人工智能之机器学习

机器学习概述

主讲人：李老师

机器学习概述

01 机器学习的定义

02 机器学习、人工智能和深度学习的关系

03 机器学习基本概念和常用的应用场景

04 机器学习、数据分析、数据挖掘的区别与联系

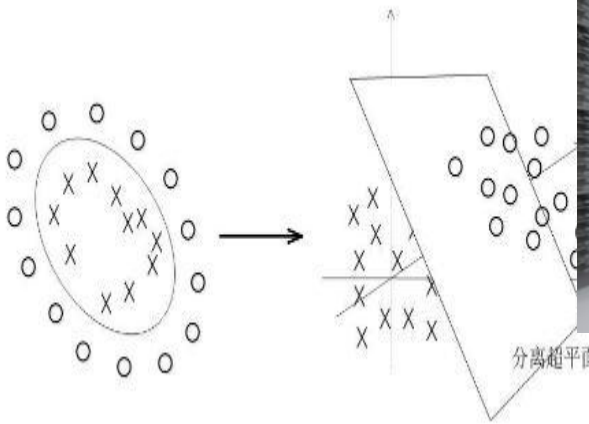
05 机器学习分类

06 机器学习数据处理流程

Machine Learning

What society thinks I do

Wha



/hat my parents thinks I do

SVM:

$$\min_{\mathbf{w},\xi,b}\left\{\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{n}\xi_i\right\}$$

subject to (for any $i=1,\dots n$)

$$y_i(\mathbf{w}\cdot\mathbf{x_i}-b)\geq 1-\xi_i,\qquad \xi_i\geq 0$$

LR:

$$\min_{\theta}\sum_{i=1}^{M}-\log p(y^{(i)}|\mathbf{x}^{(i)};\theta)+\beta\|\theta\|_1.$$

$$P_w(y|x)=\frac{\exp w^\top\Phi(x,y)}{\sum_{y'\in\mathrm{GEN}(x)}\exp w^\top\Phi(x,y')}.$$

What other programmers thinks I do

What I thinks I do

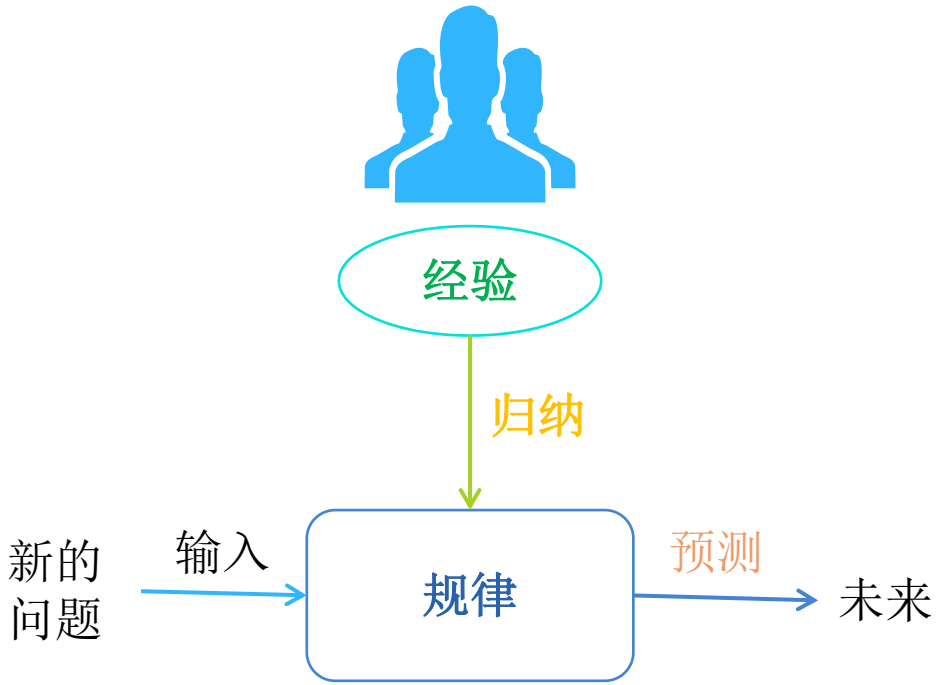
```
In [1]: from sklearn import svm

In [ ]:
```

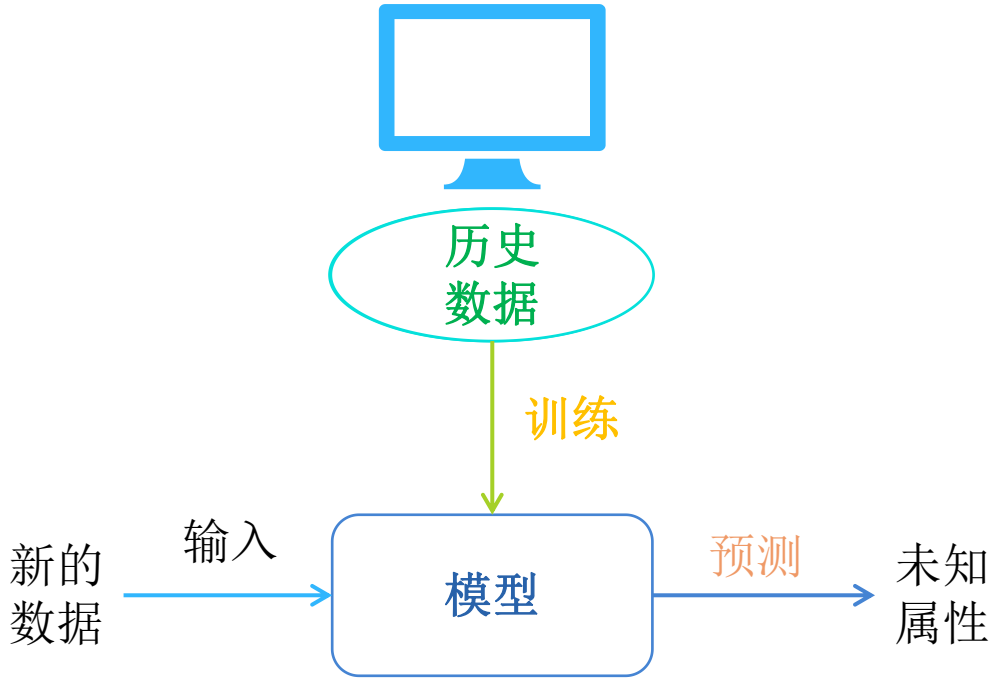
What I really do

机器学习定义

- **Machine Learning**(ML) is a scientific discipline that deals with the construction and study of algorithms that can learn from data.
- 机器学习是一门从数据中研究算法的科学学科。
- 机器学习直白来讲，是根据已有的数据，进行算法选择，并基于算法和数据构建模型，最终对未来进行预测;
- 备注：机器学习就是一个模拟**人决策过程**的一种程序结构。

机器学习/人工智能理性认识





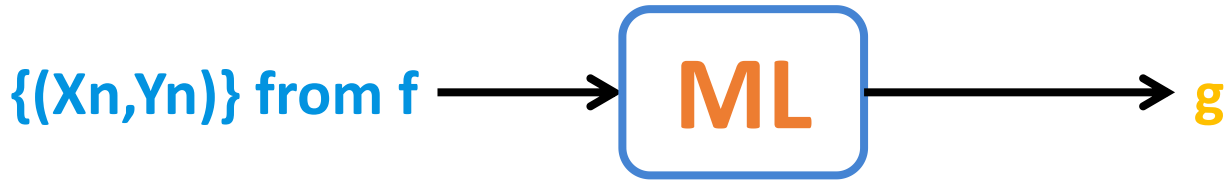
机器学习理性认识

- 基本概念
  - 输入：$x \in X$（属性值, 特征属性）
  - 输出：$y \in Y$(目标值）
  - 获得一个目标函数(target function):

$$f : X \rightarrow Y(\text{理想的公式})$$

  - 输入数据: $D = \{(x_1, y_1), (x_2, y_2), \cdots\cdots, (x_n, y_n)\}(\text{历史记录信息})$
  - 最终具有最优性能的假设公式:

$$g : X \rightarrow Y(\text{学习得到的最终公式})$$



机器学习概念

- 拟合：构建的算法模型符合给定数据的特征
- x_i :x向量的第i维度的值
- $x^{(i)}$:表示第i个样本的x向量
- 鲁棒性：也就是健壮性、稳健性、强健性,是系统的健壮性；当存在异常数据的时候，算法也会拟合数据
- 过拟合：算法太符合样本数据的特征，对于实际生产中的数据特征无法拟合
- 欠拟合：算法不太符合样本的数据特征

机器学习概念

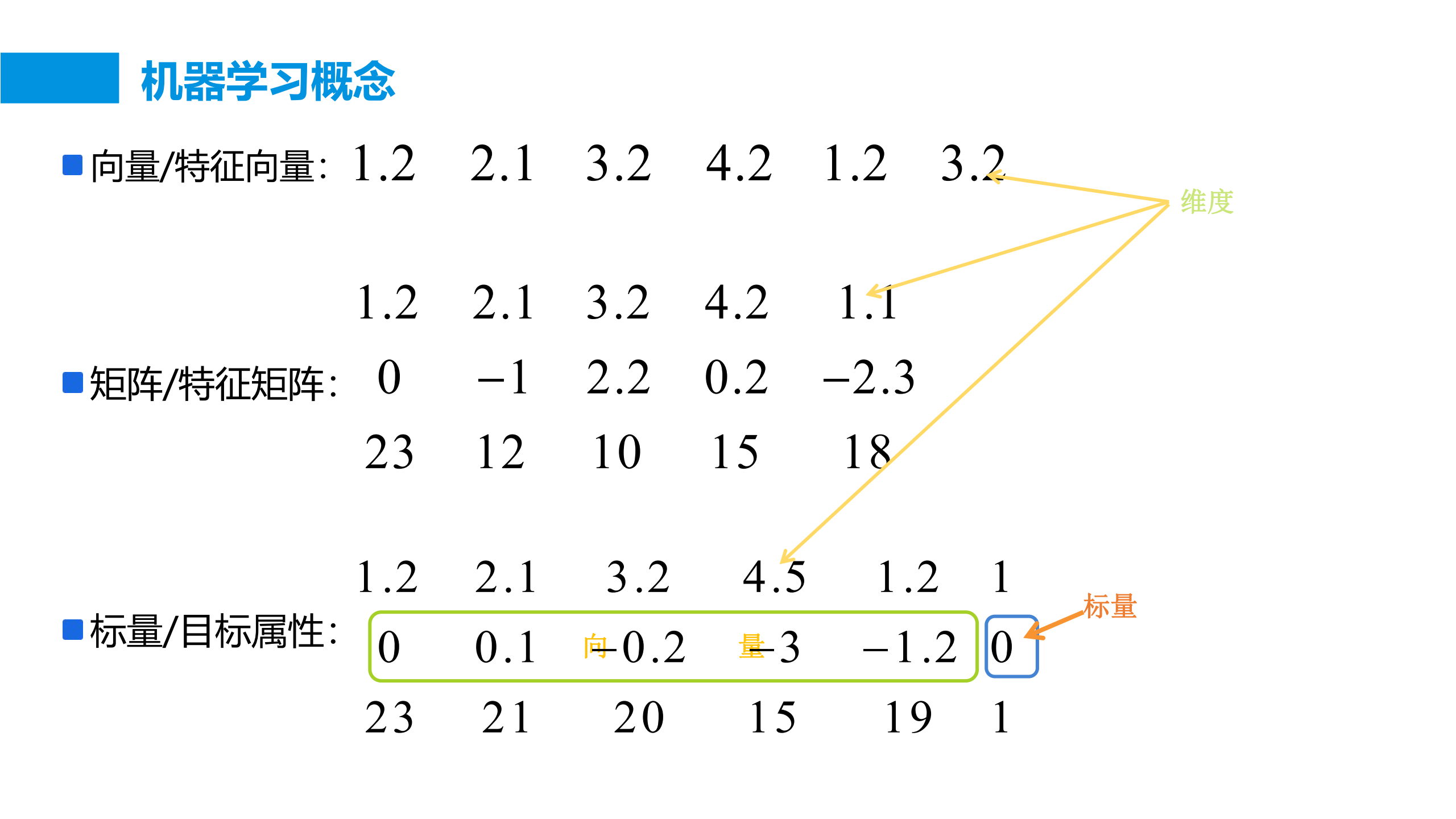
- 向量/特征向量：$1.2 \quad 2.1 \quad 3.2 \quad 4.2 \quad 1.2 \quad 3.2$

$$\begin{matrix} 1.2 & 2.1 & 3.2 & 4.2 & 1.1 \\ 0 & -1 & 2.2 & 0.2 & -2.3 \\ 23 & 12 & 10 & 15 & 18 \end{matrix}$$

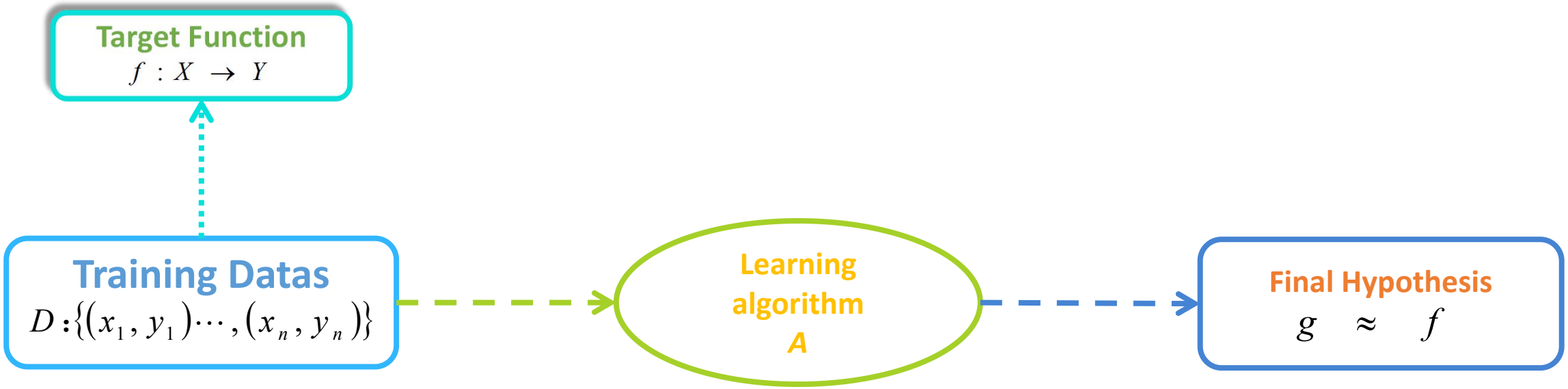
- 矩阵/特征矩阵：

- 标量/目标属性：

$$\begin{matrix} 1.2 & 2.1 & 3.2 & 4.5 & 1.2 & 1 \\ 0 & 0.1 & 0.2 & 3 & -1.2 & 0 \\ 23 & 21 & 20 & 15 & 19 & 1 \end{matrix}$$



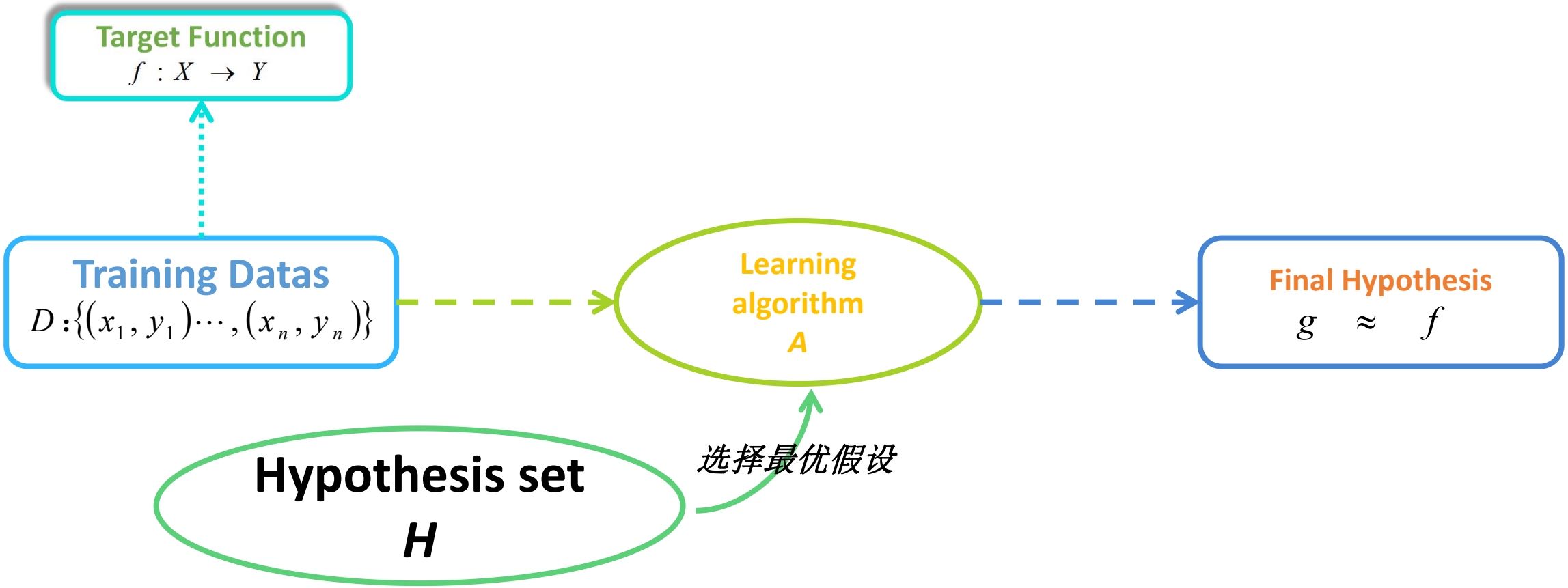
机器学习理性认识



- 目标函数f未知（无法得到）
- 假设函数g类似函数f，但是可能和函数f不同

机器学习中是无法找到一个完美的函数f

机器学习理性认识



- 机器学习

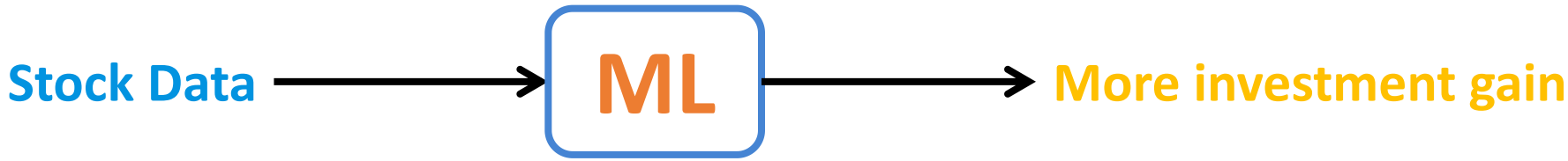
从数据中获得一个假设的函数g，使其非常接近目标函数f的效果。

机器学习概念

- A program can be said to learn from experience E with respect to some class of tasks T and performance measure P , If its performance at tasks in T, as measured by P, improves with experience E.
- 对于某给定的任务T，在合理的性能度量方案P的前提下，某计算机程序可以自主学习任务T的经验E；随着提供合适、优质、大量的经验E，该程序对于任务T的性能逐步提高。
- 其中重要的机器学习对象：
- 任务Task T，一个或多个、经验Experience E、度量性能Performance P
- 即：随着任务的不断执行，经验的累积会带来计算机性能的提升。
- 美国卡内基梅隆大学（Carnegie Mellon University）机器学习研究领域的著名教授Tom Mitchell对机器学习的经典定义

机器学习概念

- 算法(T)：根据业务需要和数据特征选择的相关算法， 也就是一个数学公式
- 模型(E)：基于数据和算法构建出来的模型
- 评估/测试(P)：对模型进行评估的策略

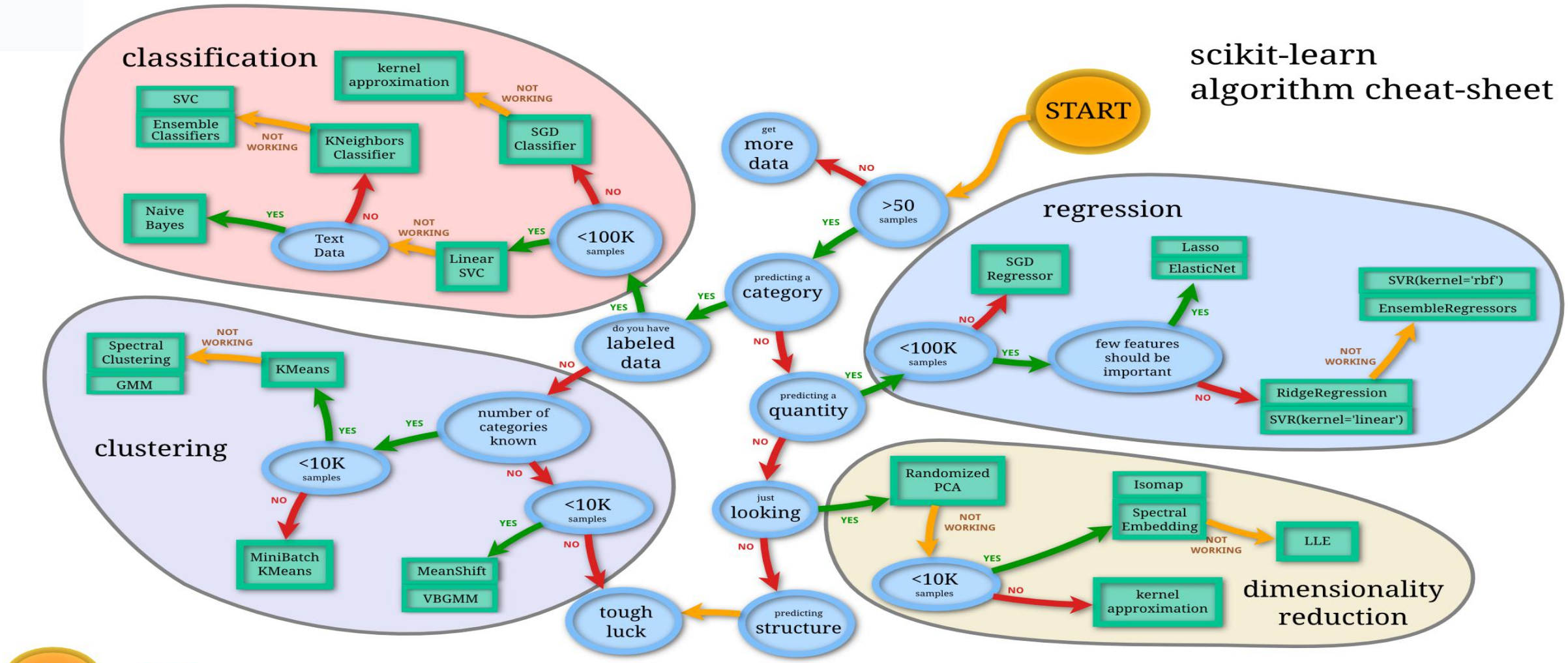


机器学习概念性含义

- 机器学习是人工智能的一个分支。我们使用计算机设计一个**系统**，使它能够根据提供的**训练数据**按照一定的方式来**学习**；随着训练次数的增加，该系统可以在性能上不断学习和改进；通过参数优化的学习模型，能够用于预测相关问题的输出。

机器学习之常见应用框架

- scikit-learn(Python)(授课环境)
  - http://scikit-learn.org/stable/
  - 建议Anaconda安装方式；pip不建议，pip案例命令：pip install scikit-learn==0.19.1
- Mahout(Hadoop生态圈基于MapReduce)
  - http://mahout.apache.org/
- Spark MLlib
  - http://spark.apache.org/



http://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

机器学习之商业场景

一 商业场景概览

1、数据挖掘
2、风控
3、CTR、搜索排序、量化投资
4、CV、NLP、无人车

二 工业4.0

AI视觉缺陷检测优势
1、准确率高
2、实时性
3、可对缺陷分类，及时预警设备维护，极大减少停产风险

三 精准营销

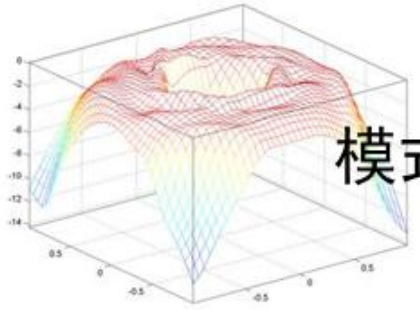
营销响应预测

1、数据清洗
2、特征工程
3、模型建立
4、模型评估
5、模型调优

四 买家秀

机器学习+深度学习

机器学习之商业场景



模式识别

计算机视觉

数据挖掘

机器学习

语音识别

统计学习

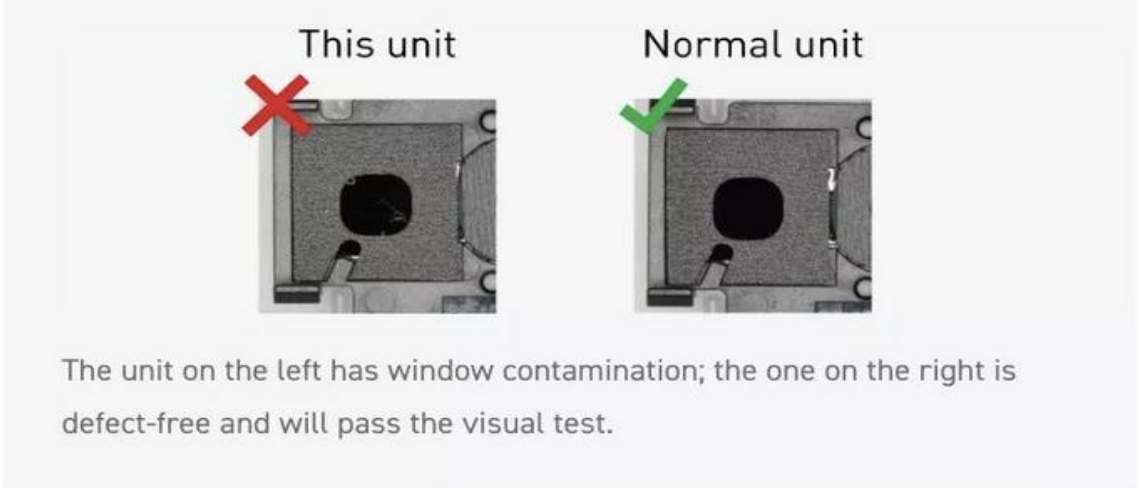
自然语言处理

机器学习、数据分析、数据挖掘区别与联系

- **数据分析**：数据分析是指用适当的统计分析方法对收集的大量数据进行分析，并提取有用的信息，以及形成结论，从而对数据进行详细的研究和概括过程。在实际工作中，数据分析可帮助人们做出判断；数据分析一般而言可以分为统计分析、探索性数据分析和验证性数据分析三大类。
- **数据挖掘**：一般指从大量的数据中通过算法搜索隐藏于其中的信息的过程。通常通过统计、检索、机器学习、模式匹配等诸多方法来实现这个过程。
- **机器学习**：是数据分析和数据挖掘的一种比较常用、比较好的手段。

智能制造=工业4.0

传统自动光学检测(AOI)方法问题：
1、贵；
2、误报率过高，误报率高达50%，需要安排人力后端验证；



图|硅谷公司 Instrumental 专攻 AI 瑕疵检测市场（来源：Instrumental 网站）

机器学习**算法优势**：
1、准确率提升；
2、检测速度快，**且可分类缺陷；**
3、后续可以进一步诊断瑕疵出现的原因

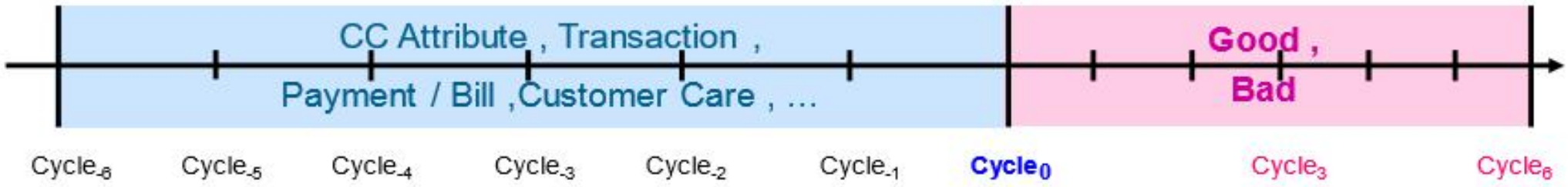
https://tianchi.aliyun.com/competition/entrance/231748/introduction?spm=5176.12281949.1003.1.68152448xsT4Kz
https://tianchi.aliyun.com/competition/entrance/231693/introduction?spm=5176.12281973.1005.4.3dd54c2aFDwFL5

精准营销-客户响应预测

建模逻辑 & 取数逻辑

项目描述

- 基于客户过去 18 个月以来的消费数据和app行为数据 等
- 预测 15 天内，若给该客户发送营销信息（电话、短信、app push），客户是否会产生消费 （是 = 1,否 = 0)



观察窗口

分析客户群：

往来时长：>=6个月

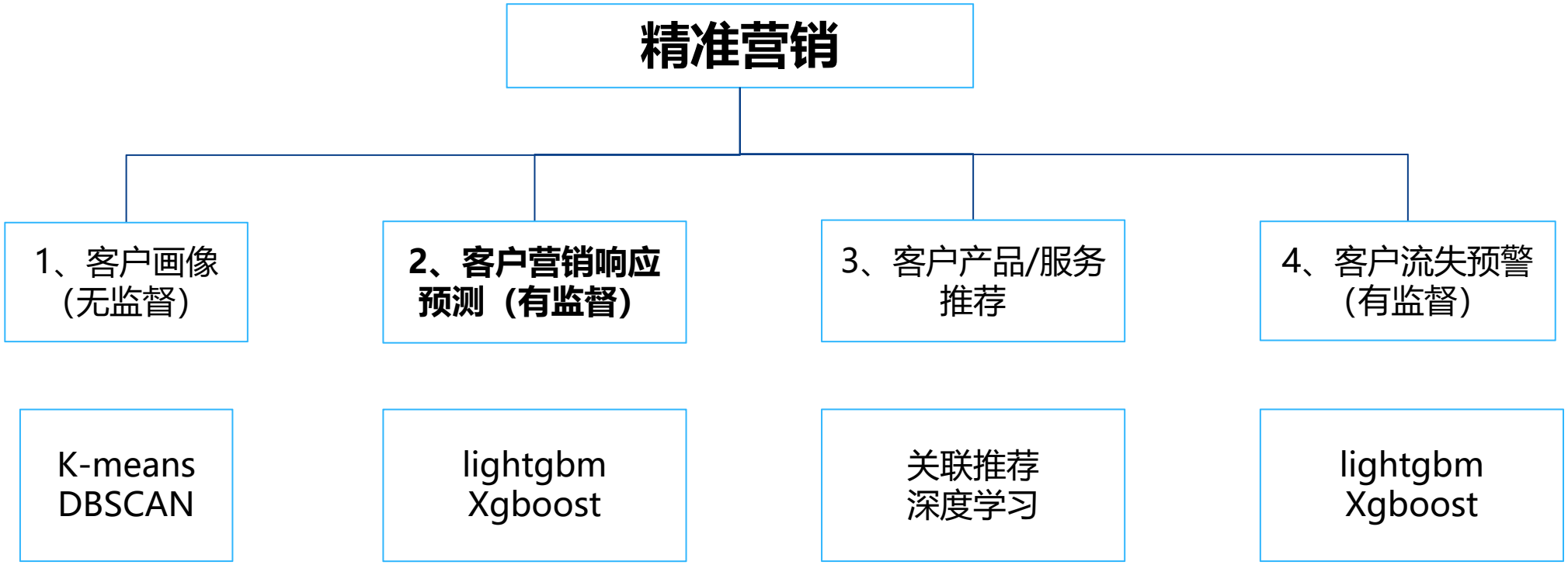
活跃客户：至少有一次购买记录

预测窗口

未来15天内，是否会产生购买

精准营销-客户响应预测

客户响应预测模型：根据客户的基础信息，历史消费信息、app 行为等，预测客户对营销活动的响应程度（响应1，不响应0）



精准营销-客户响应预测

数据准备

预测变量 (features)

客户人口统计信息	客户运营商基本信息	客户历史消费记录	客户app行为信息
▪年龄 ▪收入 ▪性别 ▪教育水平 ▪婚姻状况 ▪所在城市 ▪所在城市人均GDP和可支配收入 ▪城市外来人口数量占总人口数量 ▪城市等级 ……	▪客户手机在网时长/手机品牌、手机型号 ▪客户所拥有手机号码数量 ▪流量数据（流量和流量时段分布） ▪手机账单信息（min/max账单金额，欠费金额） ▪app安装类别，总时长、活跃天数 ……	订单总次数（sku等）（1/3/6/9/12/18） ▪平均消费消费金额 ▪近1/3/6/9/12/18个月平均消费金额 ▪近1/3/6/9/12/18个月购买促销产品次数/总购买次数 ▪近1/3/6/9/12/18个月购买促销产品金额/总购买金额 ▪客户收藏产品总数 ▪分期笔数/总支付笔数 ▪实付金额笔数/总订单次数 ▪收货地址特征（小区，小区房价，工作还是住宅，收货地址个数） ……	▪登录频率（日周月） ▪登录时间、时段（工作日/节假日 上午/下午/晚上） ▪客户评论产品次数/好评次数 ▪客户评论emoji表情分类 ▪客户平均每次登录时长，平均浏览商品数量，平均浏览页面数量 ……

精准营销-客户响应预测

建模步骤

数据清洗

1、分类变量编码错误
2、异常值处理
3、缺失值处理

特征构造和特征选择

1、特征构造 vox-box 变换(取 log sqrt 1/x 等等)，多项式变换
2、剔除共线性
3、降维

建立基础模型、调参

1、Lightgbm 追求单模型精度
2、调参：手动、GridSearch Hyperopt等

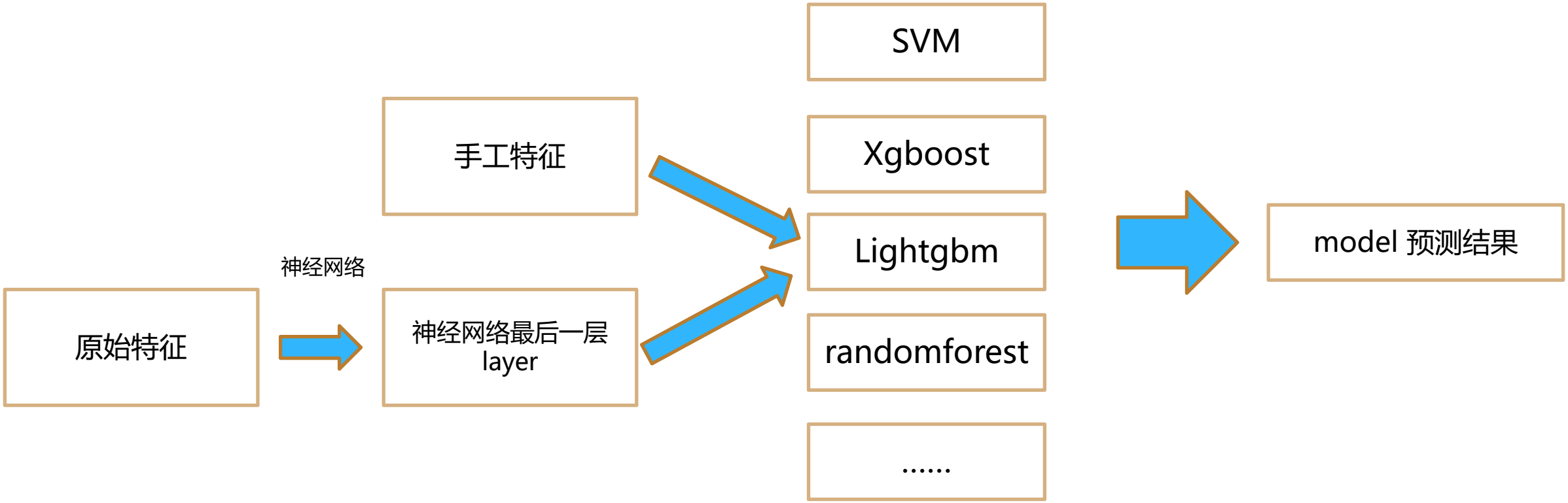
模型集成

1、单模型 精度ok 的基础上加 ensemble，提升模型精度
2、选择 ensemble 算法

评估模型

1、ROC曲线
2、Accuracy--Fscore
3、模型实施监测：

精准营销-客户响应预测


SVM
手工特征
Xgboost
神经网络
Lightgbm
model 预测结果
原始特征
神经网络最后一层
layer
randomforest
……

买家秀选择

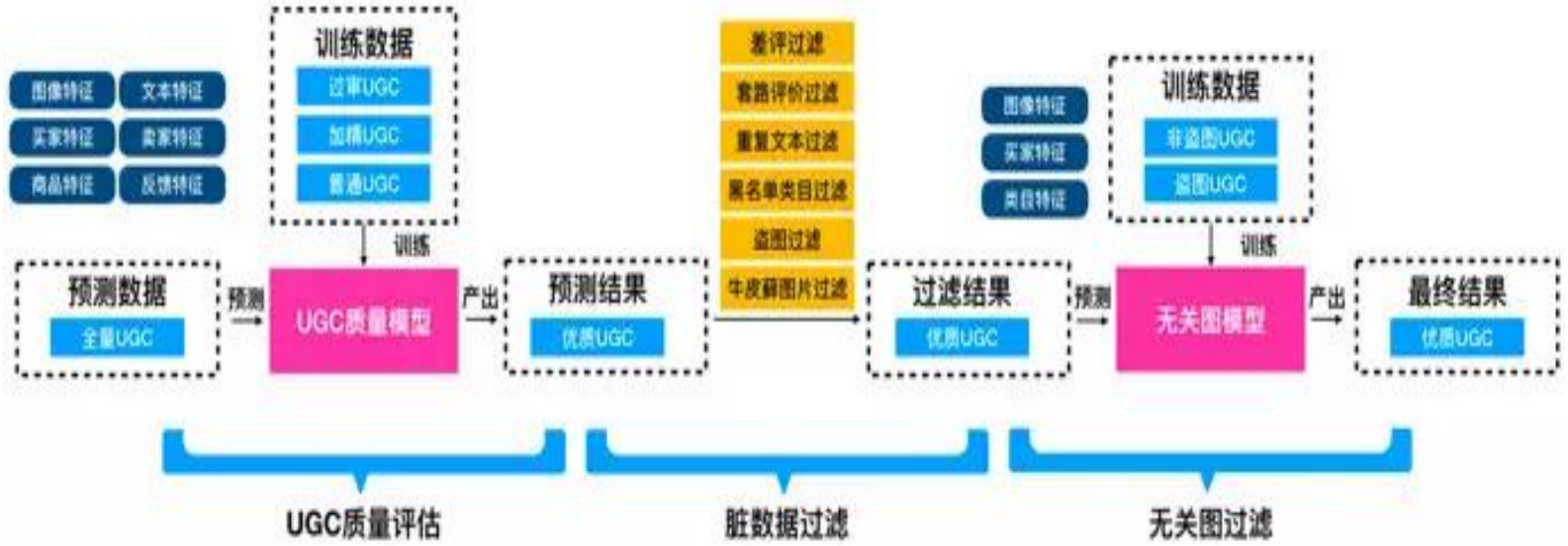
业务需求：

从当前的买家秀中抽取一批高质量的图文内容(UGC)，作为社区运营的启动数据

买家秀选择

- 全量UGC（User Generated Content）是指所有含图或含视频的买家秀
- 过审UGC是指最终**审核通过**的高质量买家秀
- 加精UGC是指商家认可的买家秀
- 普通UGC则是上述两种情况以外的其他买家秀

Target：通过机器学习技术来预测UGC内容的好坏，挖掘更多的高质量的UGC内容。



买家秀选择 ♥

特征构造

文本长度	文本段落个数	图片个数
视频个数	商品个数	商品总赞数
商品平均赞数	商品赞数中位数	商品最大赞数
商品UGC总数	商品卖出总数	买家总赞数
买家平均赞数	买家赞数中位数	买家UGC总数
该UGC赞数	图片平均高度和宽度	买家操作系统
买家手机价位	买家性别	买家年龄

标签： (label)

1、审核通过的数据标记为 2

2、运营审核未通过（商家已加精）的数据标记为1

3、将商家未加精的数据标记为0

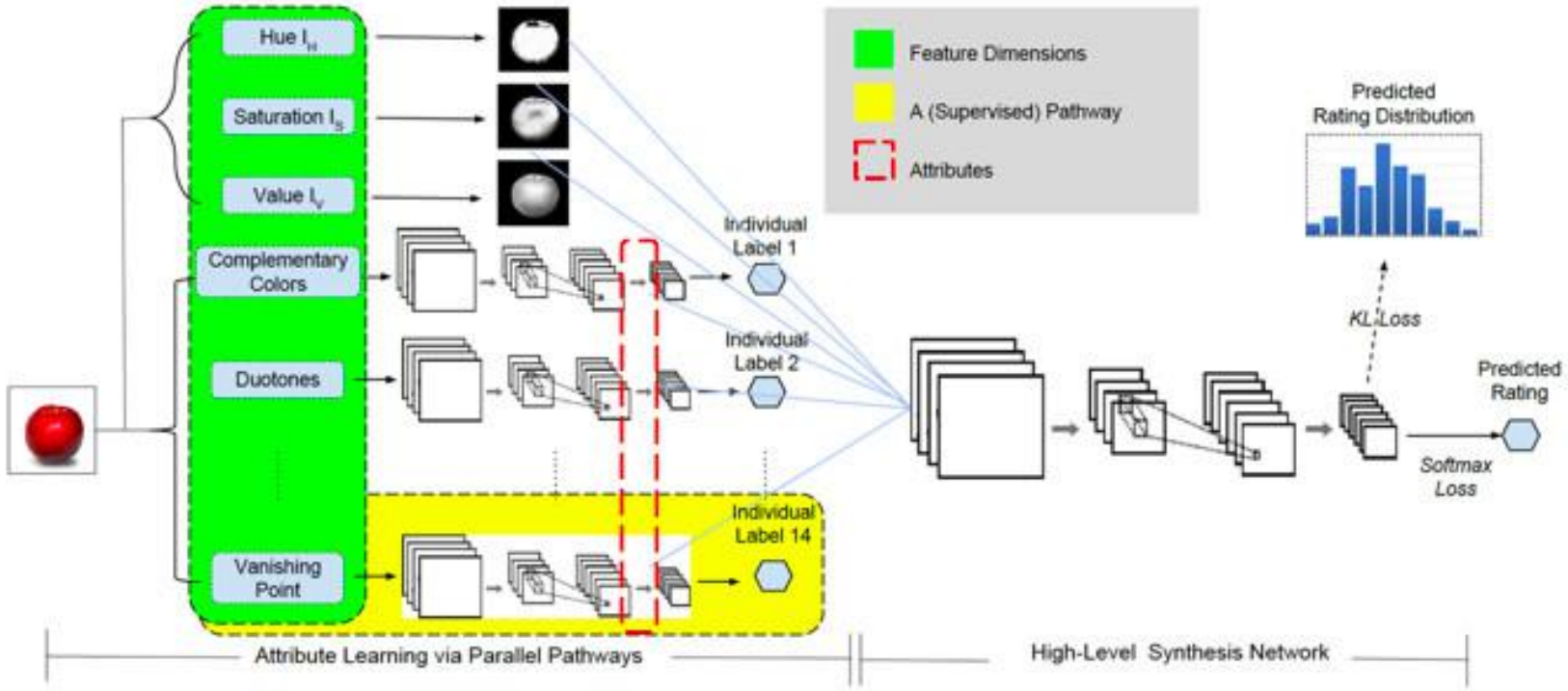
成果：使用GBDT，准确率是50%

缺点：图片不够美观！！！

买家秀选择

1、利用开源美感数据集：AVA database （A Large-Scale Database for Aesthetic Visual Analysis)
美学相关的数据库，包含25万余张图片，每张图片包含语义标注（如自然风光、天空等）、图片风格标注（如互补色、双色调等）和图片美感评分（由数十到数百人评出1-10分）；

2、通过AVA数据集提供的图片风格标签，学习图片风格的隐藏层特征，将图片风格的隐藏层特征和图片经过HSV变换后的特征结合起来，以AVA数据集提供的图片美感分为监督，学习图片的美感特征。



机器学习之商业场景总结

一 商业场景概览

1、数据挖掘
2、风控
3、CTR、搜索排序、量化投资
4、CV、NLP、无人车

二 工业4.0

AI视觉缺陷检测优势
1、准确率高
2、实时性
3、可对缺陷分类，及时预警设备维护，极大减少停产风险

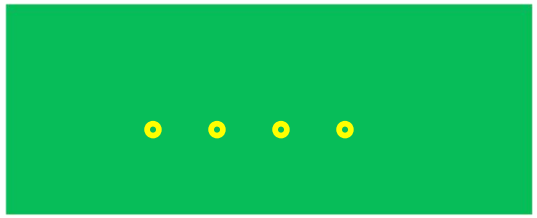
三 精准营销

营销响应预测

1、数据清洗
2、特征工程
3、模型建立
4、模型评估
5、模型调优

四 买家秀

机器学习+深度学习



机器学习分类

- 有监督学习
  - 用已知某种或某些特性的样本作为训练集，以建立一个数学模型，再用已建立的模型来预测未知样本，此种方法被称为有监督学习，是最常用的一种机器学习方法。是从**标签**化训练数据集中推断出模型的机器学习任务。
- 无监督学习
  - 与监督学习相比，无监督学习的训练集中没有人为的标注的结果，在非监督的学习过程中，数据并不被特别标识，学习模型是为了推断出数据的一些内在结构。
- 半监督学习
  - 考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的问题，是有监督学习和无监督学习的结合

有监督学习(分类类型的算法)

- **判别式模型**(Discriminative Model)：直接对条件概率p(y|x)进行建模，常见判别模型有：Logistic回归、决策树、支持向量机SVM、k近邻、神经网络等；
- **生成式模型**(Generative Model)：对联合分布概率p(x,y)进行建模，常见生成式模型有：隐马尔可夫模型HMM、朴素贝叶斯模型、高斯混合模型GMM、LDA等；
- 生成式模型更普适；判别式模型更直接，目标性更强
- 生成式模型关注数据是如何产生的，寻找的是数据分布模型；判别式模型关注的数据的差异性，寻找的是分类面
- 由生成式模型可以产生判别式模型，但是由判别式模式没法形成生成式模型

无监督学习

- 无监督学习试图学习或者**提取**数据背后的**数据特征**，或者从数据中抽取出重要的特征信息，常见的算法有聚类、降维、文本处理(特征抽取)等。
- 无监督学习一般是作为有监督学习的前期数据处理，功能是从原始数据中抽取出必要的标签信息。

半监督学习(SSL)

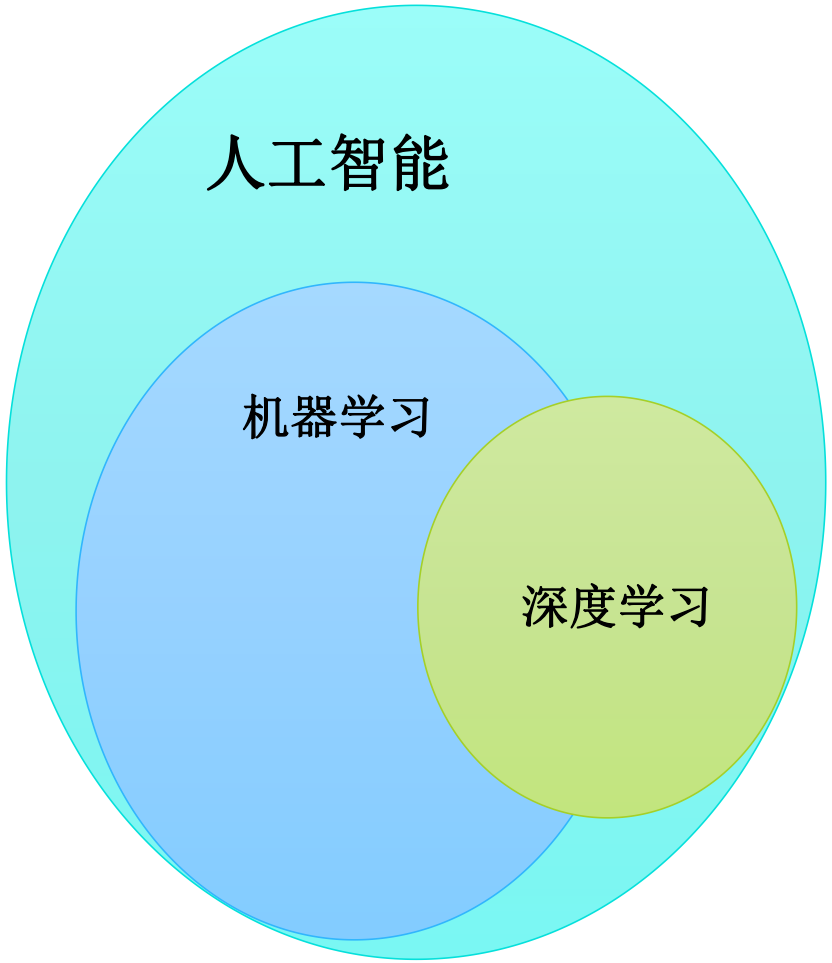
- 主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的问题。半监督学习对于减少标注代价，提高学习机器性能具有非常重大的实际意义。
- SSL的成立依赖于模型假设，主要分为三大类：平滑假设、聚类假设、流行假设；其中流行假设更具有普片性。
- SSL类型的算法主要分为四大类：半监督分类、半监督回归、半监督聚类、半监督降维。
- 缺点：抗干扰能力弱，仅适合于实验室环境，其现实意义还没有体现出来；未来的发展主要是聚焦于新模型假设的产生。

机器学习分类2

- 分类
  - 通过分类模型，将样本数据集中的样本映射到某个给定的类别中(在模型构建之前，类别信息已经确定了。)
- 聚类
  - 通过聚类模型，将样本数据集中的样本分为几个类别，属于同一类别的样本相似性比较大
- 回归
  - 反映了样本数据集中样本的属性值的特性，通过函数表达样本映射的关系来发现属性值之间的依赖关系
- 关联规则
  - 获取隐藏在数据项之间的关联或相互关系，即可以根据一个数据项的出现推导出其他数据项的出现频率。

机器学习、人工智能和深度学习的关系

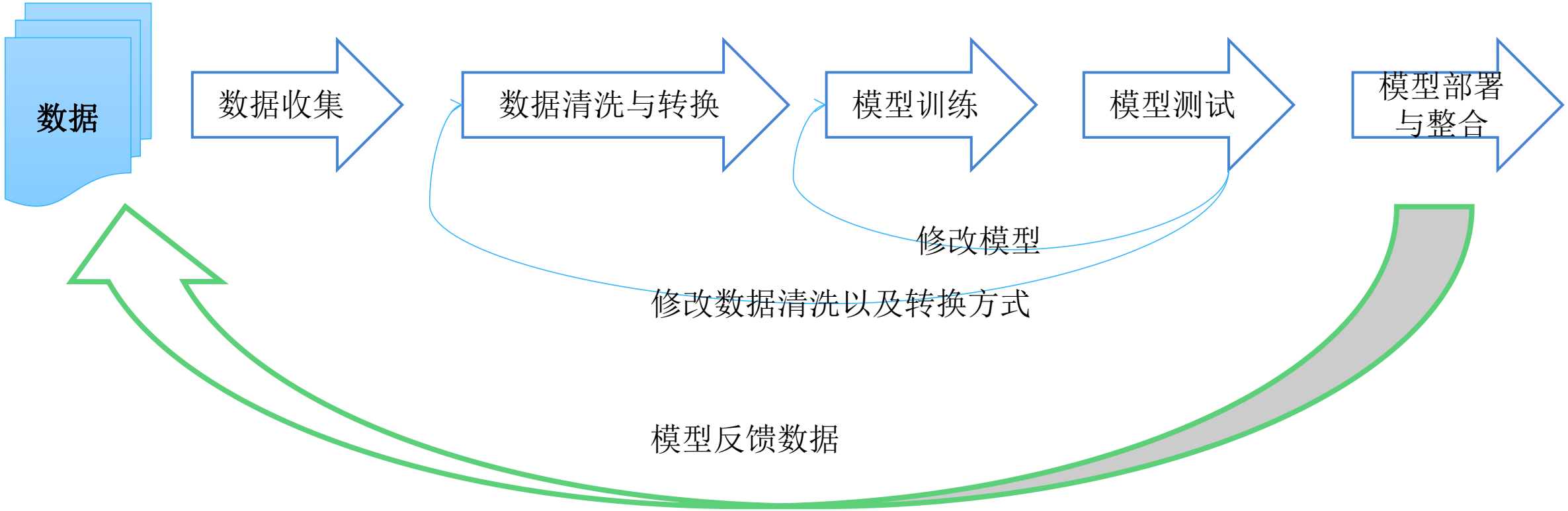
- 深度学习是机器学习的子类；深度学习是基于传统的神经网络算法发展到多隐层的一种算法体现。
- 机器学习是人工智能的一个子类；



机器学习开发流程

- 数据收集
- 数据预处理
- 特征提取
- 模型构建
- 模型测试评估
- 投入使用(模型部署与整合)
- 迭代优化

机器学习开发流程


数据
数据收集
数据清洗与转换
模型训练
模型测试
模型部署
与整合
修改模型
修改数据清洗以及转换方式
模型反馈数据

机器学习一般流程

——生活案例

数据收集 → 数据清洗 → 特征工程 → 数据建模

数据收集与存储

- 数据来源：
  - 用户访问行为数据
  - 业务数据
  - 外部第三方数据
- 数据存储：
  - 需要存储的数据：原始数据、预处理后数据、模型结果
  - 存储设施：磁盘、mysql、HDFS、HBase、Solr、Elasticsearch、Kafka、Redis等
- 数据收集方式：
  - Flume & Kafka

机器学习可用公开数据集

- 在实际工作中，我们可以使用业务数据进行机器学习开发，但是在学习过程中，没有业务数据，此时可以使用公开的数据集进行开发，常用数据集如下：
  - http://archive.ics.uci.edu/ml/datasets.html
  - https://aws.amazon.com/cn/public-datasets/
  - https://www.kaggle.com/competitions
  - http://www.kdnuggets.com/datasets/index.html
  - http://www.sogou.com/labs/resource/list_pingce.php
  - https://tianchi.aliyun.com/datalab/index.htm
  - http://www.pkbigdata.com/common/cmptIndex.html

数据清洗和转换

- 实际生产环境中机器学习比较耗时的一部分
- 大部分的机器学习模型所处理的都是特征，特征通常是输入变量所对应的可用于模型的数值表示
- 大部分情况下，收集得到的数据需要经过预处理后才能够为算法所使用，预处理的操作主要包括以下几个部分：
  - 数据过滤
  - 处理数据缺失
  - 处理可能的异常、错误或者异常值
  - 合并多个数据源数据
  - 数据汇总

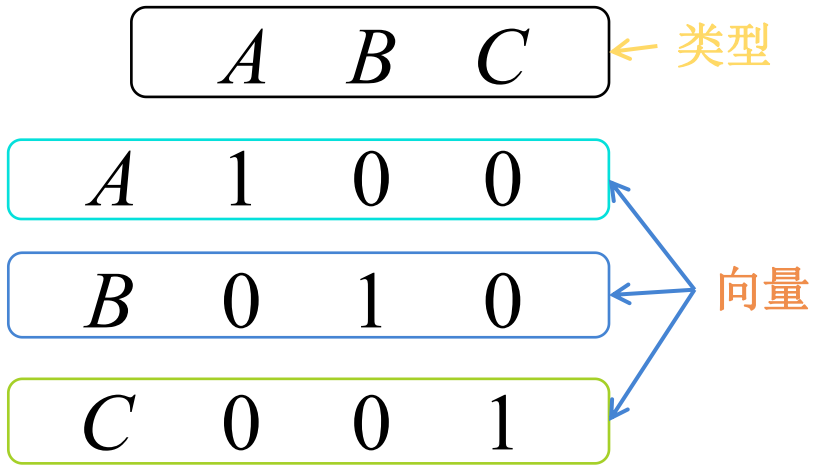
数据清洗和转换

- 对数据进行初步的预处理，需要将其转换为一种适合机器学习模型的表示形式，对许多模型类型来说，这种表示就是包含数值数据的向量或者矩阵
  - 将类别数据编码成为对应的数值表示(一般使用1-of-k\哑编码方法)
  - 从文本数据中提取有用的数据(一般使用词袋法或者TF-IDF)
  - 处理图像或者音频数据(像素、声波、音频、振幅等<傅里叶变换>)
  - 对特征进行正则化、标准化，以保证同一模型的不同输入变量的取值范围相同
  - 数值数据转换为类别数据以减少变量的值，比如年龄分段
  - 对数值数据进行转换，比如对数转换
  - 对现有变量进行组合或转换以生成新特征(基于对数据以及对业务的理解)，比如平均数 (做虚拟变量)，需要不断尝试才可以确定具体使用什么虚拟变量。

类型特征转换之1-of-k（哑编码）

- 功能：将非数值型的特征值转换为数值型的数据
- 描述：假设变量的取值有 k 个，如果对这些值用 1 到 k 编序，则可用维度为 k 的向量来表示一个变量的值。在这样的向量里，该取值所对应的序号所在的元素为1，其他元素均为0.

	T1	T2	T3
R1	*A*	1	2
R2	*A*	2	3
R3	*B*	3	3
R4	*C*	2	2
R5	C	1	2



	T1特征				
	T1-A	T1-B	T1-C	T2	T3
R1	1	0	0	1	2
R2	1	0	0	2	3
R3	0	1	0	3	3
R4	0	0	1	2	2
R5	0	0	1	1	2

文本数据抽取

- **词袋法：**将文本当作一个无序的数据集合，文本特征可以采用文本中的词条/单词T进行体现，那么文本中出现的所有词条及其出现的次数/频率就可以体现文档的特征
- **TF-IDF**: 词条的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降；也就是说**词条**在**当前文本**中出现的次数**越多**，表示该**词条**对**当前文本**的重要性**越高**，**词条**在**所有文本(语料库/训练数据集)**中出现的次数**越少**，说明这个**词条**对**文本**的重要性**越高**。TF(词频)指某个词条在文本中出现的次数，一般会将其进行归一化处理(该词条数量/该文档中所有词条数量)；IDF(逆向文件频率)指一个词条重要性的度量，一般计算方式为**语料库中总文件数目**除以包含**该词语的文件数目**，再将得到的商取对数得到。TF-IDF实际上是：**TF * IDF**

文本数据抽取

- 文档1内容：A(2)、B(1)、C(3)、D(9)、E(1)
- 文档2内容：A(1)、B(5)、C(2)、D(10)

	A	*B*	*C*	*D*	*E*
文 档 1	2	1	3	9	1
文 档 2	1	5	2	10	0

TF(*A* | 文档1)=1/8

TF(*B* | 文档1)=1/16

.....

TF(*E*|文档2)=0/18=0

IDF(A)=IDF(B)=IDF(C)=IDF(D)=2 / 2 = 1

IDF(E)=2/1=2

	A	*B*	*C*	*D*	*E*
文档1	1/8	1/16	3/16	9/16	1/8
文档2	1/18	5/18	1/9	5/9	0

模型训练及测试

- 模型选择：对特定任务最优建模方法的选择或者对特定模型最佳参数的选择。
- 在训练数据集上运行模型(算法)并在测试数据集中测试效果，迭代进行数据模型的修改，这种方式被称为**交叉验证**(将数据分为**训练集**和**测试集**，使用训练集构建模型，并使用测试集评估模型提供修改建议)
- 模型的选择会尽可能多的选择算法进行执行，并比较执行结果
- 模型的测试一般以下几个方面来进行比较，在分类算法中常见的指标分别是**准确率/召回率/精准率/F值(F1指标)**
  - 准确率(Accuracy)=提取出的正确样本数/总样本数
  - 召回率(Recall)=正确的正例样本数/样本中的正例样本数——覆盖率
  - 精准率(Precision)=正确的正例样本数/预测为正例的样本数
  - F值=Precision*Recall*2 / (Precision+Recall) (即F值为正确率和召回率的调和平均值)

模型训练及测试

		预测值	
		正例	**负例**
真实值	**正例**	真正例(A)	假负例(B)
	负例	假正例(C)	真负例(D)

A和D预测正确，B和C预测错误，测试计算结果为:

$$Accuracy = \frac{\#(A)+\#(D)}{\#(A)+\#(B)+\#(C)+\#(D)}$$

$$Recall = \frac{\#(A)}{\#(A)+\#(B)} \quad Precision = \frac{\#(A)}{\#(A)+\#(C)} \quad F = \frac{2*Recall*Precision}{Recall+Precision}$$

混淆矩阵

		predicted condition		
	total population	prediction positive	prediction negative	Prevalence = $\frac{\Sigma \text{ condition positive}}{\Sigma \text{ total population}}$
true condition	condition positive	**True Positive (TP)**	**False Negative (FN)** (type II error)	True Positive Rate (TPR), Sensitivity, Recall, Probability of Detection = $\frac{\Sigma \text{ TP}}{\Sigma \text{ condition positive}}$
	condition negative	**False Positive (FP)** (Type I error)	**True Negative (TN)**	False Positive Rate (FPR), Fall-out, Probability of False Alarm = $\frac{\Sigma \text{ FP}}{\Sigma \text{ condition negative}}$
	Accuracy = $\frac{\Sigma \text{ TP} + \Sigma \text{ TN}}{\Sigma \text{ total population}}$	Positive Predictive Value (PPV), Precision = $\frac{\Sigma \text{ TP}}{\Sigma \text{ prediction positive}}$	False Omission Rate (FOR) = $\frac{\Sigma \text{ FN}}{\Sigma \text{ prediction negative}}$	Positive Likelihood Ratio (LR+) = $\frac{\text{TPR}}{\text{FPR}}$
		False Discovery Rate (FDR) = $\frac{\Sigma \text{ FP}}{\Sigma \text{ prediction positive}}$	Negative Predictive Value (NPV) = $\frac{\Sigma \text{ TN}}{\Sigma \text{ prediction negative}}$	Negative Likelihood Ratio (LR−) = $\frac{\text{FNR}}{\text{TNR}}$

				Diagnostic Odds Ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$

http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matrix/confusion_matrix.html
https://en.wikipedia.org/wiki/Precision_and_recall

模型评估

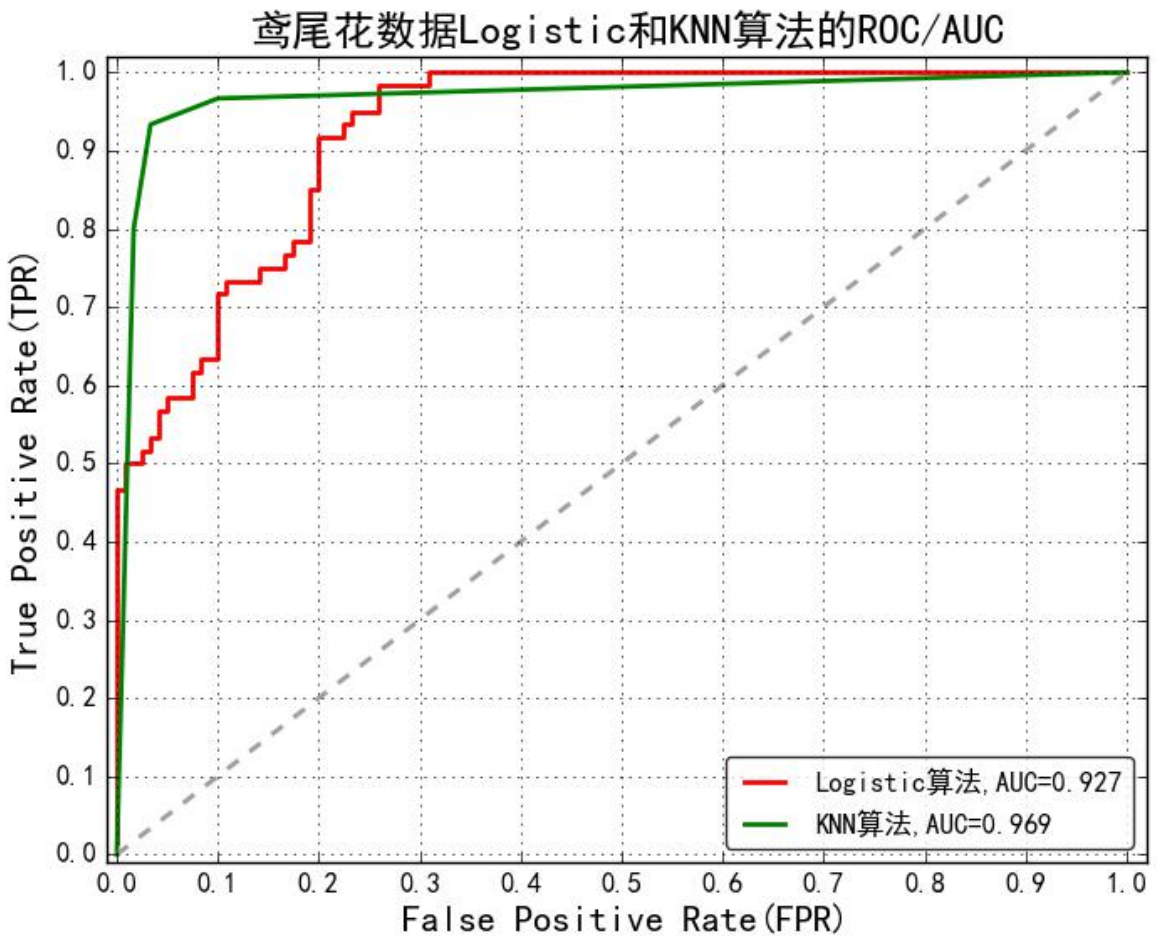
- 准确率 Accuracy
  - $\text{Accuracy} = \frac{\#(True\ positive)+\#(True\ negative)}{\#(True\ positive)+\#(True\ negative)+\#(False\ positive)+\#(False\ negative)}$
- 召回率 Recall
  - $\text{Recall} = \frac{\#(True\ positive)}{\#(True\ positive)+\#(False\ negative)}$
- 精确率 Precision
  - $Precision = \frac{\#(True\ positive)}{\#(True\ positive)+\#(False\ positive)}$
- F1指标 F1 measure
  - $F1\ measure = \frac{2}{\frac{1}{Precision}+\frac{1}{Recall}}$

ROC

- ROC（Receiver Operating Characteristic）最初源于20世纪70年代的信号检测理论，描述的是分类混淆矩阵中FPR-TPR两个量之间的相对变化情况，ROC曲线的纵轴是“真正例率”（True Positive Rate 简称TPR），横轴是“假正例率”（False Positive Rate 简称FPR）。
- 如果二元分类器输出的是对正样本的一个分类概率值，当取不同阈值时会得到不同的混淆矩阵，对应于ROC曲线上的一个点。那么ROC曲线就反映了FPR与TPR之间权衡的情况，通俗地来说，即在TPR随着FPR递增的情况下，谁增长得更快，快多少的问题。TPR增长得越快，曲线越往上屈，AUC就越大，反映了模型的分类性能就越好。当正负样本不平衡时，这种模型评价方式比起一般的精确度评价方式的好处尤其显著。

ROC

- ROC曲线


鸢尾花数据Logistic和KNN算法的ROC/AUC
True Positive Rate(TPR)
False Positive Rate(FPR)
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Logistic算法, AUC=0.927
KNN算法, AUC=0.969

AUC

- AUC的值越大表达模型越好
- AUC（Area Under Curve）被定义为ROC曲线下的面积，显然这个面积的数值不会大于1。又由于ROC曲线一般都处于y=x这条直线的上方，所以AUC的取值范围在0.5和1之间。使用AUC值作为评价标准是因为很多时候ROC曲线并不能清晰的说明哪个分类器的效果更好，而AUC作为数值可以直观的评价分类器的好坏，值越大越好。
- AUC = 1，是完美分类器，采用这个预测模型时，不管设定什么阈值都能得出完美预测。绝大多数预测的场合，不存在完美分类器。
- 0.5 < AUC < 1，优于随机猜测。这个分类器（模型）妥善设定阈值的话，能有预测价值。
- AUC = 0.5，跟随机猜测一样（例：丢铜板），模型没有预测价值。
- AUC < 0.5，比随机猜测还差；但只要总是反预测而行，就优于随机猜测。

模型评估

- 回归结果度量
  - explained_variance_score：可解释方差的回归评分函数

$$\text{explain\_variance}(y_i, \hat{y}_i) = 1 - \frac{var(y-\hat{y})}{var(y)}$$

  - mean_absolute_error：平均绝对误差 $MAE = \frac{1}{m}|y_i - \hat{y}_i|$
  - mean_squared_error：平均平方误差 $MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$
  - r2_score：R^2值 $R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}$ $\bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$

模型评估总结_分类算法评估方式

指标	描述	scikit-learn函数
Precision	精确度	from sklearn.metrics import precision_score
Recall	召回率	from sklearn.metrics import recall_score
F1	F1指标	from sklearn.metrics import f1_score
Confusion Matrix	混淆矩阵	from sklearn.metrics import confusion_matrix
ROC	ROC曲线	from sklearn.metrics import roc
AUC	ROC曲线下的面积	from sklearn.metrics import auc

模型评估总结_回归算法评估方式

指标	描述	scikit-learn函数
Mean Square Error (MSE, RMSE)	平均方差	from sklearn.metrics import mean_squared_error
Absolute Error (MAE, RAE)	绝对误差	from sklearn.metrics import mean_absolute_error, median_absolute_error
R-Squared	R平方值	from sklearn.metrics import r2_score

模型部署和整合

- 当模型构建好后，将训练好的模型进行部署
  - 方式一：直接使用训练好的模型对数据做一个预测，然后将预测结果保存数据库中。
  - 方式二：直接将模型持久化为磁盘文件的形式，在需要的代码处从磁盘中恢复模型对象，然后使用恢复的模型对象对数据做一个预测。
  - 方式三：直接将模型参数保存到数据库中，然后在需要的代码处直接从数据库把模型参数加载到代码中，然后根据模型算法原理使用模型参数对数据做一个预测。
- 模型需要周期性的进行修改、调优：
  - 一个月、一周

模型的监控与反馈

- 当模型一旦投入到实际生产环境中，模型的效果监控是非常重要的，往往需要关注业务效果和用户体验，所以有时候会进行测试
- 模型需要对用户的反馈进行响应操作，即进行模型修改，但是要注意异常反馈信息对模型的影响，故需要进行必要的数据预处理操作

本课程软件环境

- Anaconda3-5.0.1-Windows-x86_64
- Python 3.6.3
- scikit-learn 0.19.1 --> 只要版本不要比我的低就可以了