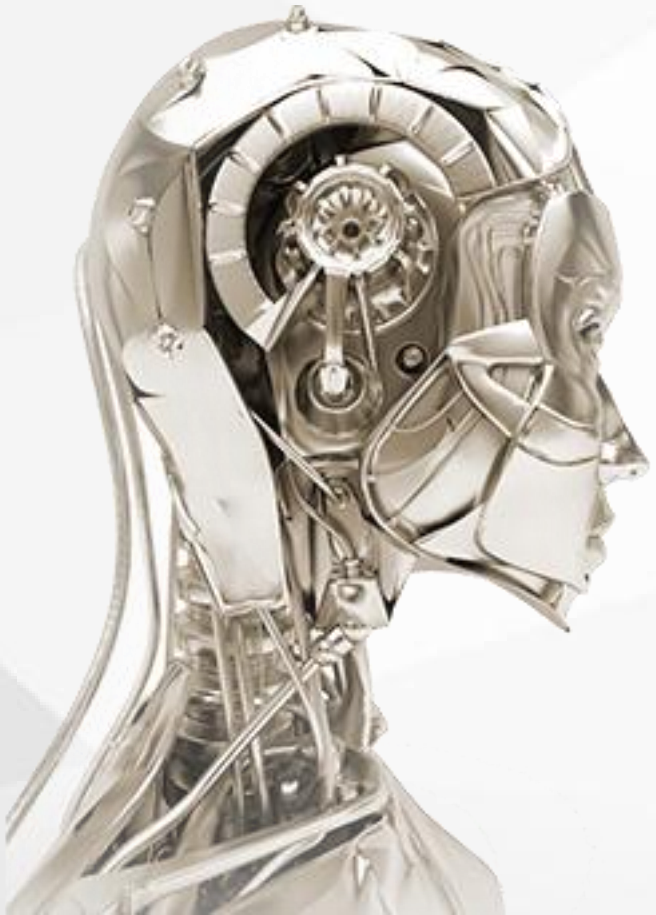
人工智能之机器学习



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、数据挖掘的区别与联系

05 机器学习类型

06 机器学习数据处理流程

Machine Learning

What society thinks I do

What my friend thinks I do

What 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 $i=1,\dots n$)

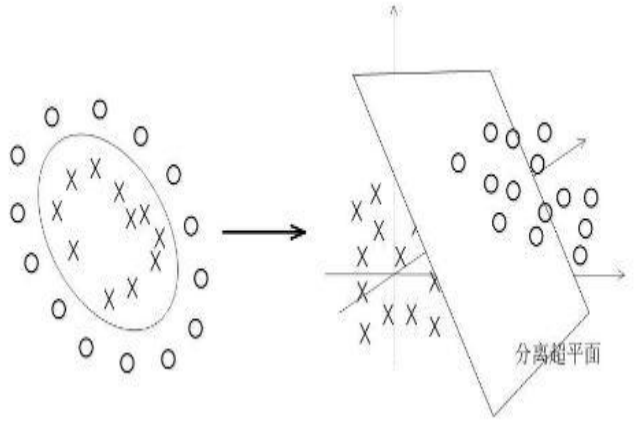
$$y_i(\mathbf{w}\cdot\mathbf{x_i}-b)\geq 1-\xi_i,\qquad \xi_i\geq 0$$

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

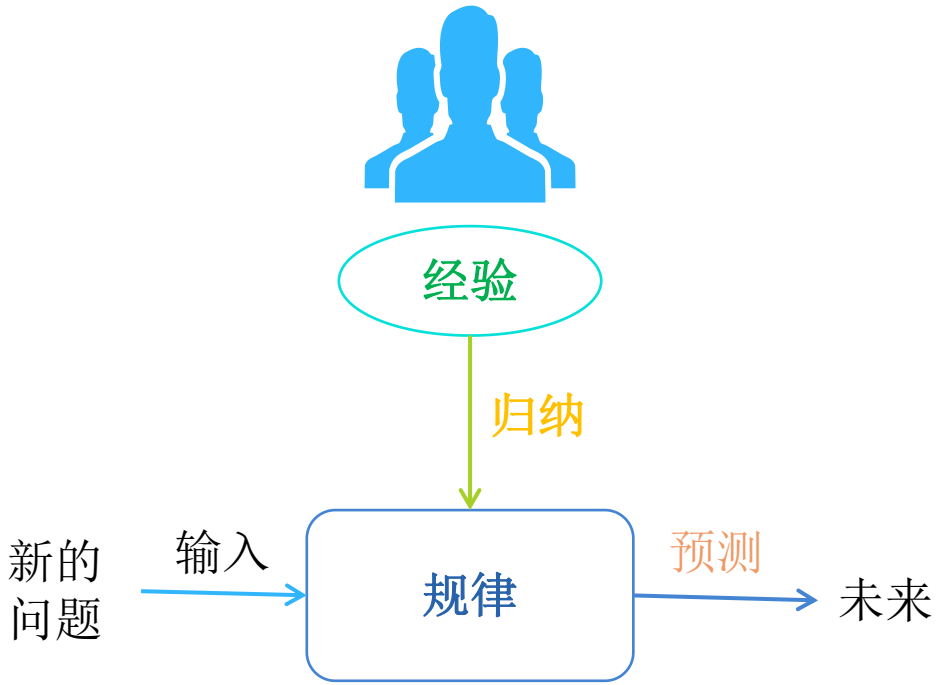
```
In [1]: from sklearn import svm

In [ ]:
```

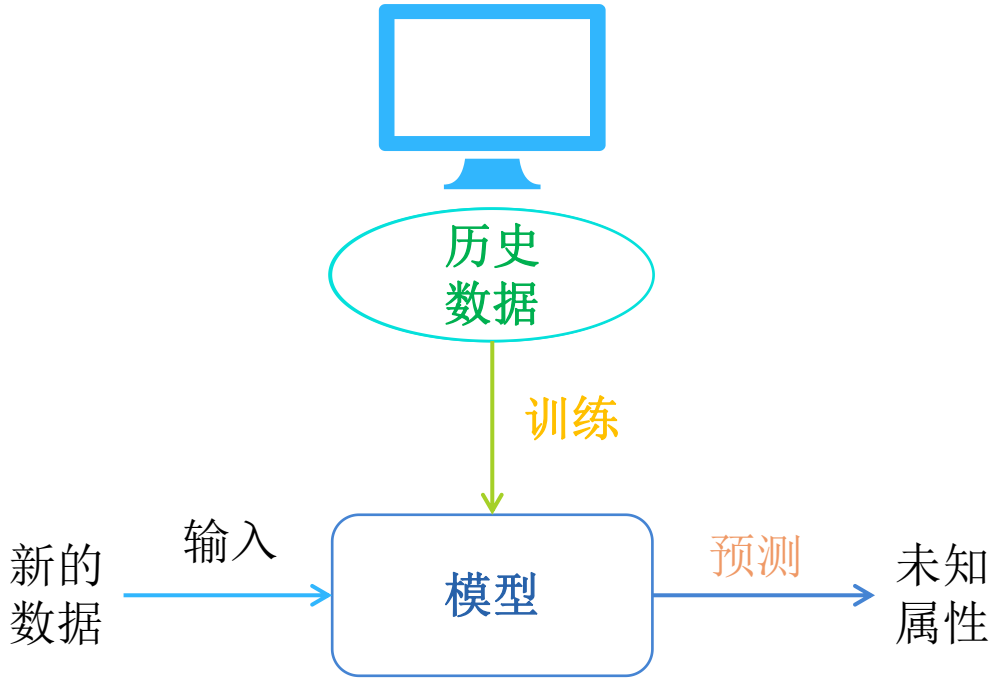
What I really do

机器学习定义

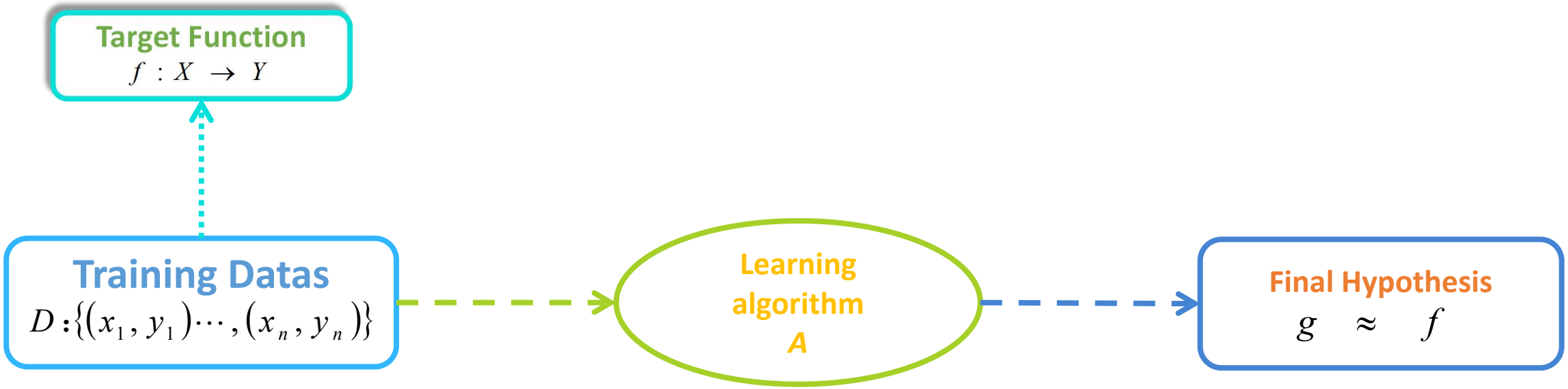
- **Machine Learning**(ML) is a scientific discipline that deals with the construction and study of algorithms that can learn from data.
- 机器学习是一门从数据中研究算法的科学学科。
- 机器学习直白来讲，是根据已有的数据，进行算法选择，并基于算法和数据构建模型，最终对未来进行预测;
- 备注：机器学习就是一个模拟**人决策过程**的一种程序结构。

机器学习/人工智能理性认识





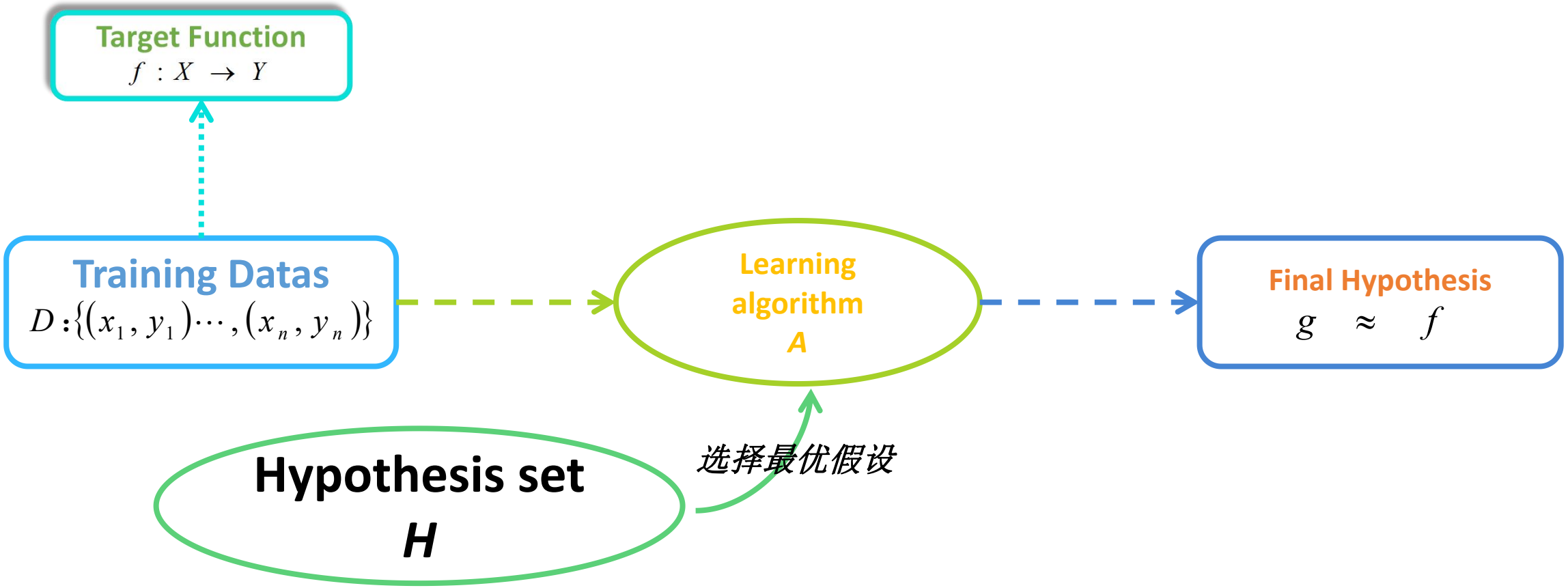
机器学习理性认识



- 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

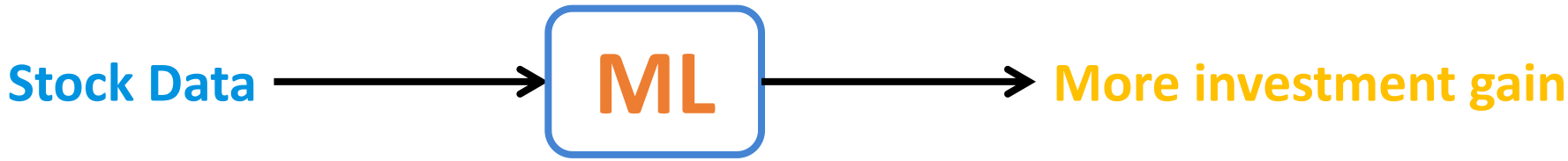


- 机器学习

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

- 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

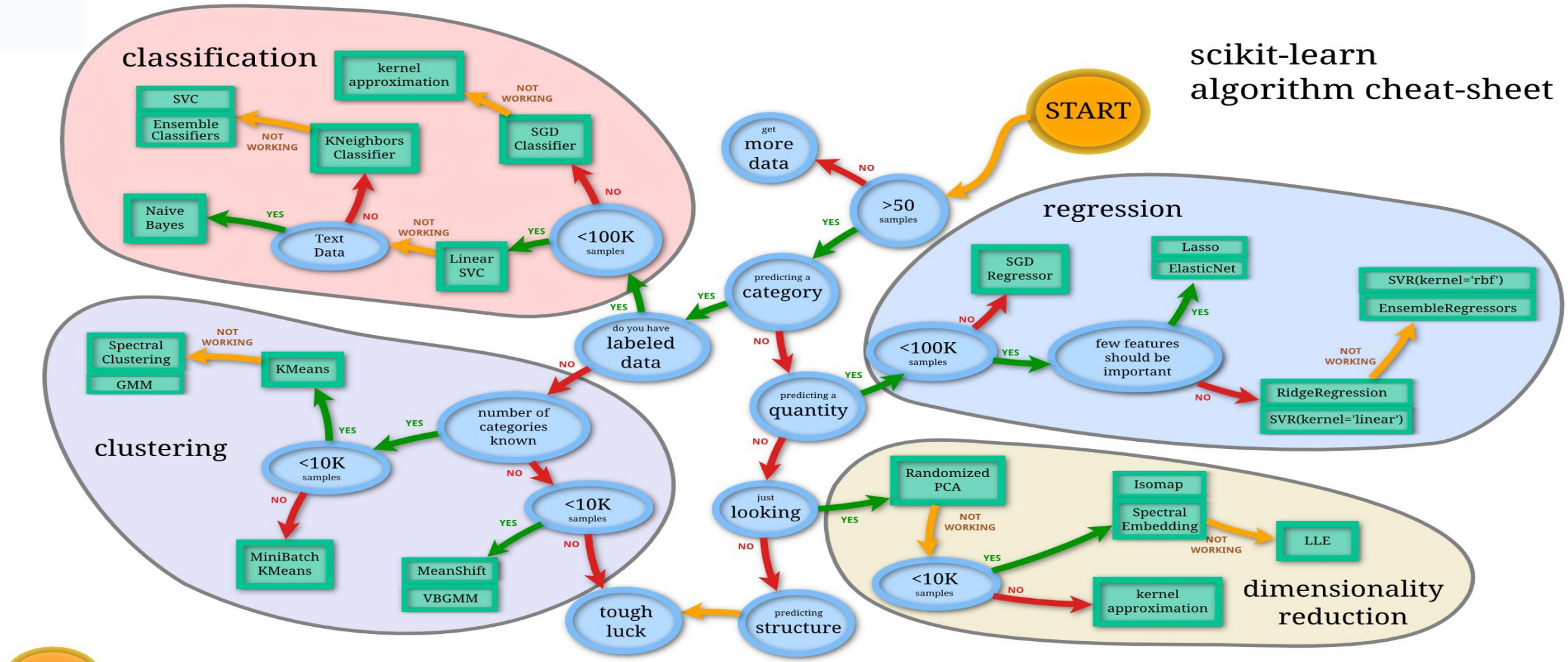


机器学习之常见应用框架

- 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

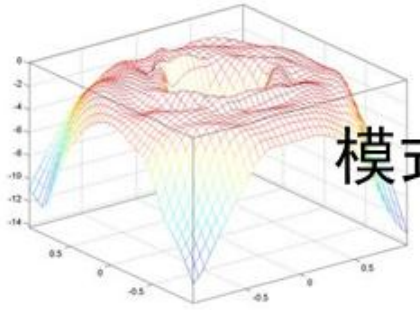
- Mahout(Hadoop生态圈基于MapReduce)
  - http://mahout.apache.org/

- 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



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

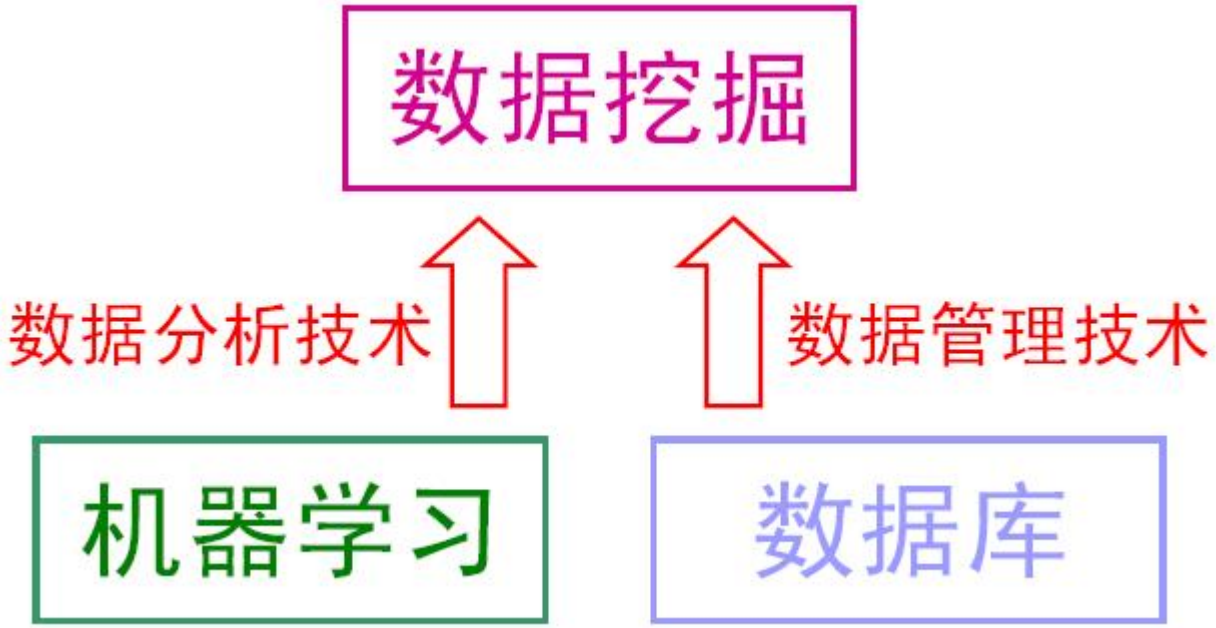
机器学习应用

- 数据挖掘
- 计算机视觉
- 自然语言处理
- 生物特征识别
- 搜索引擎
- 医学诊断
- 检测信用卡欺诈
- 证券市场分析
- DNA序列测序
- 语音和手写识别
- 战略游戏
- 机器人
- ...

机器学习和数据挖掘区别与联系

- 机器学习是数据挖掘的重要工具。
- 数据挖掘不仅仅要研究、拓展、应用一些机器学习方法，还要通过许多非机器学习技术解决数据仓储、大规模数据、数据噪音等等更为实际的问题。
- 机器学习的涉及面更宽，常用在数据挖掘上的方法通常只是“从数据学习”，然则机器学习不仅仅可以用在数据挖掘上，一些机器学习的子领域甚至与数据挖掘关系不大，例如增强学习与自动控制等等。
- 数据挖掘试图从海量数据中找出有用的知识。
- 大体上看，数据挖掘可以视为机器学习和数据库的交叉，它主要利用**机器学习界提供的技术来分析海量数据，利用数据库界提供的技术来管理海量数据**。

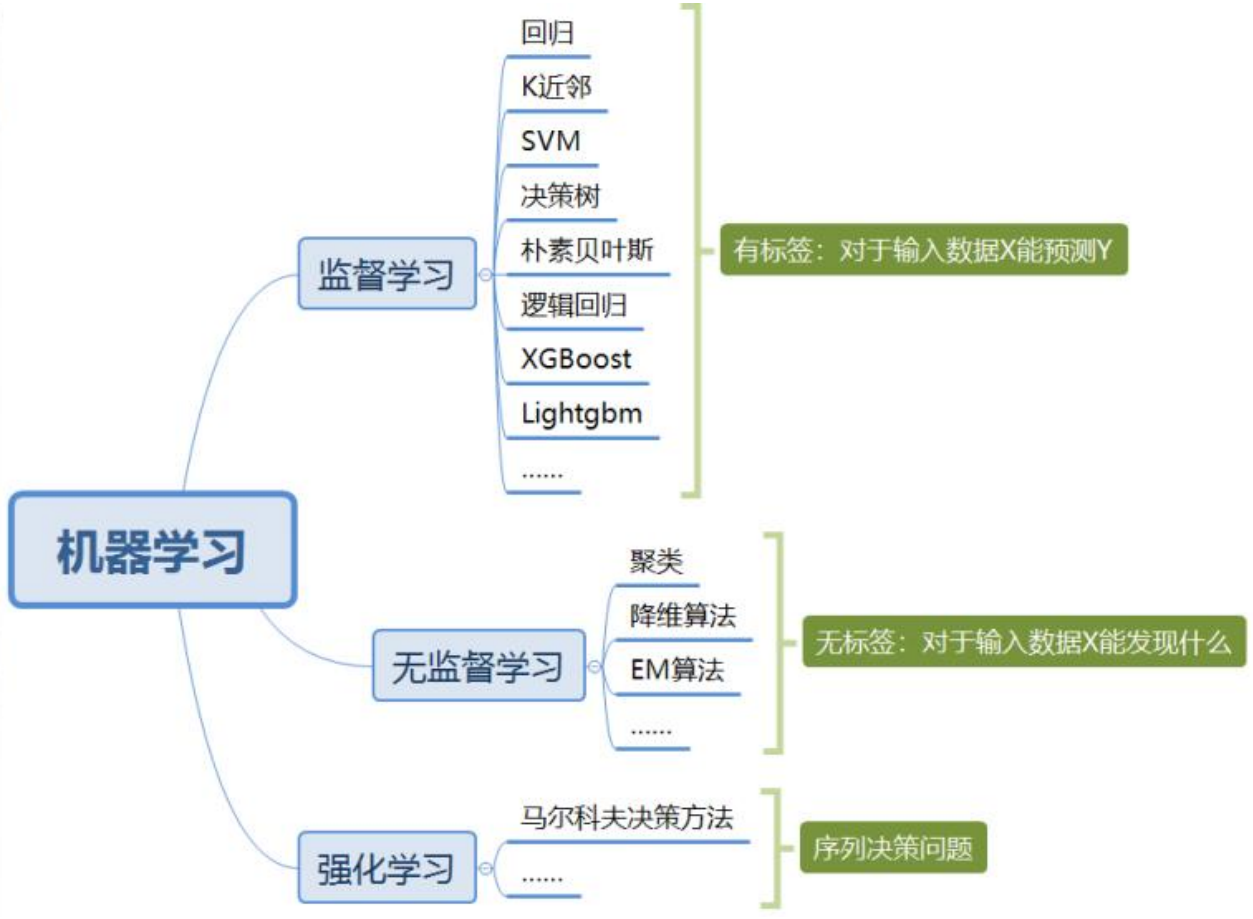
机器学习和数据挖掘区别与联系



机器学习类型

- **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
- **强化学习**
  - 强化学习（Reinforcement Learning, RL），又称再励学习、评价学习或增强学习，是机器学习的范式和方法论之一，用于描述和解决智能体（agent）在与环境的交互过程中通过学习策略以达成回报最大化或实现特定目标的问题

机器学习类型


机器学习
监督学习
回归
K近邻
SVM
决策树
朴素贝叶斯
逻辑回归
XGBoost
Lightgbm
.....
有标签：对于输入数据X能预测Y
无监督学习
聚类
降维算法
EM算法
.....
无标签：对于输入数据X能发现什么
强化学习
马尔科夫决策方法
.....
序列决策问题

机器学习的类型-监督学习

- 分类（Classification）
  - 根据肿瘤的体积、患者的年龄来判断良性或恶性？
  - 预测的标签是离散值
- 回归（Regression、Prediction）
  - 如何预测上海浦东的房价？
  - 预测的标签是连续值

机器学习的类型-监督学习

- 输入实例x的特征向量： $x=(x^{(1)},x^{(2)},\cdots,x^{(i)},\cdots,x^{(n)})^{\mathrm{T}}$
- $x^{(i)}$与x_i 不同,后者表示多个输入变量中的第i个

$$x_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)})^{\mathrm{T}}$$

- 训练集：

$$T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$$

- 输入变量和输出变量：
  - 分类问题、回归问题、标注问题

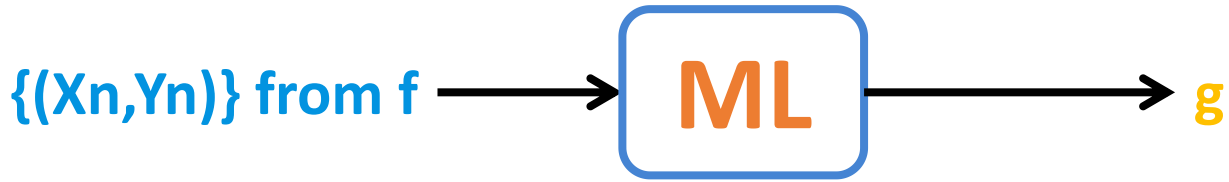
机器学习的类型-监督学习

- 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：$1.2 \quad 2.1 \quad 3.2 \quad 4.2 \quad 1.2 \quad 3.2$

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

- 标量/目标属性：$\begin{matrix} 1.2 & 2.1 & 3.2 & 4.5 & 1.2 & 1 \\ 0 & 0.1 & 向-0.2 & 量-3 & -1.2 & 0 \\ 23 & 21 & 20 & 15 & 19 & 1 \end{matrix}$

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)

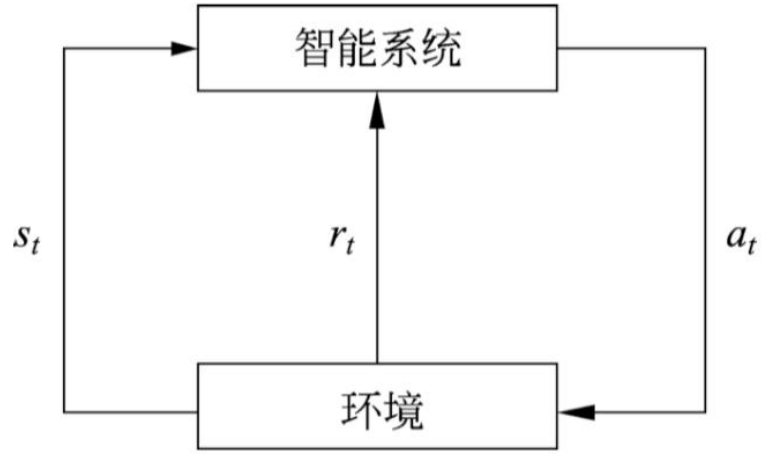
- 主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的问题。半监督学习对于减少标注代价，提高学习机器性能具有非常重大的实际意义。
- SSL的成立依赖于模型假设，主要分为三大类：平滑假设、聚类假设、流行假设；其中流行假设更具有普片性。
- SSL类型的算法主要分为四大类：半监督分类、半监督回归、半监督聚类、半监督降维。
- 缺点：抗干扰能力弱，仅适合于实验室环境，其现实意义还没有体现出来；未来的发展主要是聚焦于新模型假设的产生。

强化学习

强化学习的马尔可夫决策过程是状态、奖励、动作序列上的随机过程，由五元组 $\langle S, A, P, r, \gamma \rangle$ 组成。

- S 是有限状态（state）的集合
- A 是有限动作（action）的集合
- P 是状态转移概率（transition probability）函数:

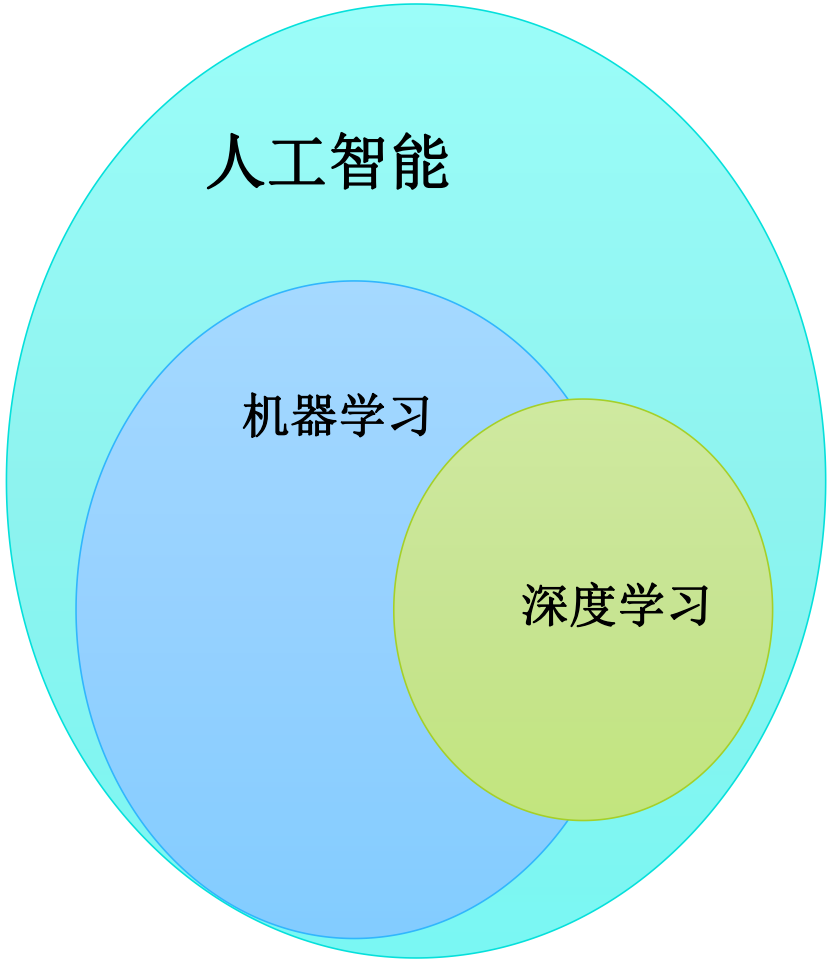
$$P(s'|s,a) = P(s_{t+1} = s'|s_t = s, a_t = a)$$



- r 是奖励函数（reward function）：$r(s,a) = E(r_{t+1}|s_t = s, a_t = a)$
- γ 是衰减系数（discount factor）：$\gamma \in [0,1]$

机器学习、人工智能和深度学习的关系

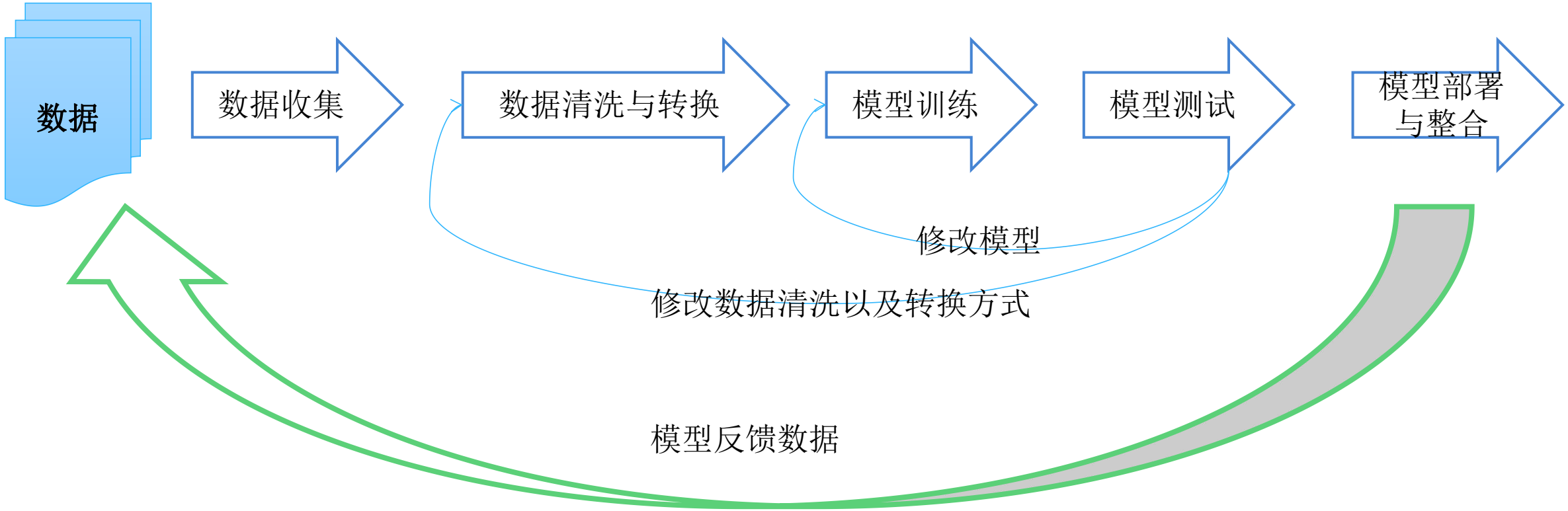
- 深度学习是机器学习的子类；深度学习是基于传统的神经网络算法发展到多隐层的一种算法体现。
- 机器学习是人工智能的一个子类；



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.php
  - 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。

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值为精准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+) = $\frac{\text{TPR}}{\text{FPR}}$	Diagnostic Odds Ratio (DOR) = $\frac{\text{LR+}}{\text{LR−}}$
	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−) = $\frac{\text{FNR}}{\text{TNR}}$	

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

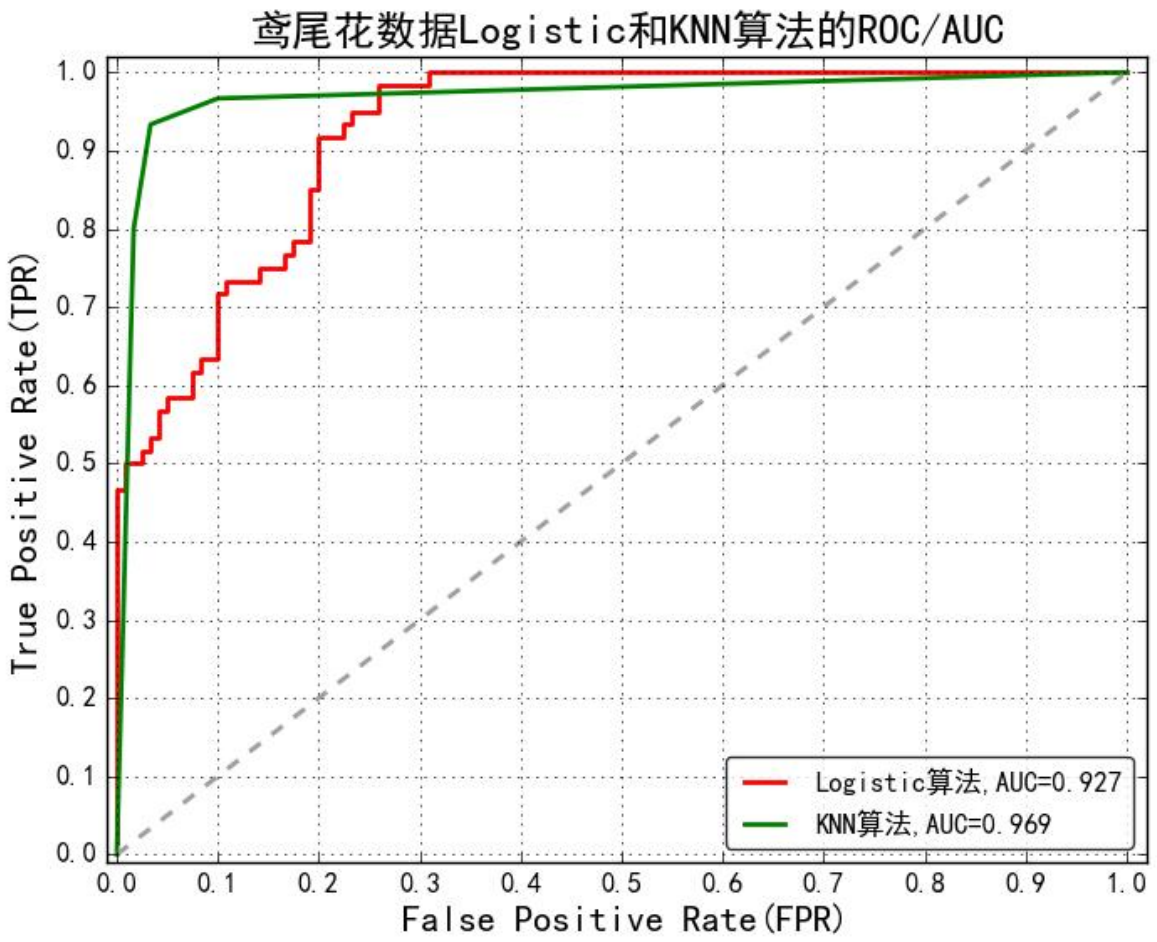
- 准确率 Accuracy
  - $\text{Accuracy} = \frac{\#(True\ positive)+\#(True\ negative)}{\#(True\ positive)+\#(True\ negative)+\#(False\ positive)+\#(False\ negative)}$
- 召回率 Recall
  - $\text{Recall} = \frac{\#(True\ positive)}{\#(True\ positive)+\#(False\ negative)}$
- 精确率 Precision
  - $Precision = \frac{\#(True\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

Thanks