Python、R、matlab 是三种常用的数据分析工具

Python 数据分析中常用的第三方类库：

1、 Numpy，支持大量的多维度数组和矩阵运算，此外也针对数组运算提供大量的数学函数库，NumPy 主要提供以下内容：

（1）快速高效的多维数组对象 ndarray

（2）广播功能函数，广播是一种对数组执行数学运算的函数，其执行的是元素级计算。广播提供了算数运算期间处理不同形状的数组的能力

（3）读/写硬盘上基于数组的数组集的工具

（4）线性代数运算、傅里叶变换及随机数生成功能

（5）将 C、C++、Fortran 代码集成到 python 工具中

NumPy 除了为 python 提供快速的数组处理能力外，NumPy 还有另一个作用，及作为算法之间传递数据的容器。对于数值型数组，使用 NumPy 数组存储和处理数据要比使用内置的 Python 数据结构要高效的多。此外，有其他语言（如 C 语言）编写的库可以直接操作 Numpy 数组中的数据，无需进行数据是复制工作。

2、 Pandas，python 的数据分析核心库，为时间序列分析提供了很好的支持，Pandas 兼具 NumPy 高性能的数组计算功能以及电子表格和关系型数据库（如 SQL）的灵活数据处理功能，它提供了复杂精细的索引功能，以便便捷地完成重塑、切片和切换、聚合及选取数据子集等操作

3、 Matplotlib 是最流行的用于绘制数据图形的 Python 库，他以

各种硬拷贝格式和跨平台的交互式环境生成高质量的图形

4、 Sklearn 可以供用户在各种环境下重复使用，建立在上述三种第三方库的基础上，目前，Sklearn 的基本模块主要由数据预处理、模型选择、分类、聚类、数据降维和回归 6 种

5、 其他 xlrd 和 openpyxl 是读取 excel 文件需要的类库，Seaborn 和 Matplotlib 类似，主要作用是绘制图形，但是 Seaborn 自带了一些数据集，可以用来练习

6、 花式索引是可以通过整数列表或数组进行索引，也可以使用 np.ix()函数完成同样的操作【不太懂】

7、 NumPy.around(a)表示对于 a 中的浮点数取整到最近的整数，但不改变浮点数类型

NumPy.around(a,decimals=1)保留一位小数

NumPy.around(a,decimals=-1) decimals 为 n 对输入近似后保留小数点后 n 位，默认为 0，若值为-n，则对小数点左边第 n 位近似；

8、 NumPy.power(A,B)使用第二个数组作为指数，计算第一个数组中的元素

NumPy.power(A,2)表示对 A 中的元素平方

NumPy.sqrt(A)对 A 中数字开方

NumPy.floor()计算小于或者等于元素的最大整数

NumPy.ceil()计算小于或等于元素的最小整数

NumPy.abs() 绝对值

NumPy.mod() 取余

NumPy.add() NumPy.subtract() NumPy.multiply()

NumPy.divide()加减乘除

【注：dot 使叉乘，数组和矩阵对应位置相乘 *是点积，对数组执行对应位置相乘，必要时使用广播规则】

9、 NumPy 库支持对整个数组或按指定轴向的数据进行统计计算

以下是几个基本的数组统计函数

sum() 求和

mean()算术平均数

std(),var() 标准差 方差

min() max() 最小值和最大值

argmin() argmax() 最小和最大元素的索引

cumsum() 所有元素的累计和【梯形累计，即逐步叠加】

cumprod() 所有元素的累计积【梯形累计，即逐步叠加】

统计函数具有 axis 参数，用于计算指定轴方向的统计值，axis 默认值为 None,此时把数组当成一维数组

10、 数组的元素大多是元素级的，数组相乘的结果是个对应元素的积组成的数组，但是矩阵相乘使用的是点积，NumPy 库提供用于矩阵乘法的 dot()函数。另外，NumPy 库的 linalg 模块来完成具有线性代数运算方法。线性代数函数有：

dot() 两个数组的点积，即元素对应相乘

vdot() 两个向量的点积

det()　数组的行列式

solve()　求解线性矩阵方程

inv()　计算矩阵的乘法逆矩阵

计算特征值时，可以求助于 numpy.linalg 程序包提供的 eigvals 函数和 eig()函数，其中函数 eigvals()返回矩阵的特征向量，eig()返回一个元组，其元素为特征值和特征向量

11、 NumPy 中通过 numpy.savetxt()函数对数组进行存储，通过 numpy.loadtxt()函数对存储数组文件进行读取，并将其加载到一个数组中

12、 Pandas 有 3 中数据结构:系列(Series)、数据帧(DataFrame)、面板(Panel)，这些数据可以构建在 NumPy 数组之上

Series 系列：系列是具有均匀数据的一维数组结构，其特点是：均匀数据、尺寸大小不变、数据的值可变。系列是能够保存任何类型的数据（整数、字符串、浮点数、Python 对象等）的一维标记数组

DataFrame 数据帧：是一个具有异构数据的二维数组，其特点是异构数据、大小可变、数据可变。数据帧是 Pandas 使用最多的数据结构。数据以行和列表示，每行是一条记录（对象），每列是一个属性，属性数据具有数据类型。

Panel 面板是具有异构数据的三维数据结构。其特点是:异构数据、大小可变、数据可变。

13、 Numpy 的广播规则

规则 1：如果两个数组的维度不相同，那么小维度数组的形状将会在最左边补 1.

规则 2：如果两个数组的形状在任何一个维度上都不匹配，那么数组的形状会沿着维度为 1 扩展以匹配另外一个数组的形状。

规则 3：如果两个数组的形状在任何一个维度上都不匹配并且没有任何一个维度为 1，那么会引起异常。

14、 读取外部数据分为读取文件、数据库、网络数据

保存数据的文件主要由 CSV、Excel、txt 和 json

数据库的读取分为两部分：建立连接和 执行 SQL 语句

网络数据的读取使用最多的是网络爬虫，不过 Pandas 提供了 read_html()函数读取网页数据

CSV 格式文件是指以纯文本形式存储的表格数据，巨量的数据常使用 CSV 格式。Pandas 提供了处理数据量巨大的 CSV 文件功能

read_table()函数于 read_csv()函数大同小异，不同之处在于 read_table 默认分隔符为制表符，而 read_csv 默认分隔符为英文逗号

15、 函数应用与映射运算的作用是将其他函数或者是自定义函数应用于 Pandas 对象，函数主要包括：pipe() apply()

applymap() map()

16、 数据预处理是一项极其重要又非常繁琐的工作，数据预处理的好坏对数据分析结果有决定性作用，同时在实际的数据分析和建模中，大约 80%的时间是花费在数据准备和预处理上的。

17、 数据清洗主要是处理原始数据中的重复数据、缺失数据和异常数据，使数据分析不受无效数据的影响。

18、 原始数据往往会出现重复数据，对于重复的数据通常需要删除多余的记录，保留一份即可。因此在处理重复值时的过程为先检测重复值，然后再处理重复值。

19、 数据采集中由于设备或人为原因可能造成部分数据缺失，数据缺失会对数据分析造成不利影响，因此必须加以处理。

（1）检测缺失值，在处理缺失值前，需要先找到缺失值，isnull()函数可以检查数据中的缺失值，返回一个布尔型矩阵，每一个布尔值表示对应位置的数据是否缺失

（2）处理缺失值的四种方法：删除法（dropna（）），固定值替换（replace（））效果不好，填充法（fillna（）），插值法（效果好，有三种常用的方法，线性插值、多项式插值、样条插值）

20、原始数据中可能会出现明显违背自然规律的数据，这些数据属于噪声，对数据分析会造成很大的干扰，对分析结果具有巨大的不良影响。

（1）异常值的检测通常采用绘制图形，从图形观察数据分布情况，找出离群点。离群点可能是异常值，但是不绝对，有些时候离群点的数据也可能是正常的。异常值的判断需要有行业背景和业务知识。

（2）处理异常值主要有三种方法：

1、删除，删除包含有异常值的纪律

2、视为缺失值，将异常值视为缺失值，利用缺失值的处理方法

进行处理

3、平均值修正。使用前后两个值的平均值或者整列的数据平均值修正异常值

21、连接类型有内连接，左连接，右连接，外连接。

内连接（inner）是最常用的连接，左右两个 DataFrame 数据主键具有相等关系时，左 DataFrame 的记录才会和右 DataFrame 的记录合并。

左连接（left）：以左 DataFrame 为主，当右 DataFrame 具有对应数据时，和 inner 连接相同，当右 DataFrame 没有数据和左 DataFrame 相同时，则右 DataFrame 的值取 NaN

右连接（right）与左连接相反

外连接（outer）：左连接和右连接的和，左 DataFrame 没有数据与右 DataFrame 对应时，左 DataFrame 的值设为 NaN，反之亦然。

22、数据的合并连接和重塑

方法函数有：merge()合并，concat()合并，combine_first()合并

数据重塑是将 DataFrame 的行或列进行旋转的操作，stack()函数将 DataFrame 的列旋转为行，unstack()函数将 DataFrame 的行旋转为列。对于层次化索引，数据重塑的操作默认从最内层旋转，当然可以通过设置 stack()或 unstack()的参数指定重塑的索引层次。

23、数据变换是对原始数据按照一定的规则进行变换的数据处理，将数据转换成适合分析的形式，以满足数据分析的需要，提升分析效果。

虚拟变量：在某些数据中，是一些选项而不是一些可计算的值，比如

性别，班级等，像这样的值在计算过程中需要转化成可量化的数值，这时就需要虚拟变量将其转化成可量化的数值。

连续属性离散化：有些数据分析算法需要数据是离散的值，如果对应的原始数据是连续值，需要把连续属性离散化，离散化的方法主要有两种：等宽法、等频法。等宽法：将属性的值域分成具有相同宽度的区间，区间的个数有数据本身的特点决定。等频法：将区间划分为指定个数的区间，将相同数量的记录放进每个区间，即每个区间具有相同的数据个数。

规范化：在数据分析过程中，不同评价指标往往具有不同的量纲，数值间的差别可能很大，如果不加处理直接使用，通常会影响数据分析的结果。数据规范化（归一化）的主要作用就是消除指标之间的量纲和取值范围差异的影响，是数据分析的基本工作。数据规范化按照比例进行缩放，使之落入一个特定的区域，便于进行综合分析，通常把标量取值映射到[0, 1]或者[−1, 1]内，主要的规范化方法有 3 个：

（1）最小−最大规范化：最小−最大规范化也称离差规范化，是对原始数据进行线性变化，将数值映射到[0, 1]之间，转化公式为：

$$X^* = \frac{x - min}{max - min}$$

（2）零−均值规范化：零−均值规范化也称标准差规范化，经过处理的数据的均值为 0，标准差为 1. 转化公式为：$x^* = \frac{x-\bar{x}}{\sigma}$，$\sigma$为样本数据标准差，零−均值规范化是目前使用最多的数据标准化方法。

（3）小数定标规范化：小数定标规范化通过移动属性值的小数位数，

将属性值映射到[-1,1]内，移动的小数位数取决于属性值绝对值的最大值。转化公式为：$x^* = \frac{x}{x^k}$

随机采样:随机采样是从原始数据中随机选出一部分的数据，需要两个函数配合使用来实现。

① numpy.random.permutation(n)函数可以产生 0—n 范围内的 n 个随机数，输出形式为 numpy 数组

② df.take(np.random.permutation(len(df))[:m])函数可以从原有的 n 行数据中随机抽取 m 行数据。

24、机器学习是通过让机器对已知样本进行学习，然后对更多的未知样本进行预测的过程。机器学习可以分成两类：监督学习和无监督学习。

监督学习：通过所有特征已知的训练集让机器学习其中的规律，然后再向机器提供有一部分特征未知的数据集，让机器帮补全其中未知部分的一种方法。主要包括分类和回归两类。（1）分类：分类是指根据样本数据中已知的分类进行学习，对未知分类的数据进行分类的算法（2）回归。回归是指根据样本中离散的特征描述出一个连续的回归曲线，之后只要能给出其他任意几个维度的值就能够确定某个缺失的维度值的方法就称为回归。例如可以根据人类的性别、年龄、家族成员信息建立一个甚高的回归方程，可以预测新生儿各个年龄阶段的身高。

无监督学习：无监督学习不会为机器提供正确的样本学习，而是靠机器自己去寻找可以参考的依据，通常使用距离函数或者凸包理论等方

式对给定的数据集进行聚类。例如，对按照用户访问网站的行为将用户分成不同的类型。

训练集和测试机：通常在有监督的机器学习中会有一组已知其分类或结果值的数据，一般来说不能把这些数据全部用来进行训练，如果使用全部的数据进行训练，可能导致过拟合，而且也需要用一部分数据来验证算法的效果

过拟合：过拟合训练后的算法虽然严格的符合训练集，但可能会在面的真正的数据时效果变得很差，在过拟合的训练结果中将会使算法在测试时表现得完美无缺，但是实际应用时却很不理想。

正确率/召回率/ROC 曲线：正确率/召回率/ROC 曲线是用来衡量机器学习算法效果的 3 个指标，正确率是指提取出的正确信息条数与提取出的信息条数的比率，但这实际上掩盖了样本是如何被分错的。召回率是指提取出的正确信息条数与样本中的信息条数的比值，召回率越大表示被错判的正例就越少。ROC 曲线则是“被正确分为正例的正例与被错误分为正例的负例”的曲线

降维：当数据的维度特别大时(如自然语言处理)，需要对数据进行降维以减小对计算的需求，同时对结果的影响又不会太大。

25、机器学习的过程

（1）将实际问题抽象成数学问题：该过程明确目标是一个分类还是回归或者是聚类的问题，如果都不是，如何抽象成其中的某一类问题。

（2）获取数据：数据在机器学习中作用巨大，决定机器学习结果的上线，而算法只是尽可能地逼近这个上限。

①获取数据包括获取原始数据以及从原始数据中经过特征工程从原始数据中提取训练、测试数据。学习过程中原始数据都是直接提供的，但是解决实际问题时需要自己获得原始数据。

②数据要具有代表性，例如，对于分类问题、数据偏斜，不同类别的数据数量不要相差太大。

③ 还有对评估数据的量级、样本数量、特征数，估算训练模型对内存的消耗，如果数据量太大，还可以考虑减少训练样本、降维或者使用分布式机器学习系统

（3）特征工程

特征工程包括从原始数据中进行特征构建，特征提取，特征选择。特征工程做的好能发挥原始数据的最大效力，往往能够使得算法的效果和性能得到显著提升。

数据预处理、数据清洗、筛选显著特征、摒弃非显著性特征都是特征工程的重要内容，

（4）训练模型，诊断，调优

模型诊断中至关重要的是判断过拟合、欠拟合，常见的方法是绘制学习曲线，进行交叉验证。通过增加训练的数据量、降低模型复杂度来降低过拟合的风险，提高特征的数量和质量，增加模型复杂来防止欠拟合。　　诊断后的模型需要进行进一步调优，调优后的新模型需要重新进行诊断，这是一个反复迭代、不断逼近的过程，需要不断的尝试，达到最优的状态。

（5）模型验证，误差分析

通过测试数据验证模型的有效性，观察误差样本，分析误差产生的原因，往往能使得我们找到提升算法性能的突破点，误差分析主要是分析出误差来源与数据、特征、算法。

（6）模型融合

提升算法准确度的主要方法是模型的前端（特征工程，清洗，预处理，采样）和后端的模型融合。

（7）上线运行

工程上是结果导向，模型在线上运行的效果直接决定模型的成败。除了准确程度、误差等情况外，还要考虑运行速度，资源消耗程度和稳定性等。

26、Sklearn 是 Scipy 的扩展，是建立在 NumPy 和 Matplotlib 库基础上的一个机器学习算法库。Sklearn 主要包括特征提取、数据处理和模型评估三大模块;主要功能有:Classification 分类、Regression 回归、Clustering 非监督聚类、Dimensionality reduction 数据降维、Model Selection 模型选择、Preprocessing 数据预处理等。

（1）Sklearn 的特点

①简单高效，能够在复杂环境中重复使用

②利用 NumPy、Scipy 和 Matplotlib 的优势，可以大大提高机器学习的效率

③拥有完善的文档，上手容易，拥有丰富的 API

④封装大量的机器学习算法，包括 LIBSVM 和 LIBINEAR 等

⑤内置大量的数据集，节省了获取和整理数据集的时间

（2）Sklearn 算法选择

Sklearn 算法主要有四类：分类、回归、聚类、降维

① 常用的回归：线性、决策树、SVM、KNN

② 集成回归：随机森林、Adaboost、GradientBoosting、Bagging、ExtraTrees

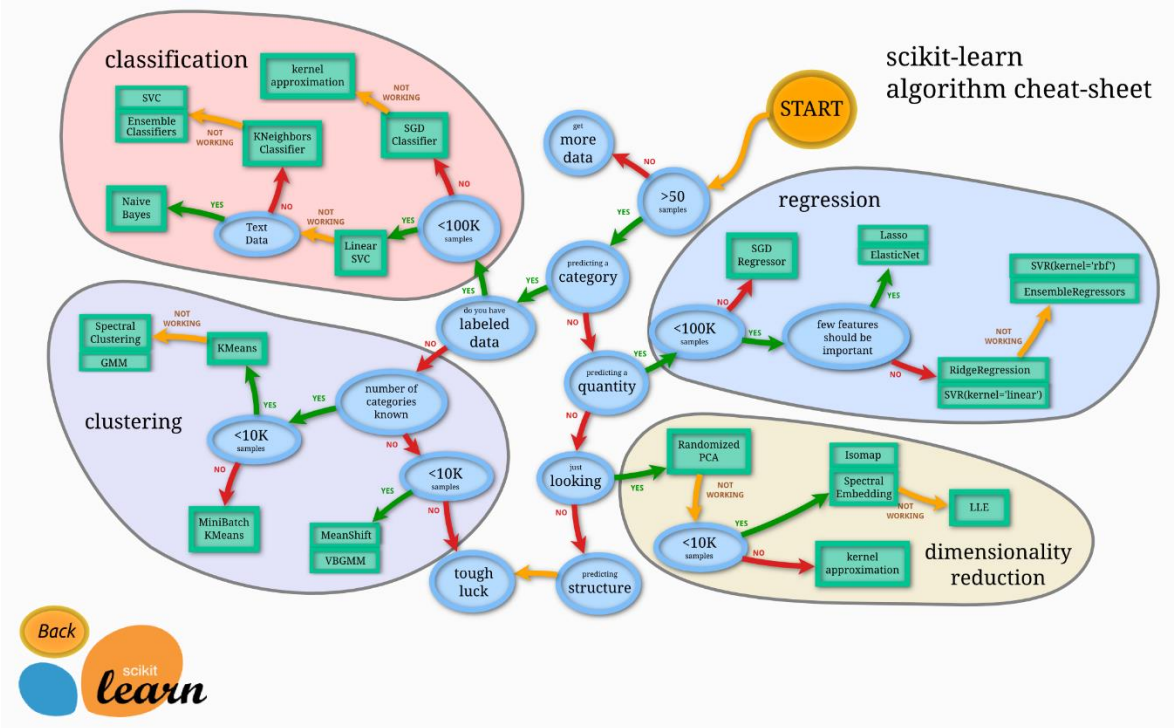
③ 常用的分类：线性、决策树、SVM、KNN、朴素贝叶斯

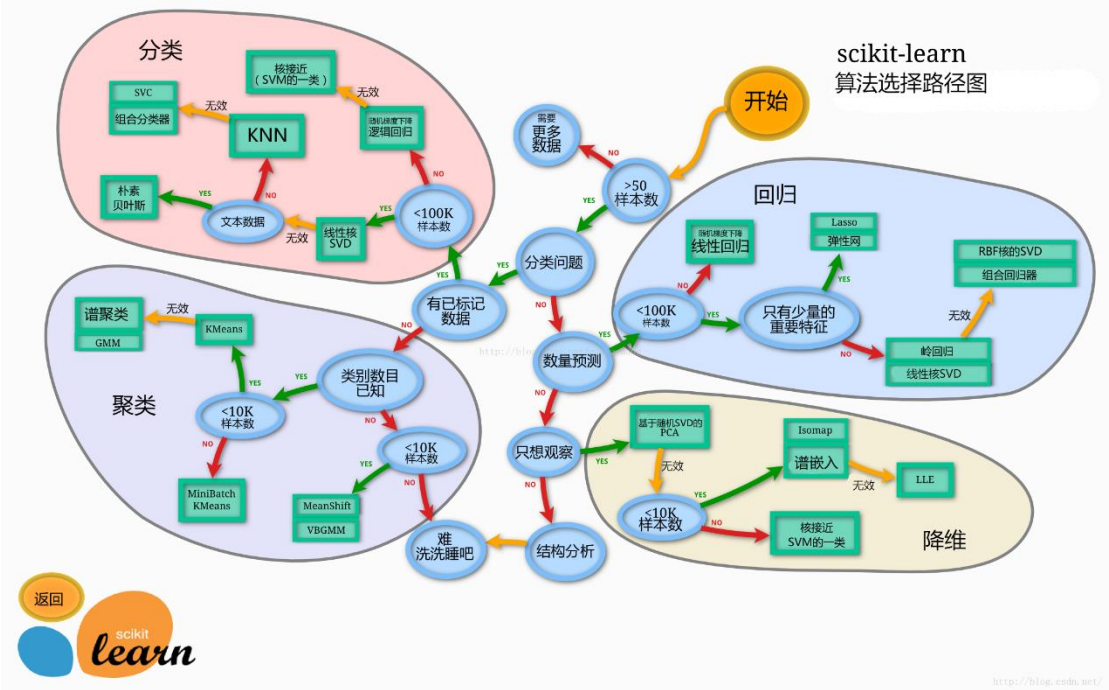
④ 集成分类：随机森林、Adaboost、GradientBoosting、Bagging、ExtraTrees

⑤ 常用聚类：k 均值（k-means）、层次聚类（Hierarchical clustering）、DBSCAN

⑥ 常用降维：LinearDiscriminantAnalysis、PCA

27、sklearn 算法选择路径图





28、Sklearn 数据集

Sklearn 的数据集主要有如下五种：

① 自带的小数据集（Packaged Dataset）：

sklearn.datasets.load_<name>

② 可在线下载的数据集（Downloaded Dataset）：

sklearn.datasets.fetch_<name>

③ 计算机生成的数据（Genered Dataset）：

sklearn.datasets.make_<name>

④ svmlight/libsvm 个数的数据集：

sklearn.datasets.load_svmlight_file(…)

⑤ data.org 在线下载获取的数据集：

sklearn.datasets.fetch_mldata（…）

其中<name>表示数据集的名字

名称	描述	适合任务
load_iris()	鸢尾花数据集	分类、聚类
load_breast_cancer()	乳腺癌数据集	分类、聚类
load_digits()	手写数字数据集	分类
load_diabets ()	糖尿病数据集	回归
load_boston()	波士顿房价数据集	回归
load_linnerud()	体能训练数据集	分类
load_wine()	葡萄酒数据集	分类、聚类

29、make 数据集

生成数据集：可以用来分类任务、回归任务和聚类任务

用于分类任务和聚类任务的函数：产生样本特征向量矩阵以及对应的类别标签集合。常用函数有 make-blobs() 函数、make_classification() 函数、生成球形判决界面数据集、make_gaussian_quantiles(将一个单高斯分布的点集划分为两个数量均等的点集，作为两类)、make_hastie-10-2（产生一个相似的二元分类数据表，有 10 个维度）

30、sklearn 算法

Sklearn 对机器学习常用算法都做了实现，只需要调用即可。

线性回归：是利用数理统计中回归分析，来确定两种或者两种以上变量间相互依赖的定量关系的一种统计分析方法，运用十分广泛，其表达形式为$y = w' + b$,回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种分析称为一元线性回归分

析。如果回归分析中包括两个或两个以上的自变量，且隐变量和自变量之间是线性关系，则称为多元线性回归分析。线性分析实际上使用一条直线去拟合一大推数据，求出系数w和截距b，获得直线方程，然后可以使用函数求出其他未知的值。Sklearn 中线性回归使用最小二乘法实现，使用起来非常简单，线性回归是回归问题，其回归模型的评价函数 score 使用 R2 系数作为评价标准。

逻辑回归：是一种广义线性回归（Generalized Linear Model），因此与多重线性回归分析有很多相同之处，他们的模型形式基本上相同，都具有$y = w' + b$，其中w和b是待求参数，其区别在于他们的因变量不同，多重线性回归直接将$w' + b$作为因变量，即$y = w' + b$，而逻辑回归则通过函数 L 将$w'x + b$对应一个隐状态 p，p=$L(w'x + b)$，然后根据 p 与 1-p 的大小决定因变量的值。如果 L 是逻辑函数，就是逻辑回归，如果 L 是多项式函数就是多项式回归。

朴素贝叶斯：贝叶斯分类是一类分类算法的总称，这类算法均以贝叶斯定理为基础，故统称为贝叶斯分类。朴素贝叶斯分类是贝叶斯分类中最简单，也是常见的一种分类方法。朴素贝叶斯的核心便是贝叶斯公式：$P(B|A) = P(A|B)P(B)/P(A)$，即在 A 条件下，B 发生的概率，换个角度可以表示：$P(\text{类别}|\text{特征}) = P(\text{特征}|\text{类别})P(\text{类别})/P(\text{特征})$，朴素贝叶斯求解的就是$P(\text{类别}|\text{特征})$

决策树：决策树是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这

种决策分支化成图像很向一棵树的枝干，故称决策树。在机器学习中，决策树是一个预测模型，它代表的是对象属性与对象值之间的一种映射关系。决策树是一种树形结构，其中每个内部节点表示一个属性上的测试，每个分支代表一个测试输出，每个叶节点代表一个类别。

SVM(支持向量机)：支持向量机是通过求解最大化间隔解决分类问题。支持向量机将向量映射到一个更高维的空间，在这个空间里建立一个最大间隔超平面，在分开数据的超平面的两边建有两个互相平行的超平面，建立方向合适的分隔超平面使两个与之平行的超平面间的距离最大化。其假定为，平行超平面间的距离或差距越大，分类器的总误差越小。SVM 的关键在于核函数，低维无法线性划分的问题放到高维就可以线性划分，一般用高斯函数。

神经网络：神经网络是一种模仿动物神经网络行为特征，进行分布式并行信息处理的算法数学模型。这种网络依靠系统的复杂程度，通过调整内部大量结点之间相互连接的关系，从而达到处理信息的目的，并具有自学习和自适应能力。

KNN(K-近邻算法)：在训练集中数据和标签已知的情况下，输入测试数据，将测试数据的特征与训练集中对应的特征相互比较，找到训练集中与之最为相似的前 K 个数据，则该测试数据对应的类别就是 K 个数据中出现次数最多的那个分类。

KNN 是分类算法，K-Means 是聚类算法。

算法的描述如下：

① 计算测试数据与各个训练数据之间的距离

② 按照距离的递增关系进行排序

③ 选取距离最小的 K 个点

④ 确定前 K 个点所在类别出现的频率

⑤ 返回前 K 个点中出现频率最高的类别作为测试数据的预测分类

31、sklearn 中提供的数据预处理功能

（1）标准化：为了使得训练数据的标准化规则与测试数据的标准化规则同步，preprocessing 中提供了很多的 Scaler

①基于 mean 和 std 的标准化

②最小-最大标准化值

（2）归一化：归一化的过程是将每个样本缩放到单位范数（单位样本的范数为 1），其思想是对每个样本计算其 p-范数，然后对该样本中每个元素除以该范数,这样处理的结果是使得每个处理后样本的 p-范数（绝对值的和，欧几里得距离）等于 1

（3）one-hot 编码（独热编码）：是一种对离散特征值的编码方式，在 LR 模型中常用到，用于给线性模型增加非线性能力。

32、数据集拆分：通常把数据集拆分成训练集和测试集（验证集），这样有助于模型参数的选取。

33、模型评估

模型评估是对预测质量的评估，即模型的评价，主要有 3 种方法：

（1）模型自带 score()函数，其值为[0,1]之间的数，1 表示最好

（2）cross_val_score()函数，得到一个对于当前模型的评估得分

（3）评估函数

34、Sklearn 常用方法

Sklearn 中所有的模型都有 4 个固定的且常用的方法。

（1）模型训练，model.fit(X_train,y_train)

（2）模型预测，model.predict(X_test)

（3）获得这个模型的参数，model.get_params()

（4）为模型进行评价，model.score(data_X,data_y)

35、模型的保存和载入

模型的保存和载入方便我们将训练好的模型保存和共享,模型保存和载入方法如下：

```
Import sklearn.externals as sk_externals
sk_externals.joblib.dump(model,’model.pickle’)  # 保存
model= sk_externals.joblib.dump(model,’model.pickle’)
```

36、降维：降维是对数据高维度特征的一种预处理方法，降维将高维度的数据保留下最重要的一部分特征，去除噪声和不重要特征，从而实现提升数据处理速度的目的。在实际的生产和应用中，降维在一定的损失范围内，可以节省大量的时间和成本，降维也成为应用非常广泛的数据预处理方法，降维具有以下优点：

①使得数据集更易使用

②降低算法的计算开销

③去除噪声

④ 使得结果容易理解

PCA（主成分分析法），是一种最广泛的数据压缩算法，是无监督的学习方法。

在 PCA 中，数据从原来的坐标系转换到新的坐标系，由数据本身决定，转换坐标系时，以方差最大的方向作为坐标轴方向，因为数据的最大方差给出了数据的最重要信息，第一个新坐标轴选择的原始数据中方差最大的方法，第二个新坐标轴选择的是与第一个新座标轴正交且方差次大的方向，重复该过程，重复次数为原始数据的特征维数。

信号具有较大的方差，噪声具有较小的方差，因此 PCA 的目标是新坐标系上数据的方差越大越好。

LDA（线性评价分析），LDA 基于费舍尔准则，即同类样本尽可能聚合在一起，不同类样本尽量扩散，即同类样本具有较好的聚合度，类别间具有较好的扩散度。

37、回归：回归分析是确定两种或者两种以上变量间相互依赖的定量关系的一种统计分析方法，应用十分广泛，回归分析按照涉及的自变量多少，分为回归和多重回归分析，按照自变量的多少分为一元回归分析和多元回归分析；按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。如果在回归分析中，只包括一个自变量和因变量，且二者的关系可用一条直线近似表示，则这种回归分析可以称为一元线性回归分析。如果回归分析中包括两个或者两个以上的自变量，且因变量和自变量之间是线性关系，则称为多重线性回归分析。

线性回归是利用数理统计中的回归分析，来确定两种或两种以上变量

间相互依赖的定量关系的一种统计分析方法。

逻辑回归与多重线性回归同属于一个广义线性模型家族，因此有很多相同之处，但是也有不同之处，最大的区别就在于他们的因变量不同。

① 如果是连续的，就是多重线性回归

② 如果是二项分布，就是逻辑回归

③ 如果是 Poisson 分布，就是 Poisson 回归

④ 如果是负二项分布，就是负二项回归

逻辑回归的因变量可以是二分类的，也可以是多分类的，但是二分类的更为常用，也更容易解释，因此在实际中最常用的就是二分类的逻辑回归。逻辑回归主要在流行病学中应用较多，比较常用的情形是探索某疾病的危险因素，根据危险因素预测某疾病发生的概率等。例如，想探讨胃癌发生的危险因素，可以选择两组人群，一组是胃癌组，一组是非胃癌组，两组人群肯定有不同的体征和生活方式等，这里的因变量就是是否胃癌，即“是”或“否”，自变量就可以包括很多，如年龄、性别、饮食习惯等，自变量既可以是连续的，也可以是分类的。

回归问题的常规步骤如下：

① 寻找 h() 函数（即 hypothesis）

② 构造 j() 函数（损失函数）

③ 想办法使得 j 函数最小并求得回归参数（θ）

回归树是可以用于回归的决策树模型，一个回归树对应着输入空间（即特征空间）的一个划分以及在划分单元上的输出值. 与分类树不同的是，回归树对输入空间的划分采用一种启发式的方法，会遍历所

有输入变量，找到最优的切分变量 j 和最优的切分点 s，即选择第 j 个特征x^j和它的取值 s 将输入空间划分为两部分，然后重复这个操作。

38、在机器学习和统计中，分类是基于包含其类别成员已知的实例训练集来识别新观察所属的一组类别中的哪一个问题。

朴素贝叶斯：是基于贝叶斯定理与特征条件独立假设的分类方法，最广泛的两种分类模型是决策树模型和朴素贝叶斯模型，朴素贝叶斯的基本方法：在统计数据的基础上，依据条件概率公式，计算当前特征的样本属于某个分类的概率，选最大的概率分类。理论上，朴素贝叶斯分类器（NBC）模型与其他分类方法相比具有最小的误差率，但实际上并非如此，这是因为 NBC 模型假设属性之间相互独立，这个假设在实际应用中往往不成立，这给 NBC 模型的正确分类带来一定的影响

分类决策树：决策树是在已知各种情况发生概率的基础上，通过构建决策树来求取净现值的期望值大于或者等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。由于这种决策分支画成图形很像一棵树的枝干，故称决策树。决策树是一种树形结构，其中每个内部结点表示一个属性上的测试，每个分支代表一个测试输出，每个叶结点代表一个类别。

支持向量机（SVM）是一种分类算法，但是也可以做回归，根据输入的数据不同可做不同的模型。若输入标签为连续值则做回归，若输入标签为分类值则用 SVC()做分类。

通过寻求结构化的风险最小来提高学习泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。

SVM 是一种二类分类模型，其基本模型定义为特征空间上的间隔最大的线性分类器，即支持向量机的学习策略便是间隔最大化，最终可转化成一个凸二次规划问题的求解。

神经网络是一种应用类似于大脑神经突触连接的结构进行信息处理的数学模型，在工程与学术界也常直接简称为人工神经网络或者类神经网络，神经网络是一种运算模型，由大量的结点（或称神经元）和结点之间相互连接构成。每个结点代表一种特定的输出函数，称为激励函数。每两个结点间的连接都代表一个对于通过该连接信号的加权值，称为权重，这相当于人工神经网络记忆网络的记忆。网络的输出则依网络的连接方式、权重值和激励函数的不同而不同。而网络自身通常都是对自然界某种算法或者函数的逼近，也可能是对一种逻辑策略的表达。他的构筑理念是受到生物（人或其他动物）神经网络功能的运作启发而产生的。人工神经网络通常是通过一个基于数学统计学类型的学习方法得以优化，所以人工神经网络也是数学统计学方法的一种实际应用，通过统计学的标准数学方法我们能够得到大量的可以用函数来表达的具部结构空间，另一方面，在人工智能学的人工感知领域，我们通过数学统计学的应用可以来做人工感知方面的决定问题，通过统计学方法，人工神经网络能够类似人一样具有简单的决定能力和简单的判断能力。

K 近邻（KNN）算法的核心思想是如果一个样本在特征空间中的 k 个最相似（即特征空间中最邻近）的样本中的大多数属于某一个类别，则该样本也属于这个类别，KNN 算法不仅可以用于多类别分类，还可以用于回归。通过找出一个样本的 k 个最近邻居，将这些邻居的属性的平均值赋给该样本，作为预测值。

39、聚类分析指将物理对象或抽象对象的集合分组为由类似的对象组成的多个类的分析过程，聚类分析的目标就是在相似的基础上收集数据来分类。聚类与分类的不同在于聚类所要求划分的类是未知的。聚类是将数据分类到不同的类或者簇这样的一个过程，所以同一个簇中的对象有大的相似性，而不同簇间的对象有很大的相异性。从统计学的观点看，聚类分析是通过数据建模简化数据的一种方法，聚类分析是一种探索性的分析，能够从样本数据出发，自动进行分类。聚类分析所使用的方法不同，通常会得到不同的结论。

K-means 算法的基本思想：以空间中 K 个点为形心进行聚类，对最靠近他们的兑现归类。通过迭代的方法，逐次更新各簇的形心值，知道得到最好的聚类结果。K-means 算法最大的优势在于简洁和快捷，算法的关键在于初始中心的选择和距离公式。

DBSCAN 算法是通过样本的紧密程度来确定样本的分布，适用于集群在任何形状的情况下。DBSCAN 算法中的一个重要概念就是核心样本（具有较高的紧密度），该算法有两个参数 min_samples 和 eps，高 min_samples 或者低 eps 代表着在形成聚类时，需要较高的紧密度。该算法的简单描述：先任意选择数据集中的一个核心对象为“种子”，

在由此发出确定相应的聚类簇，再根据给定的领域参数（min_samples,eps）找出所有的核心对象，在以任意一个核心对象出发，找出由其密度可达的样本生产聚类簇，知道核心对象均被访问为止。