**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**

图示

描述已自动生成

**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。）

图示

描述已自动生成

Database Language：

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

范式：

1NF:

图形用户界面, 文本

描述已自动生成

2NF:

图形用户界面, 文本

描述已自动生成

3NF:

手机屏幕截图

中度可信度描述已自动生成

BCNF:

图片包含 图形用户界面

描述已自动生成

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.  联表，自然连接

**π** name,job(**σ**name=’harry’∧job='police'（SC×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：

文本

描述已自动生成

图形用户界面, 文本, 应用程序, 聊天或短信

描述已自动生成

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

图示

描述已自动生成

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

图片包含 表格

描述已自动生成

数据科学的伦理学

文本

描述已自动生成

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：

文本

描述已自动生成

文本

描述已自动生成