4. relation model ： ER 图
5. object oriented model ： 软件工程里的类图
6. graph model

数据抽象层级：view level -> logical level -> physical level

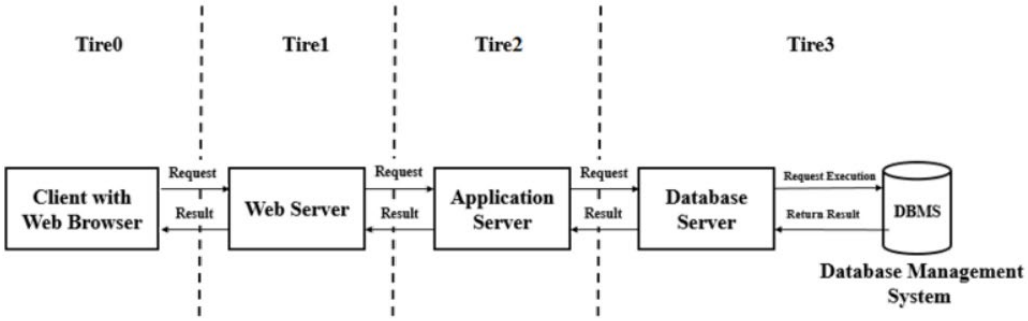
Architecture:

两层结构



N 层结构（在两层中，客户端为一层，客户端服务器为一层，DBMS 为另一层。

在更多的层中，我们可以把服务器分为不同的部分/层和 DBMS。）



Database Language:

1. Data-definition language：定义数据库结构
2. Data-manipulation language：增删查改

范式:

1NF:

- All attribute domains are atomic
- All attributes can only have a single value
- All attributes are uniquely determined by the primary key

2NF:

- It is in 1NF
- All non-key attributes are uniquely determined by the whole primary key, not by a part of it.

3NF:

- It is in 2NF
- All non-key attributes depend on the key[s] and nothing but the key[s]
- Eg, would updating one attribute independently cause issues (such as updating a person's postcode independent of city)?

BCNF:

- It is in 3NF
- All attributes depend on the key[s] and nothing but the key[s]

关系代数:

1. Union (∪) 两列相同的表去并集
2. 2. Intersection (∩) 两列相同的表去交集
3. 3. Difference (−) 例子：A-B，在 A 里但不在 B 里的
4. 4. Selection (σ) sql 里的 where
5. 5. Projection (Π) sql 里的 select
6. 6. Join (⋈) 联表，自然连接

$$\pi_{name,job}(\sigma_{name='harry' \wedge job='police'}(SC \times SC))$$

Information security

CIA Triad ：
confidentiality（保密性）: prevent unauthorized disclosure of information
Integrity (完整性): assure that data cannot be modified by unauthorized manner
Availability (有效性): should be available for authorized users
Authenticity (真实性): proof of identity
Non-repudiation (不可抵赖性): cannot deny what you did
Accountability (问责制): Traceability of actions to a specific entity

Reliability（可靠性）：Consistently performs to specifications
攻击的类型：

- Interruption: Attack on Availability
- Interception: Attack on Confidentiality
- Modification: Attack on Integrity
- Fabrication: Attack on Authenticity

- Passive Attacks:
  - Release of message contents
  - Traffic analysis
- Active Attacks:
  - Masquerade
  - Replay
  - Modification of message contents
  - Denial of service

加密方式：
对称加密（symmetric）
发送方接受方都有密钥
Encryption 密文=Ek(明文)
Decryption 明文=Dk(密文)
Dk(Ek(明文))=明文 正确的明文此公式一定对
Ek(Dk(明文))=明文 不总是对的

DES （data encryption standard）,AES (advanced encryption standard)
DES vs AES:
更快,更安全,更灵活因为 AES 有三种不同密钥尺寸

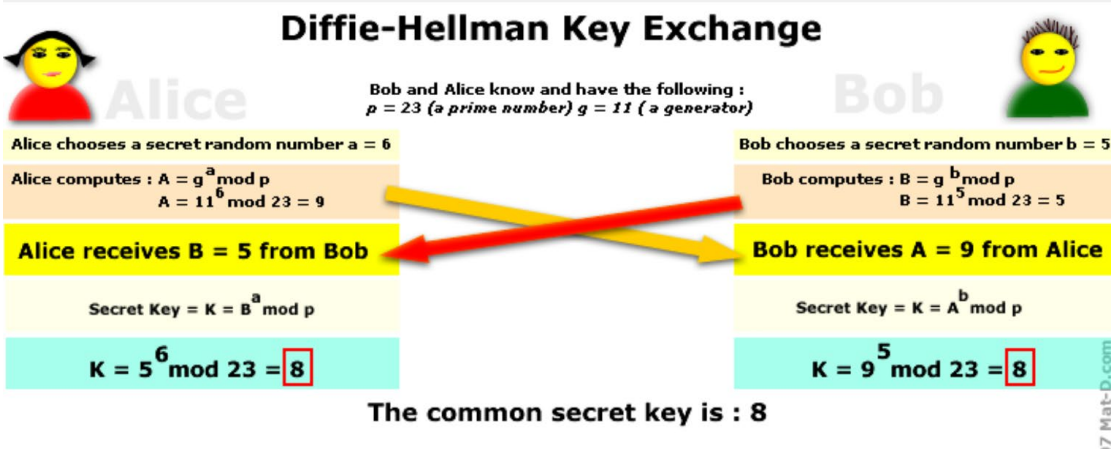
非对称加密（asymmetric）
有公钥私钥，加密解密可能用不同的密钥，不同的算法，有数字签名功能
Public key：公钥可被所有人知道，或者存储于可信的第三方，用于加密与解签名
Private key：私钥只有自己知道，用于解密或签名

常见的非对称加密：RSA：(基于大数质因数分解难被破解)
选两个大质数 p,q, n=pq
选个随机数 e 满足,e 与(p-1)(q-1)互质，1<e<(p-1)(q-1)
Public key : (e,n)

找出个 d,使 ed % (p-1)(q-1) = 1
Private key (d,n)
密文 = 明文^e % n
明文 = 密文^d % n

RSA vs DES
RSA 太慢，DES 加密过程复杂，RSA 数学上难破解，DES 适用于批量处理，RSA 适用于短消息

两种交换密钥的方式：
1. 用非对称加密传递对称加密的密钥
2. Diffie–Hellman key exchange



其中 p（质数），g 是事先两人约定的数

公钥的分发：
可靠第三方颁发（ca, certificate authority）的证书（third trusted party），包括身份信息
例子：
A 与 B 想安全联系
B 获得 A 的公钥，B 获得 ca 的公钥，用 ca 公钥验证 A 的公钥，成功后从证书中提取 A 的公钥，A 做同样的事，两人可以安全联系

编码方式 xml
Well-formed ： 语法没错误
Valid ： 提前定义了数据类型，确保所有数据都符合要求（类比数据库建表时的限制）

Ethic

Some Ethical Issues

Ethical Issues arise from conflicts among stakeholders' interests

Economic

- Income Distribution
- Casualisation of Labour
- Work-Dependence of Income (cf. 'a living wage')

Environmental

- Habitat Destruction
- Climate Change

Political

- Location and Tracking

Technological

- Nuclear Power
- Robotic Warfare

Social

- Capital Punishment
- Unfair Discrimination (Race, Physical Disability)
- Gender Equality
- Continuous Disruption (Workplace, Occupations)

数据科学的伦理学

Ethical Issues in Data Science

- **Data**
  - Expropriation for Unintended Purposes
  - Data Quality Assurance
  - Data Security
- **Data Analysis** Quality Assurance
  - Unfair Discrimination, Redlining, Weblining, 'Algorithmic Discrimination'
- **Decision-Making** delegated to Artefacts
  - Transparency of Decision-Rationale
  - Due Process / Procedural Fairness
  - The Digital Surveillance Economy and 'Surveillance Capitalism'

数据保护与隐私

PIT (Privacy-Invasive technology) 隐私侵入技术

PET （privacy enhancing technology）隐私增强技术:

反 PIT 对存储的数据或传输的数据保护，authentication（证明）
Savage PET：Persistent Anonymity 持续的匿名
Gentle PETs：for Protected Pseudonymity, and hence accountability as well as freedom（保护假名，问责自由）

PET 的种类：

Categories of PETs – 1. Communications

- **Encryption**
  e.g. SSL/TLS and HTTPS Everywhere
- **Email and Instant Messaging / Chat**
  e.g. Protonmail, Hushmail, Fastmail, Signal
- **Handsets**
  e.g. Silent Circle BlackPhone
- **Search-Engines**
  e.g. DuckDuckGo, Ixquick/Startpage
- **Browsers**
  e.g. Stripped Chromium, Brave, Tor, Onion, ...
- **Social Media Services**
  e.g. Diaspora

Categories of PETs

2. Traffic Management

- **End-Point Authentication,**
  e.g. VPNs
- **End-Point Obfuscation**
  Proxy-Servers, VPNs, ToR
- **Firewalls, Malware Filters, Cleansers**
- **Meshnets**
- **Privacy-Enhancing Software Agents**

3. Data Management

- **Stored Data Encryption**
  e.g. Veracrypt
- **Secure Data Deletion**
- **Secure Dropbox**
  e.g. SecureDrop, Podzy