—武大本科生课程

第1讲　机器学习与模式识别概论

(Lecture 1　Intro to Machine Learning and Pattern Recognition)

武汉大学计算机学院机器学习课程组

2023. 02

●课程考核说明

总评成绩=平时成绩×60%+期末笔试成绩×40%

(期末笔试成绩≥55分)

平时成绩包括上课考勤、平时作业、编程实验作业：

1. 上课考勤与平时作业

每讲课后布置(每布置3次左右打包一次文件提交)，占30%

2. 编程实验作业

(1) 编程单元实验作业：4次，随课程进展发布，占20%

(2) 华为编程大作业： 1次，使用华为AI平台，占10%

3. 期末笔试

开卷，占40%

●教材与主要参考书

[01] 埃塞姆•阿培丁，机器学习导论(第3版)，机工版

[02] 齐敏等编著，模式识别导论，清华版

简明扼要，逻辑结构较清晰。适合自学和阅读。

[03] 周志华，机器学习，清华版

相关概念讲解较多，有配套的公式推导南瓜书。

[04] 李航，统计学习方法(第2版)，清华版

公式推导较完整。

[05] 邱锡鹏，神经网络与深度学习，机工版

[06] Raschka，Python机器学习(第2版)，机工版

Python语言与ML工具包。本课程单元实验参考。

[07] 陈雷，深度学习与MindSpore实践，清华版(武大-华为合作推荐参考书)

本课程实验大作业参考。

[08] 田奇，ModelArts人工智能应用开发指南，清华版(武大-华为合作推荐参考书)

本课程实验大作业参考。

[09] 李弼程，模式识别原理及应用，西电版

[10] 孙即祥，现代模式识别 (第2版)，高教版

● 人工智能与机器学习相关主要CCF顶级国际期刊和顶级国际会议

· IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), IEEE
· Journal of Machine Learning Research (JMLR), MIT Press
· International Journal of Computer Vision (IJCV), Springer
· Artificial Intelligence (AI), ELSEVIER
· IEEE Transactions on Neural Networks
· Journal of AI Research
· Machine Learning
· Pattern Recognition
· Neural Networks
· Neural Computation
· IEEE Transactions on Fuzzy Systems
· IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)
· IEEE International Conference on Machine Learning (ICML)
· AAAI Conference on Artificial Intelligence AAAI(AAAI)
· International Conference on Computer Vision (ICCV)
· International Joint Conference on Artificial Intelligence (IJCI)
. Annual Meeting of the Association for Computational Linguistics (ACL)
· European Conference on Machine Learning (ECML)
· European Conference on Computer Vision (ECCV)
· European Conference on Artificial Intelligence (ECAI)
· Annual Conference on Neural Information Processing Systems (NIPS)

……(红色部分是CCF A类，蓝色部分是CCF B类)

●机器学习的定义

目前机器学习还没有一个准确、统一的定义。

经典定义：计算机程序如何随着经验的积累自动改善自身的性能 [T.Mitchell, CMU *Book* 97]。

Definition: A computer program is 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. (对于某类任务T和性能度量P，若一个计算机程序在T上以P衡量的性能随着经验E而自我完善，那么我们称该计算机在从经验E中学习。)---从学术的角度定义

例：手写文字识别学习问题：

·任务T：识别和分类图像中的手写文字

·任务性能标准P：分类的正确率

·训练经验E：已知分类的手写文字数据库

“Machine Learning” as an Engineering Paradigm: Use data and examples, instead of expert knowledge, to automatically create systems that perform complex tasks. ---从工程的角度定义

只要是有数据的地方，就会对数据进行分析(当前热门的研究领域：大数据分析)，机器学习就无处不在(Machine Learning Everywhere)，随着该领域的发展，机器学习主要做智能数据分析。机器学习的子类----深度学习可用于大数据分析。

●模式识别的定义

Pattern recognition is the study of how machines can observe the environment, learn to distinguish patterns of interest from their background, and make sound and reasonable decisions about the categories of the patterns. (Anil K. Jain)

模式识别与机器学习的关系：模式识别≈机器学习。两者的主要区别在于前者是从工业界发展起来的概念，后者则主要源自计算机学科。在著名的《**Pattern Recognition and Machine Learning**》一书中，**Christopher M. Bishop**在开头是这样说的：模式识别源自工业界，而机器学习来自于计算机学科。不过，它们中的活动可以被视为同一个领域的两个方面，同时在过去十几年，它们都取得了长足的发展。

通俗地说，**机器学习方法**是计算机利用已有的数据/经验，通过对数据/经验的学习得出某种模型(规律/机器学习算法)，并利用此模型预测未来(推断/决策)的一种方法。

机器学习与相关学科的关系：

(1) 模式识别：模式识别≈机器学习。两者的主要区别在于前者是从工业界发展起来的概念，后者则主要源自计算机学科。

(2) 统计学习：统计学习≈机器学习。统计学习是一个与机器学习高度重叠的学科，因为机器学习中的大多数方法来自统计学。统计学习研究者重点关注的是统计模型的发展与优化，偏数学；机器学习研究者重点研究“机器学习算法”在计算机上执行的效率与准确性的提升。

(3) 数据挖掘：数据挖掘=机器学习+数据库。大部分数据挖掘中的算法是机器学习算法在数据库中的优化。

(4) 计算机视觉: 计算机视觉=图像处理+机器学习。如百度识图、指纹识别、人脸识别等。

(5) 语音识别：语音识别=语音处理+机器学习。语音识别技术一般不会单独使用，一般会结合自然语言处理的相关技术，目前的相关成功应用有苹果语音助手Siri等。

(6) 自然语言处理：自然语言处理=文本处理+机器学习。自然语言处理技术主要是让机器(即计算机)理解人类语言的一门学科领域。在自然语言处理技术中，大量使用了计算机编译原理相关的技术，例如词法分析、语法分析等，除此之外，在理解层面则使用了语义理解、机器学习等技术。

图　人工智能、机器学习、深度学习三者关系

人工智能: Artificial Intelligence, 简称AI

机器学习: Machine Learining, 简称ML

深度学习: Deep Learinig, 简称DL

●机器学习技术与系统的种类

- 机器学习技术分类(I)： 是否有带分类标记的样本（样例）
  - 有监督学习、无监督学习、半监督学习、强化学习
- 机器学习技术分类(II)：是否能动态进行增量学习
  - 在线学习、批量学习
- 机器学习技术分类(III)：是简单地将待分类数据点和已知数据点进行匹配，还是像科学与工程中常见的那样，对训练数据进行模式检测，然后建立一个预测模型
  - 基于样例的学习、基于模型的学习
- 各分类之间并不互斥，可以按合适的方式进行组合。例如，现在比较先进成熟的垃圾邮件过滤器，可能是使用深度神经网络模型对垃圾邮件和常规邮件进行分类训练，完成动态学习。这使其成为一个在线的、基于模型的监督式学习系统

第1种分类：有监督学习、无监督学习、半监督学习、强化学习

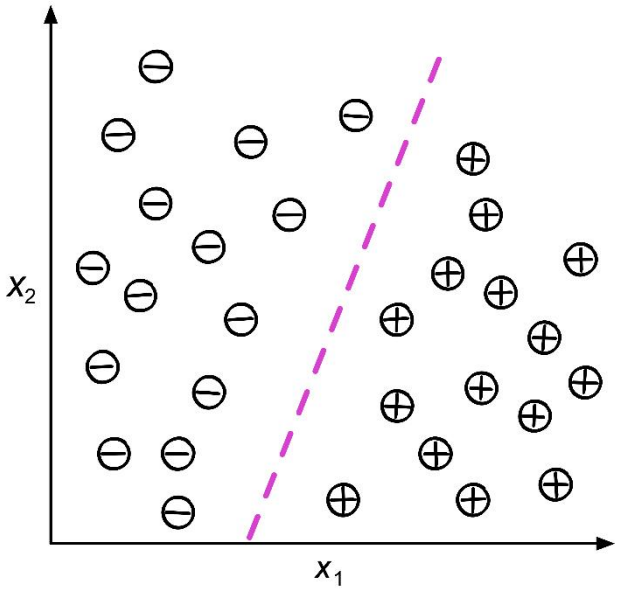
A. 有监督学习

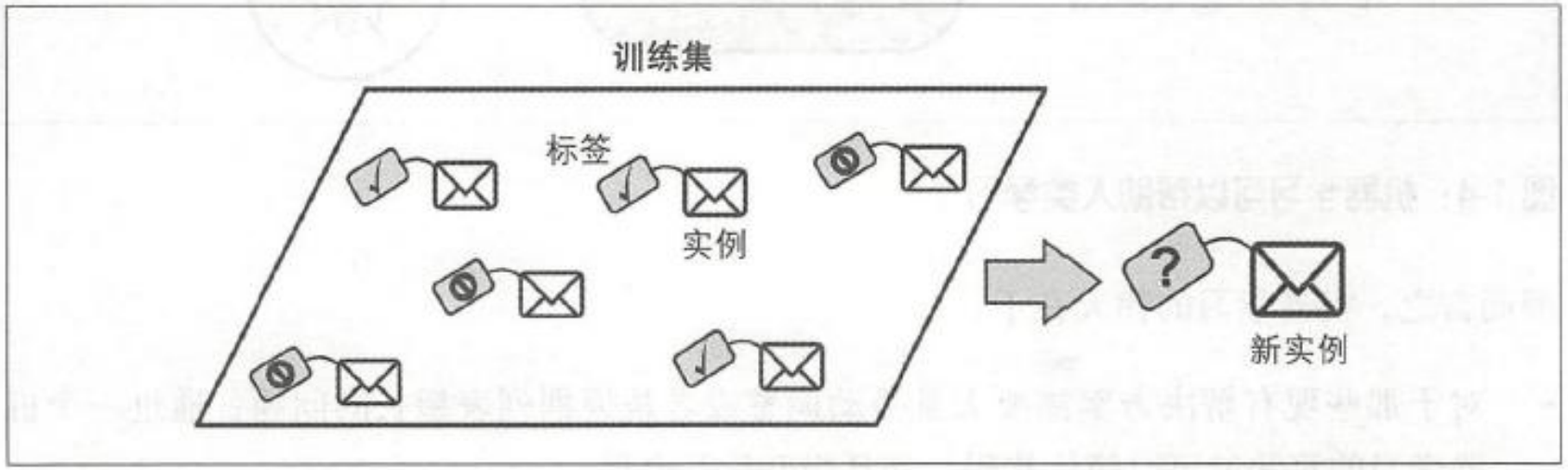
有指导者，也称为有导师的学习，常见的分类和回归问题都属于监督学习的范畴。有监督学习的目标是从已标记训练样本学习得到样本特征到样本标记的映射关系，这种映射关系要求与已标记样本情况相吻合。映射关系和标记在分类问题中分别指分类器和类别，而在回归分析(Regression Analysis)问题中就是回归函数和实值输出。

需要注意的是，在传统的监督学习中，通常都假设具有足够的已标记样本。如果已标记样本相对于维数或者标记数过少，那么，从中学习得到的映射会缺乏足够的泛化性，即对新样本进行判别分析的能力不足。

有监督学习：分类（类别预测）

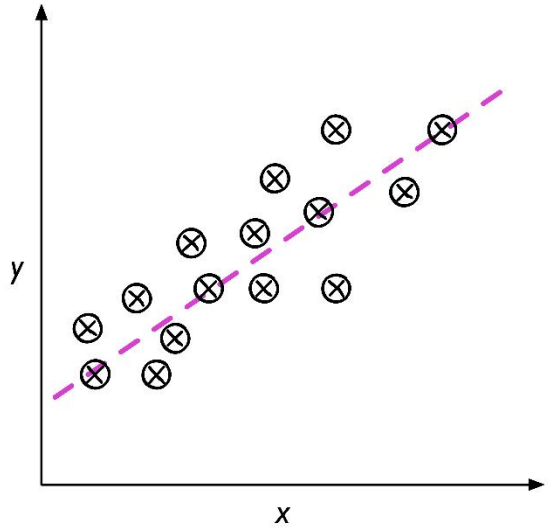
分类任务

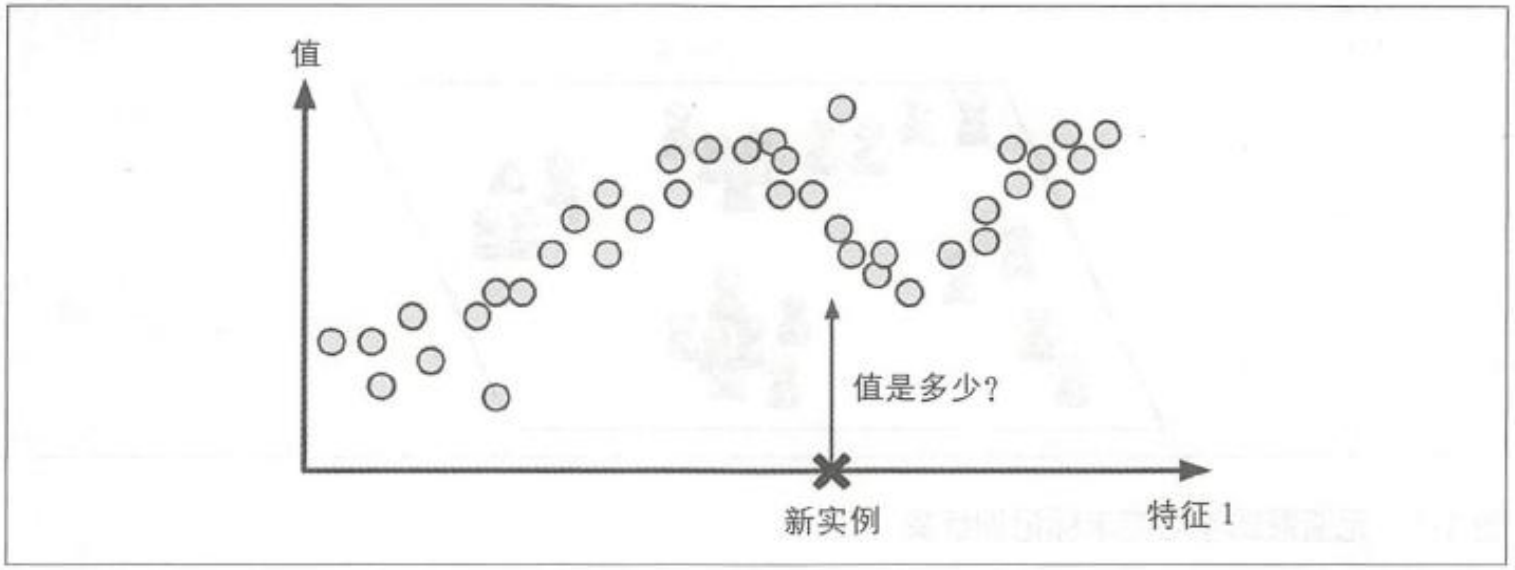




有监督学习：回归（数值预测）

回归任务





一些典型的有监督学习算法

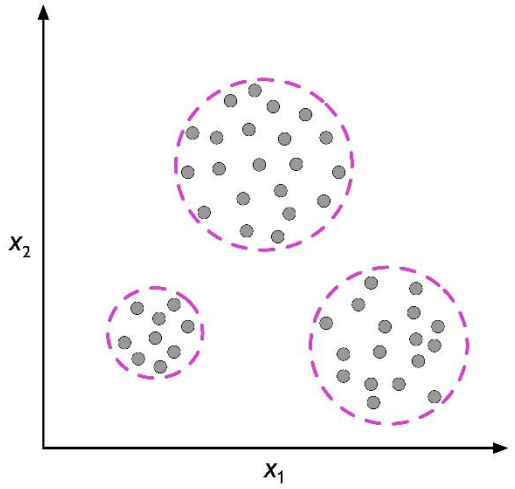
- K-近邻算法 (K-Nearest Neighbors)
- 线性回归 (Linear Regression)
- 逻辑回归 (Logistic Regression)
- 支持向量机 (Support Vector Machines, SVM)
- 决策树和随机森林 (Decision Trees & Random Forests)
- 神经网络 (Neural Networks)

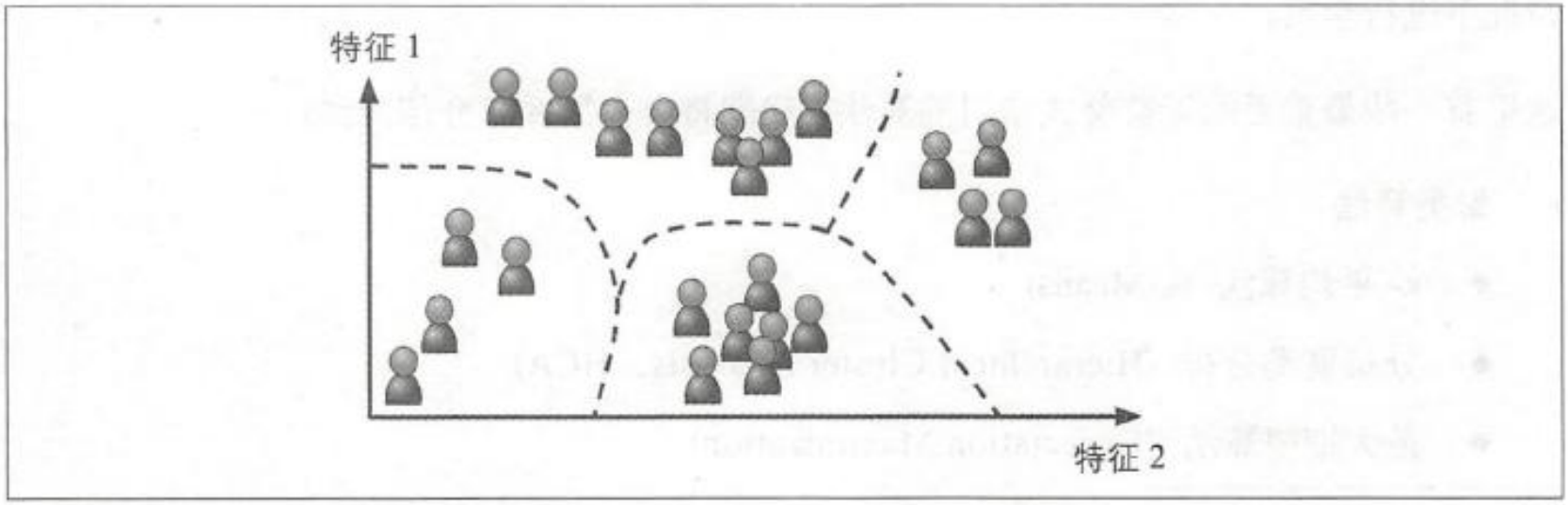
B. 无监督学习

无指导者，只有一批输入数据。其学习目标是发现输入数据集中潜藏的结构或者规律(这与概率统计中密度函数估计类似)。常见的无监督学习有聚类(Clustering)分析、降维等。

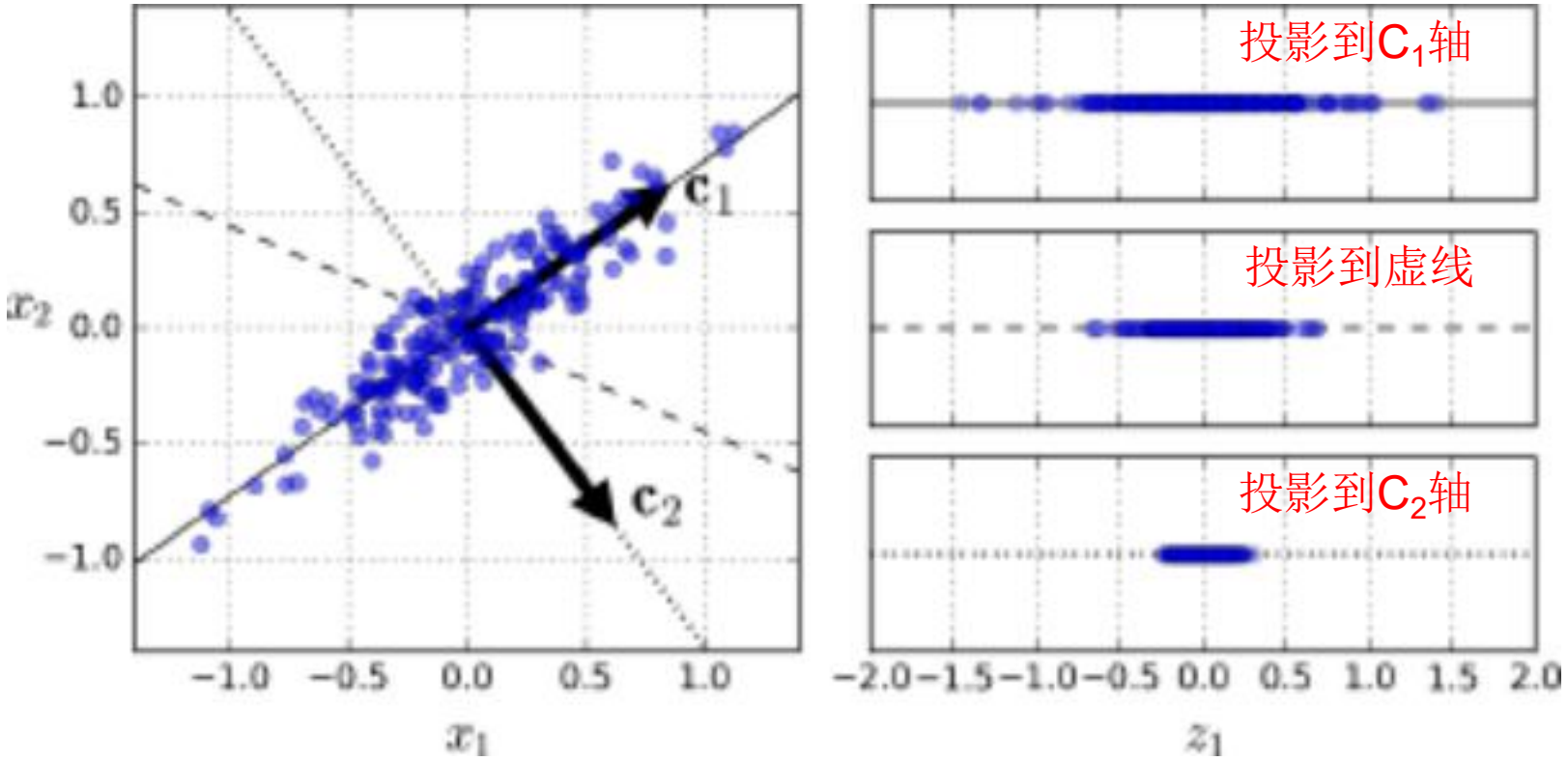
无监督学习：聚类分析

聚类分析



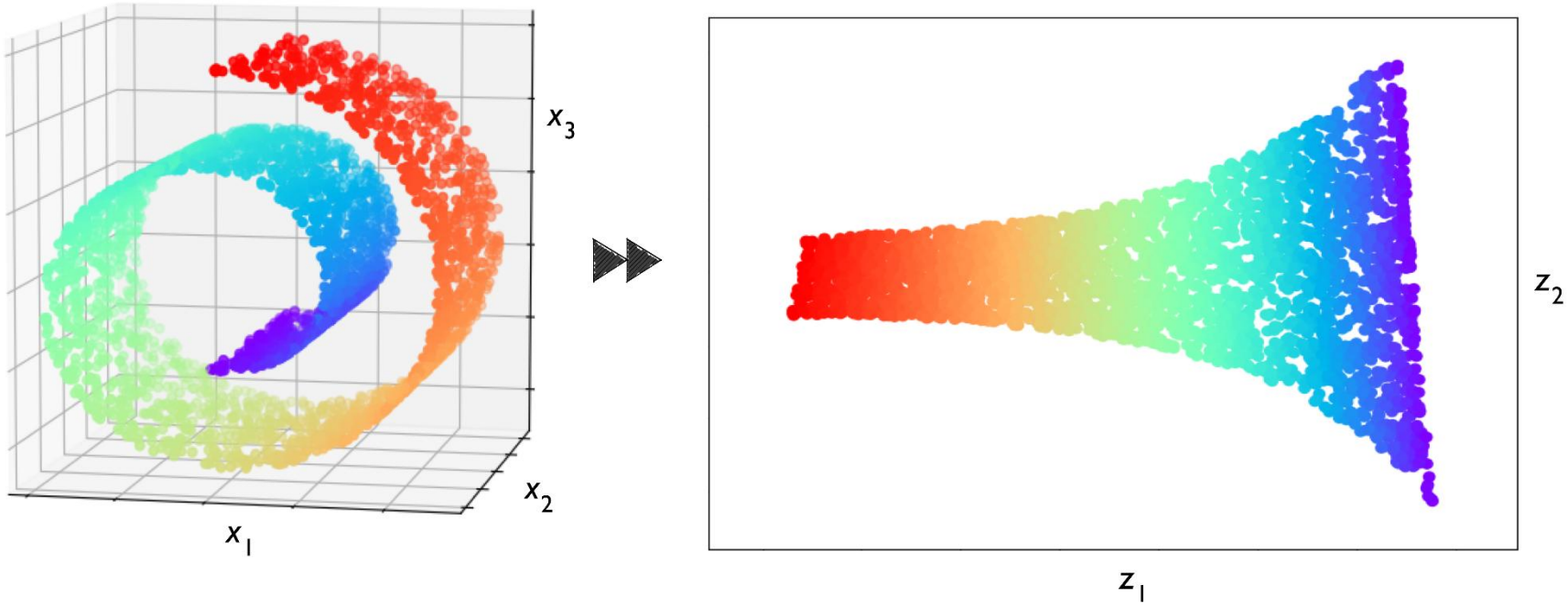


无监督学习：线性降维(PCA)



PCA是迄今为止最流行的降维算法
先找出第1主轴，投影数据的方差最大
再找出第2主轴：投影数据的方差次之
（第2主轴应该与第1主轴“正交”）

无监督学习：非线性降维(局部线性嵌入，LLE)

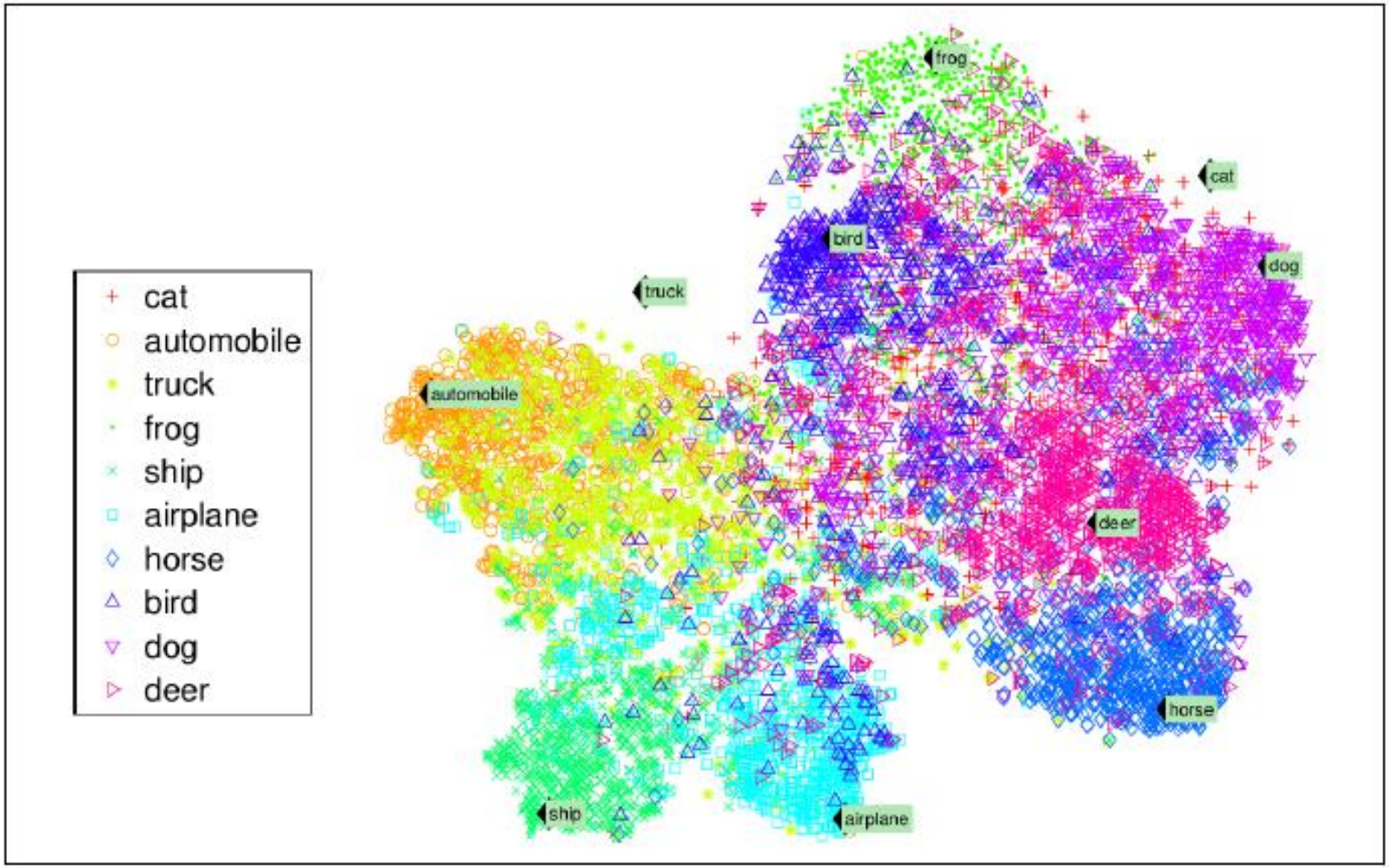


局部线性嵌入：一种流形学习技术

对高维空间的流形，先计算每个点与其邻域点的距离信息，在向低维空间投影时，最大程度保持这种局部距离信息。

LLE能有效实现，前提是高维空间的数据是本质低维的

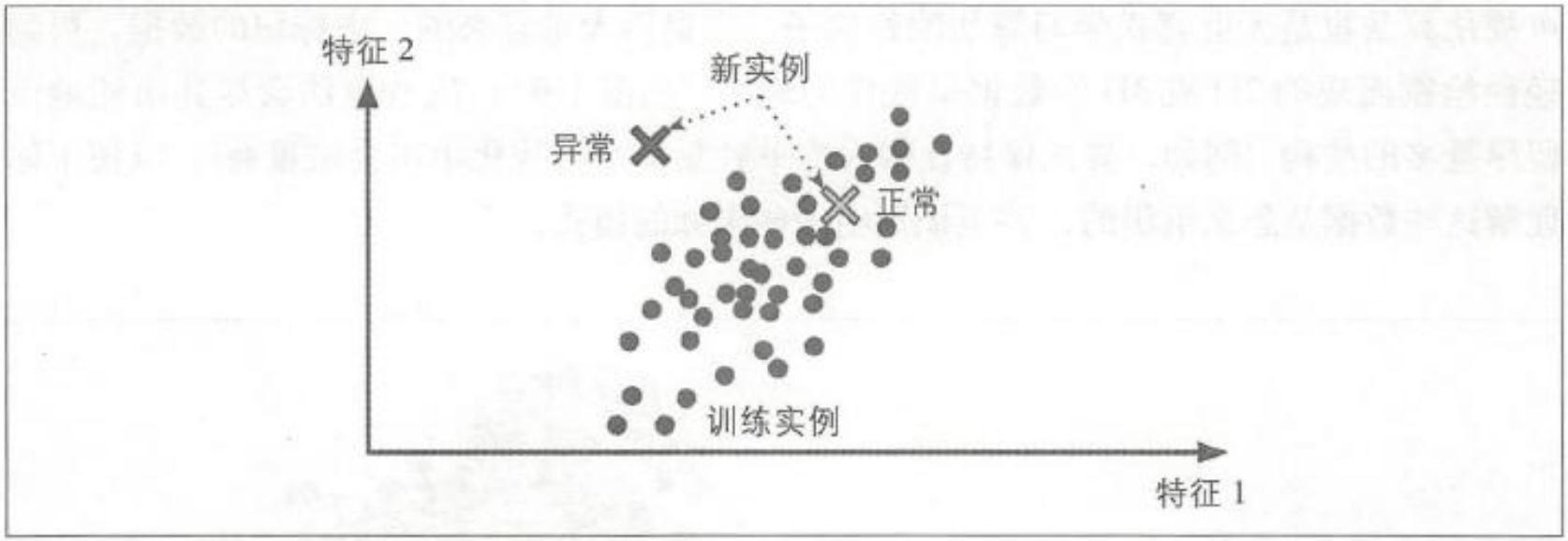
无监督学习：探索性数据分析(可视化)



一个可视化示例，突显了各种语义簇

无监督学习：异常检测

异常检测



一些典型的无监督学习算法

- **聚类算法**
  - K-均值聚类(K-Means)
  - 层次聚类分析(Hierarchical Cluster Analysis)
  - 概率聚类分析(Probabilistic Cluster Analysis)

- **降维算法**
  - 主成分分析(PCA)
  - 核主成分分析(K-PCA)
  - 局部线性嵌入(LLE)　　其他：IsoMap，MDS

- 关联规则学习算法
  - Apriori
  - Eclat

C. 强化学习/增强学习

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) 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。

强化学习的常见模型是标准的马尔可夫决策过程（Markov Decision Process, MDP）。按给定条件，强化学习可分为基于模式的强化学习（model-based RL）和无模式强化学习（model-free RL，以及主动强化学习（active RL）和被动强化学习（passive RL）。

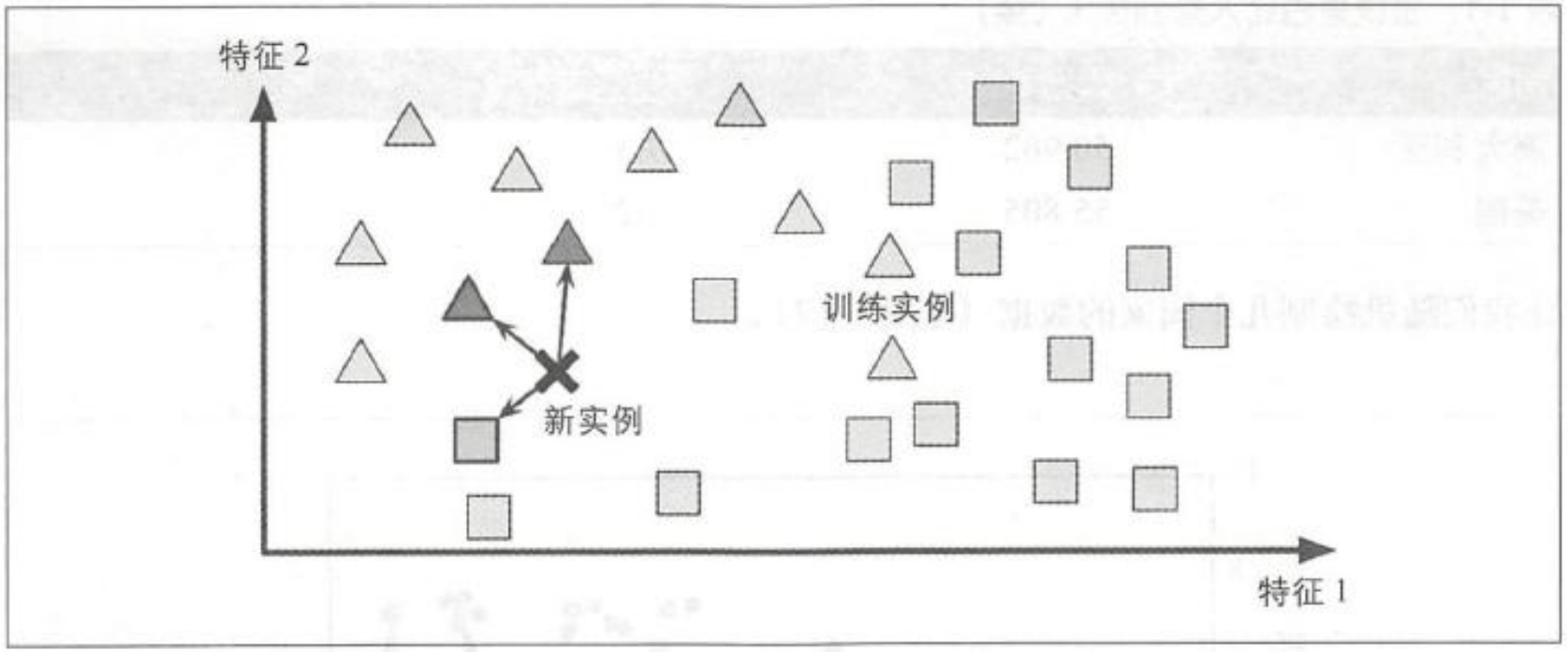
强化学习并不是某一种特定的算法，而是一类算法的统称。

D. 半监督学习

半监督学习(Semi-Supervised Learning，SSL)是机器学习和模式识别领域研究的重点问题，是监督学习与无监督学习相结合的一种学习方法。半监督学习的基本原则是通过大量无标记数据辅助少量已标记数据进行学习，从而提高学习效果。

第2种分类：基于样例的学习(*)

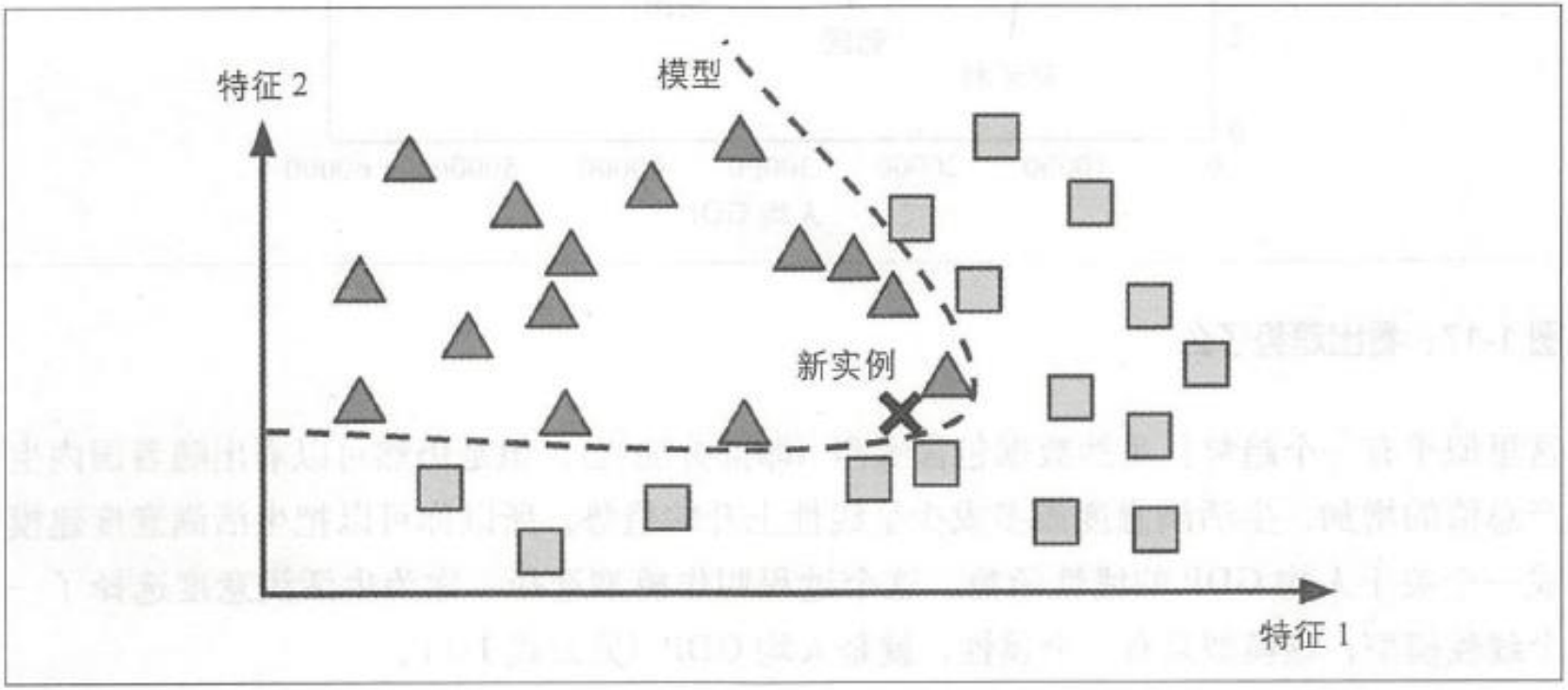
- 系统先完全记住学习样例（example），然后通过某种相似度量方式，将其泛化到新的实例（instance）　　（对新实例进行分类或回归）



基于样例的学习

第2种分类：基于模型的学习

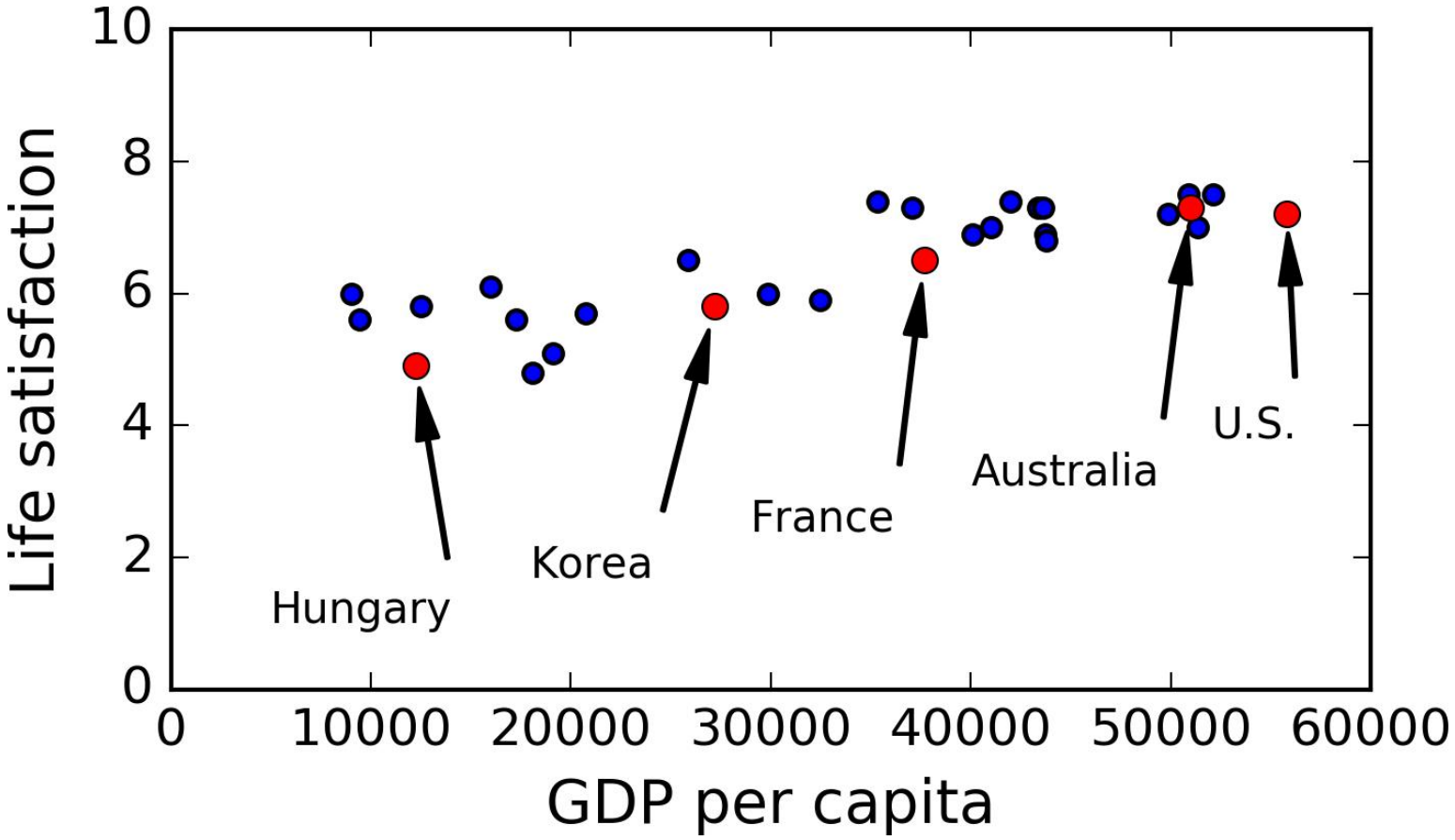
- 从样例中构建出或者能生成这些样例数据的模型（**生成式模型**），或者能基本反映这些样例中存在的某些数量关系（特征与标签）的模型（**判别式模型**），然后使用这些模型进行预测（数值回归和标签分类）



基于模型的学习

基于模型的学习：示例

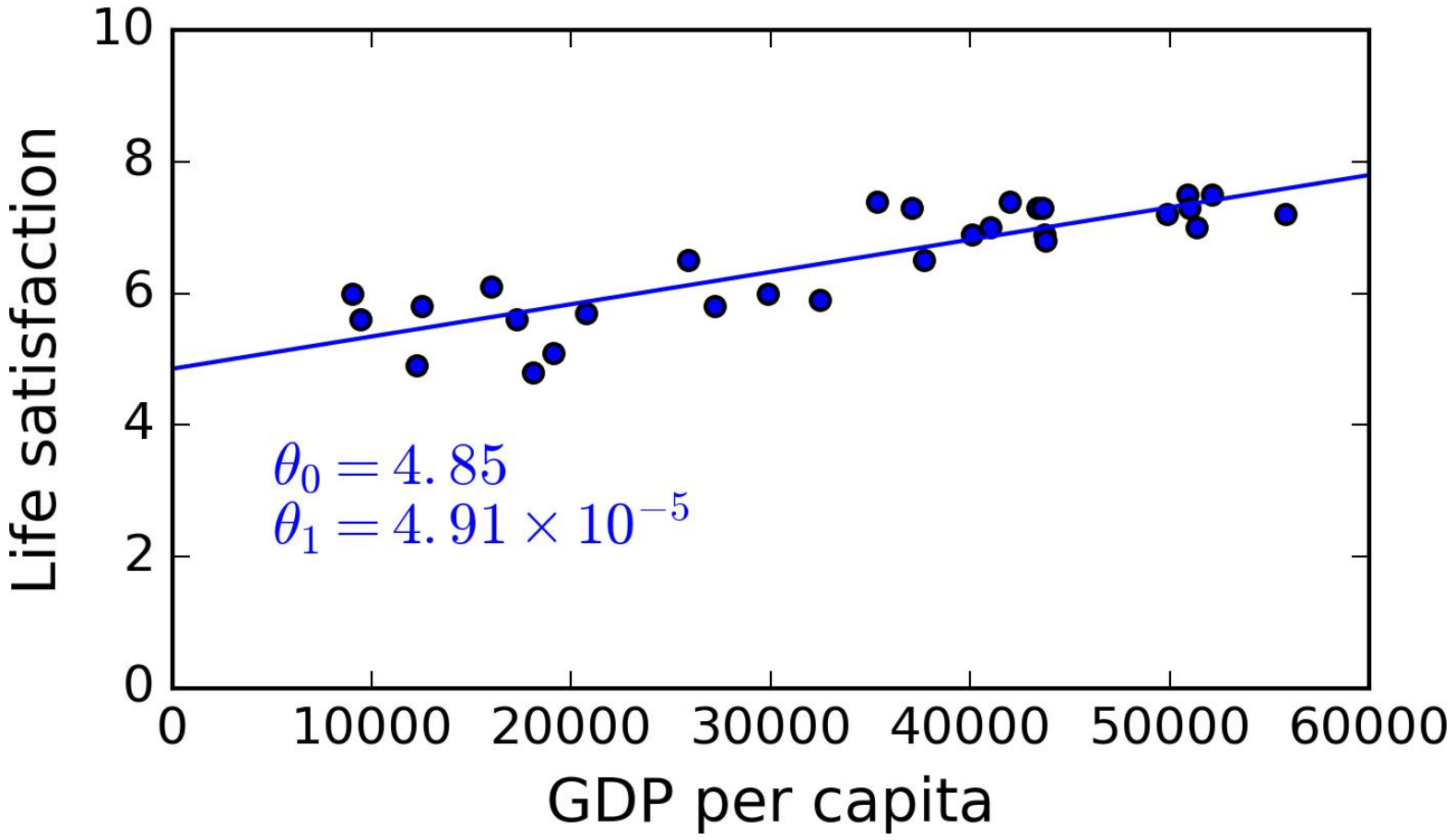
- 基于模型的回归问题：金钱能否使人感到快乐？



部分国家国民生活满意度与人均GDP的散点图

基于模型的学习：示例

- 简单线性回归模型：金钱似乎能使人感到快乐！



国民生活满意度与人均GDP的“关系”

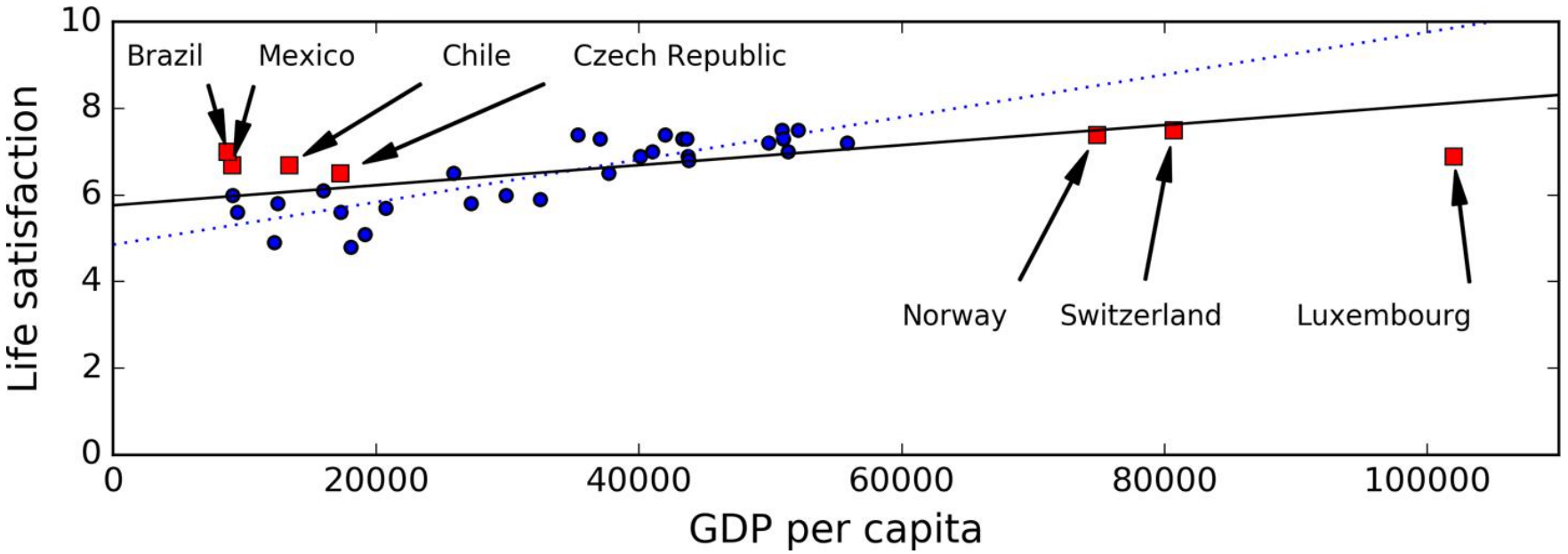
机器学习的主要挑战 1：训练数据的数量不足

- 由于构建机器学习技术的主要任务是选择一种学习算法，并对某些数据进行训练，所以可能出现的问题不外乎是“**坏算法**”和“**坏数据**”。让我们先从坏数据开始
- 训练数据的数量不足
  - 要教一个咿呀学语的小朋友什么是苹果，你只需要指着苹果说“苹果”（重复这个过程几次）就行了，然后孩子就能够识别各种颜色和形状的苹果了，样例教学的典型。简直是天才！
  - 机器学习还没达到这一步，大部分机器学习算法需要大量的数据才能正常工作。即使是最简单的问题，可能也需要成千上万个样例，而对于诸如图像识别或语音识别等复杂问题，机器可能需要上千万个样例（除非你可以重用现有模型的某些部分）

机器学习的主要挑战 2：训练数据不具代表性

- 训练数据不具代表性
  - 为了很好地实现向新数据的推广与泛化，重要的是，对于将要泛化的新实例来说， 建模用的训练数据一定要非常有代表性。不论你使用的是基于样例的学习还是基于模型的学习，都是如此
  - 例如，前面我们用来训练线性模型的国家数据集并不具备完全的代表性，有部分国家的数据缺失，下页图中显示了补充缺失国家 / 地区信息之后的数据表现
  - 如果用这个新的数据集训练线性模型，将会得到图中的实线，而虚线表示旧模型。添加部分缺失数据不仅显著地改变模型，也清楚地说明，这种线性模型可能永远不会那么准确
  - 看起来，某些非常富裕的国家并不比中等富裕国家幸福（事实上，看起来甚至是不幸福）。反之， 一些贫穷的国家也似乎比许多富裕国家更加快乐

机器学习的主要挑战 2：训练数据不具代表性



使用更具代表性数据后的线性回归模型

- 使用不具代表性的训练集拟合出的模型难以做出准确预测
- 使用具有代表性的训练集，说起来容易做起来难：
  - 如果样本集太小，将会出现采样噪声（非代表性数据被选中），容易导致模型过拟合
  - 即便是有大样本量数据，如果采样方式欠妥，同样可能导致非代表性数据集， 这就是所谓的采样偏差问题

机器学习的主要挑战 3：训练数据的质量差

- 如果训练集充满了错误、异常值和噪声（例如，差质量的测量产生的数据），系统将更难检测到底层模式，更不太可能表现良好。所以花时间来清理训练数据是非常值得的投入
- 事实上，大多数数据科学家都会花费很大一部分时间来做这项工作。如果样本集太小，将会出现采样噪声（非代表性数据被选中）
  - 数据异常：如果某些实例明显是异常情况，要么直接将其丢弃，要么尝试手动修复错误
  - 数据缺失：如果某些实例缺少部分特征（例如， 5% 的顾客没有指定年龄），你必须决定是整体忽略这些特征，还是忽略这部分有缺失的实例，又或者是将缺失的值补充完整（例如，填写年龄值的中位数）

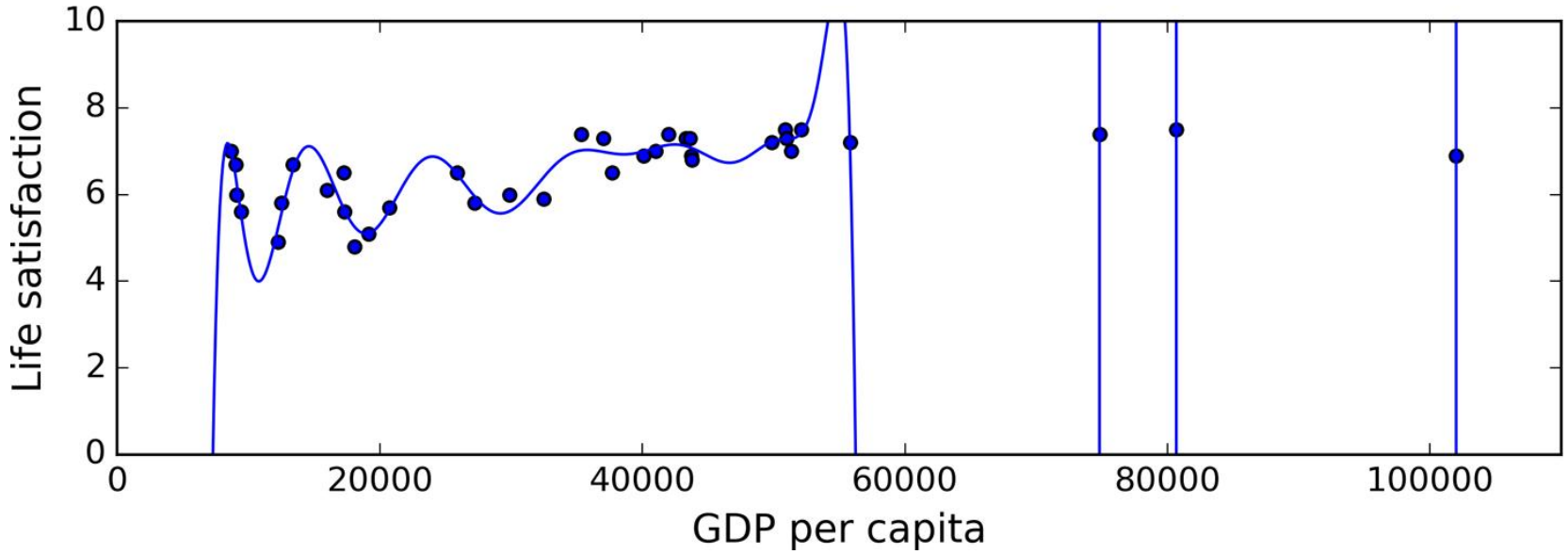
机器学习的主要挑战 4：无关特征，互相关特征

Garbage in, garbage out

- 当训练数据中包含较多**与目标变量有关的**（Relevant)特征，以及较少的**与目标变量无关的**（Irrelevant）特征，系统才能够有效学习。同时，**各特征之间要尽可能减少互相关**(Correlation)
- **互相关特征**：是指样本中的某些特征可以由其他特征，或者部分地，或者完全地表达出来
- **特征的多重共线性**：是互相关特征的极端情况，指某些特征可表达为其他特征的线性组合，导致算法极不稳定
- 一个机器学习项目成功的关键是：提取出一组好的特征子集，用于训练模型，这个工作叫**特征工程**，主要有以下几点：
  - 特征理解与选择：从现有特征中选择出最有效的特征
  - 特征清洗与增强：删除和填充含缺失值的数据，最大化数据价值
  - 持征提取与转化：将现有特征进行整合，产生更为有效的特征
  - 特征构建与学习：通过新数据创造新特征，或学习出新特征

机器学习的主要挑战 5：拟合数据过度（过拟合）

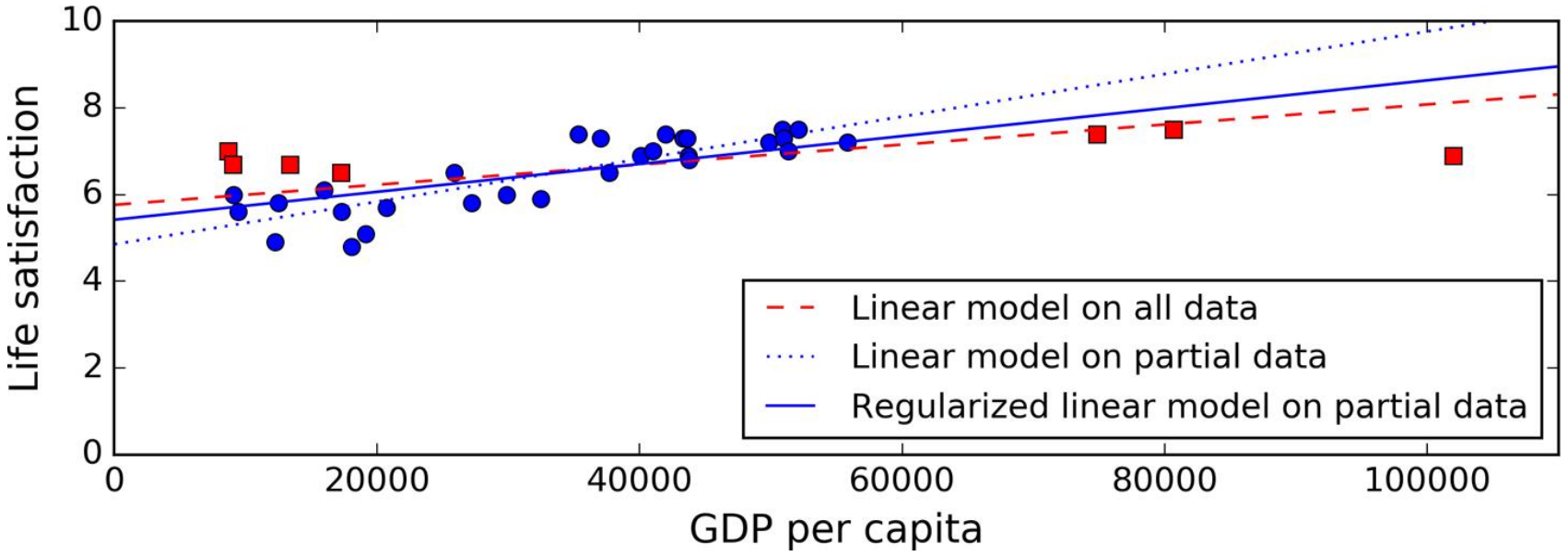
- 假设你正在国外旅游，被出租车司机狠宰了一刀，你很可能会说，那个国家的所有出租车司机都是强盗。**过度概括**是我们人类常做犯的错误。
- 图中是对前面的的生活满意度数据，用**高阶多项式**进行的模型拟合，这是一个明显的**与训练数据过拟合的**模型。虽然它在训练数据上的表现比简单的线性模型要好得多，但是**你真的相信它的预测是合理的吗**？



模型过拟合

机器学习的主要挑战 5：拟合数据过度 （过拟合）

- 蓝色**虚线**：代表一开始的原始模型，也就是缺失部分国家的数据时拟合的模型
- 红色**虚线**：代表用所有国家数据拟合出的第二个模型：
- 蓝色实线：代表的模型与第一个模型使用的训练数据相同，但是应用了模型正则化。可以看出通过正则化使得模型具有较小的斜率，虽然这略微降低了模型与训练数据的匹配程度，但是能更好地泛化到新的实例



机器学习的主要挑战 5：拟合数据过度 （过拟合）

- **过拟合** 是指在建模过程中，模型完美适应于训练集数据，从而牺牲对新数据点的预测能力的一种倾向
- 当模型相对于训练数据过于复杂时，容易发生过拟合
- 过拟合问题可能的解决方案有
  - 减少模型参数（人工手动）
  - 优化模型特征（特征工程）
  - 约束模型参数（对复杂性进行惩罚的正则化）
  - 增加训练数据（增加过拟合难度）
  - 改善数据质量（修复数据错误，填补数据缺失，清除异常数据）

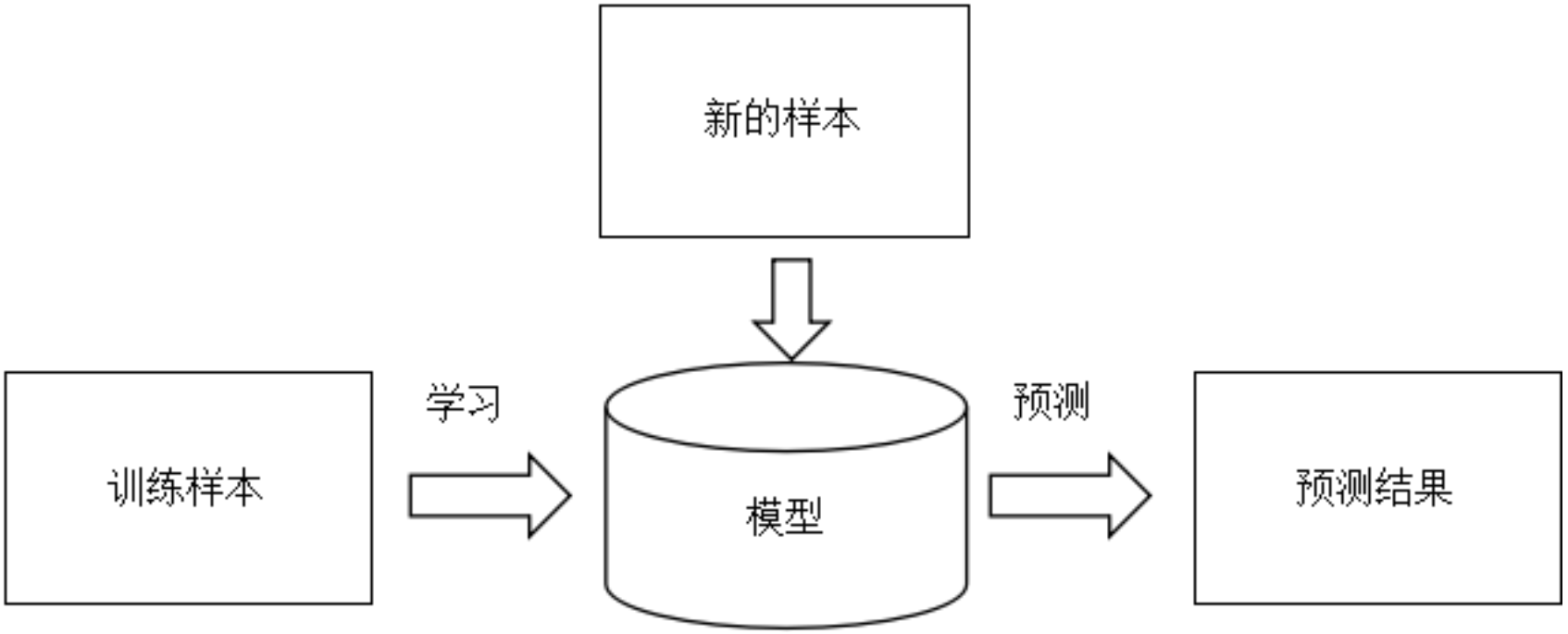
机器学习的主要挑战 6：数据拟合不足（欠拟合）

- **欠拟合** 与过拟合相反：模型太过简单，没有学习到数据中的真实结构，难以做出有效可用的预测
- 例：用线性模型来表达生活满意度与人均GDP的关系，属于拟合不足；**现实情况远比这个简单模型要复杂得多**。所以，即便对于用来训练模型的样例， 该模型产生的预测都会有很多是不准确的
- 解决欠拟合问题主要有以下一些方式：
  - 选择一个带有更多参数、更强大的模型（复杂化模型）
  - 通过特征变换，为学习算法提供更好的特征集（特征工程）
  - 放松对模型的约束（降低复杂度正则项的惩罚因子值）

回顾性小结

- 机器学习系统有很多类型：有监督的和无监督的， 批量的和在线的，基于样例的和基于模型的，等等
- 一个机器学习项目，包括从训练集中采集数据，然后将数据交给学习算法来计算等过程
  - 如果算法是**基于模型的**，它会调整一些参数来将模型适配于训练集（即对训练集本身做出很好的预测），然后算住就可以对新的场景做出合理的预测
  - 如果算法是**基于实例的**，它会记住这些样例，并根据相似度来对新的实例进行泛化
- 如果训练集的数据太少，数据代表性不够，包含太多噪声或者是被一些无关的特征或互相关特征污染，系统将不会很好地工作。最后，模型既不能太简单（这会导致欠拟合），也不能太复杂（这会导致过拟合）

机器学习任务的一般流程



机器学习任务的一般流程

●模式识别方法的分类

- **模式识别**：按理论基础划分
  - **统计**模式识别：以模式特征的概率分布为基础进行统计分析
  - **结构**模式识别：以形式语言理论对模式的结构特征进行分析
  - **模糊**模式识别：以模糊数学理论与方法进行的模式分类
  - **神经网络**模式识别：以人工神经网络理论与方法为基础
  - **人工智能**模式识别：以逻辑推理与专家系统理论为基础
- **统计模式识别**：按任务类型划分
  - **聚类分析**（Clustering Analysis）——简称：**聚类**
    - 简单聚类方法：最大最小距离法
    - 层次聚类方法：分裂式、凝聚式
    - 动态聚类方法：C-均值，ISODATA
  - **判别分析**（Discriminatory Analysis）——简称：**分类**
    - 几何分类法（判别函数分类法）：线性、分段线性、二次、支持向量机
    - 概率分类法（统计决策分类法）：判别式Discriminative、生成式Generative
    - 近邻分类法（几何分类法和概率分类法的一种融合方法）

模式识别简要历史

- 1950年代Noam Chemsky 提出形式语言理论，美籍华人 傅京荪提出句法结构模式识别。
- 1960年代L.A. Zadeh提出了模糊集理论，模糊模式识别理论得到了较广泛的应用。Fuzzy Set Theory。
- 1980年代Hopfield提出神经元网络模型理论。近些年人工神经网络ANN (Artificial Neural Networks)在模式识别和人工智能上得到较广泛的应用。
- 1990年代Vapnik等人提出的小样本学习理论，支持向量机SVM (Support Vector Machines)得到了很大的重视。

模式识别的几个学派（多明戈斯《终极算法》）

（1）**符号学派**，又称为逻辑主义、心理学派或计算机学派，其原理主要为物理符号系统假设和有限合理性原理。决策树方法，专家系统。

（2）**连接学派**，又称为**仿生学派**或生理学派，主要原理为神经网络及神经网络间的连接机制与学习算法。ANN。

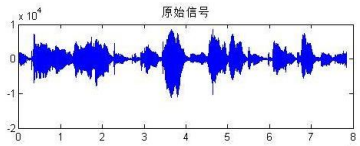
（3）**进化学派**。进化计算与遗传算法。

（4）**贝叶斯学派**：HMM，概率图模型。

（5）**类推学派**：kNN，支持向量机。

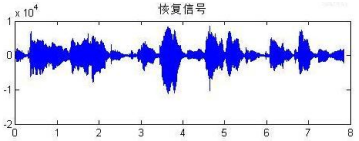
模式识别系统的组成

“处理”与“识别”两个概念的区别

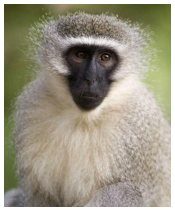


图像、语音

处理



图像、语音



动物、卫星照片

识别

具体的羊、猴

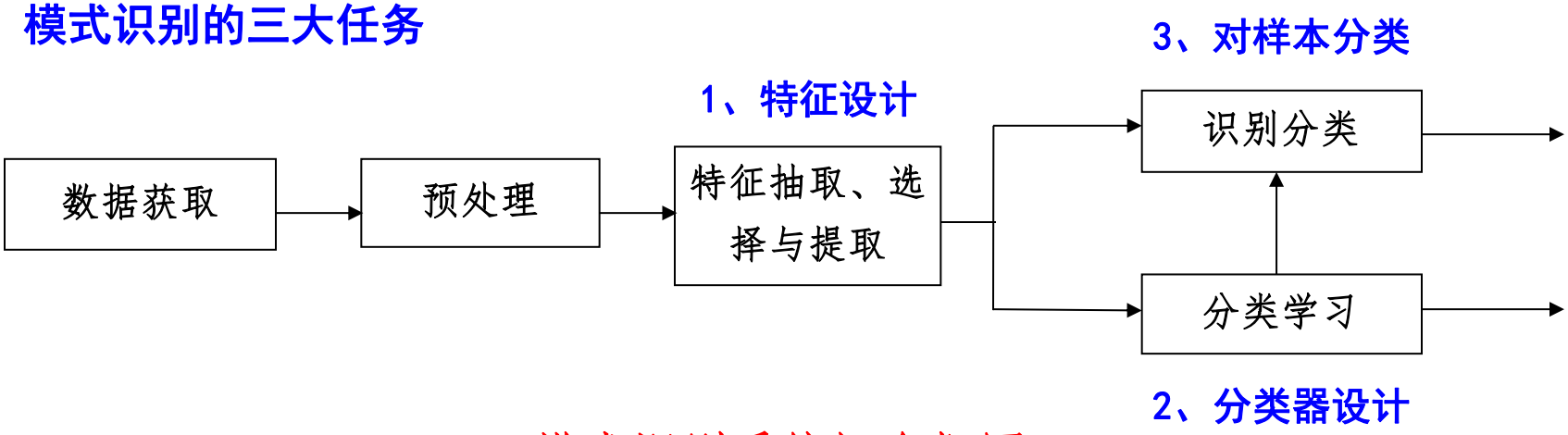
地形特点

处理：输入与输出是同样性质的数值或特征。

识别：输入的是对象特征，输出的是它的分类和描述。

模式识别系统的组成

模式识别的三大任务



模式识别系统概念框图

模式识别系统的主要环节：

- **特征抽取**：特征数据的采集。如长度测量、波形测量、…
- **特征选择**：选择尽可能少、且更有利于分类的特征。含特征变换（提取）
- **分类学习**：利用样本集建立分类和识别规则，即分类器设计
- **识别分类**：对所获得样本按建立的分类规则进行分类识别

●几个基本概念

- **样本**（sample）：一个具体的研究对象，具有一个或多个可观测量。
- **特征**（features）：能从某个方面对样本进行描述、刻画或表达的可观测量（数值型、结构型）。多个特征通常用矢量表示——特征矢量。
- **模式**（pattern）：样本特征向量的观测值，是抽象样本的数值代表。在这个意义上，“样本”与“模式”说的是同一件事情。
- **类别、类属、模式类**（class）：指在一定合理颗粒度下、有实际区分意义的基础上，主观或客观地被归属于“同一类” 的客观对象（样本、模式）的**类别代号**。数学上一般处理为整数。
- **样本集**（sample set）：多个样本的集合。训练集、验证集、测试集。
- **已知样本**（known samples，example）：事先知道类别标号的样本。
- **未知样本**（unknown samples）：指特征已知、但类别标号未知的样本，也称：**待分样本**、**待识样本**。
- 机器学习：称已知样本为**样例**(example)，未知样本为**实例**(instance)。
- **模式识别**（pattern recognition）：基于样本和应用目的设计分类器；并对性能进行验证，进行必要优化；最后用分类器对未知类属样本进行类属判定。**识别**，不是宽泛的认识或理解，而是对类别判定。

模式识别系统的组成

- **训练集**：样本的集合，用它来设计开发模式分类器。在有监督分类中，已知类别样本的集合，在无监督分类中，是未知类别样本的集合。但都被用来提取“关于各类之间的分类知识”，即使这些样本本身没有类别信息，仍然隐含着某种分类信息！
- **验证集**：既可以认为是宽泛的训练集中的一部分，也可以理解为独立于训练集的有监督样本集。在一次特定的模型训练过程中，两者不重合；但是在多次模型训练（伴随不同模型超参数设置），两者之间是可以交换部分数据的。其作用在于：通过对特定超参数或特定子模型的设置后，对训练后的模型进行“交付前测试”，以便从多个交付模型中选择泛化性能最优者。
- **测试集**：在设计与调优分类系统时没有用过的独立样本集。
- **系统评价原则**：为了更好地对模式识别系统性能进行评价，必须使用一组独立于训练集的测试集对系统进行测试

●模式识别的基本问题 (*)

1. 模式(样本)表示方法

(1)向量表示: 假设一个模式/样本有d个变量(特征)

$$x = (x_1, x_2, \ldots, x_d)^T$$

(2) 矩阵表示: N个样本，d个变量(特征)

变量 / 样本	x_1	x_2	…	x_d
$\mathbf{X}_1$	X_{11}	X_{12}	…	X_{1d}
$\mathbf{X}_2$	X_{21}	X_{22}	…	X_{2d}
…	…	…	…	…
$\mathbf{X}_N$	X_{N1}	X_{N2}	…	X_{Nd}

(3) 几何表示

1D表示

$X_1=0.5$　　$X_2=3$

2D表示

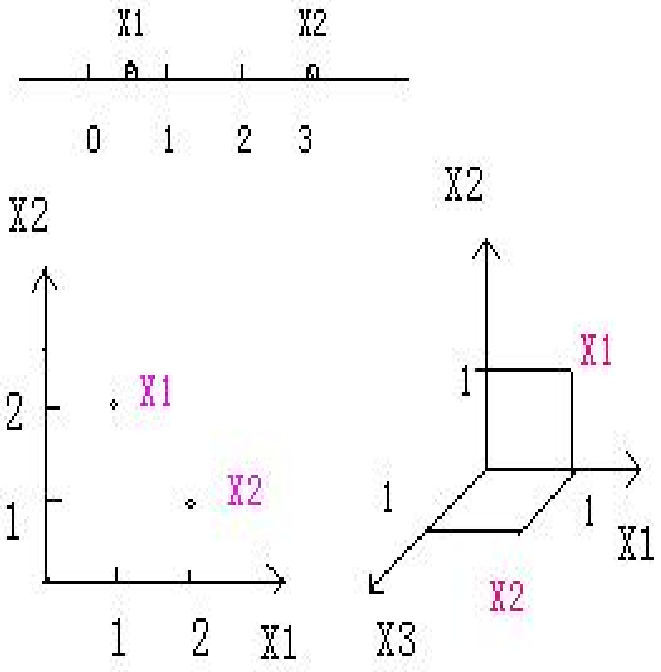
$X_1=(x_1, x_2)^T=(1, 2)^T$

$X_2=(x_1, x_2)^T=(2, 1)^T$

3D表示

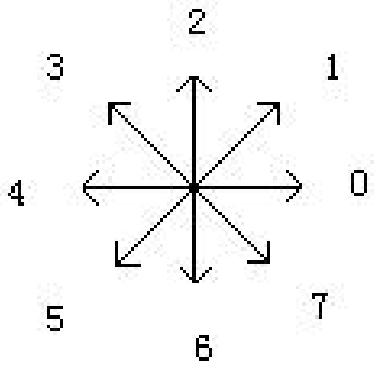
$X_1=(x_1, x_2, x_3)^T=(1, 1, 0)^T$

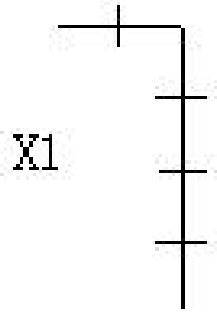
$X_2=(x_1, x_2, x_3)^T=(1, 0, 1)^T$



(4) 基元（链码）表示：

在右侧的图中八个基元分别表示0，1，2，3，4，5，6，7，八个方向和基元线段长度。



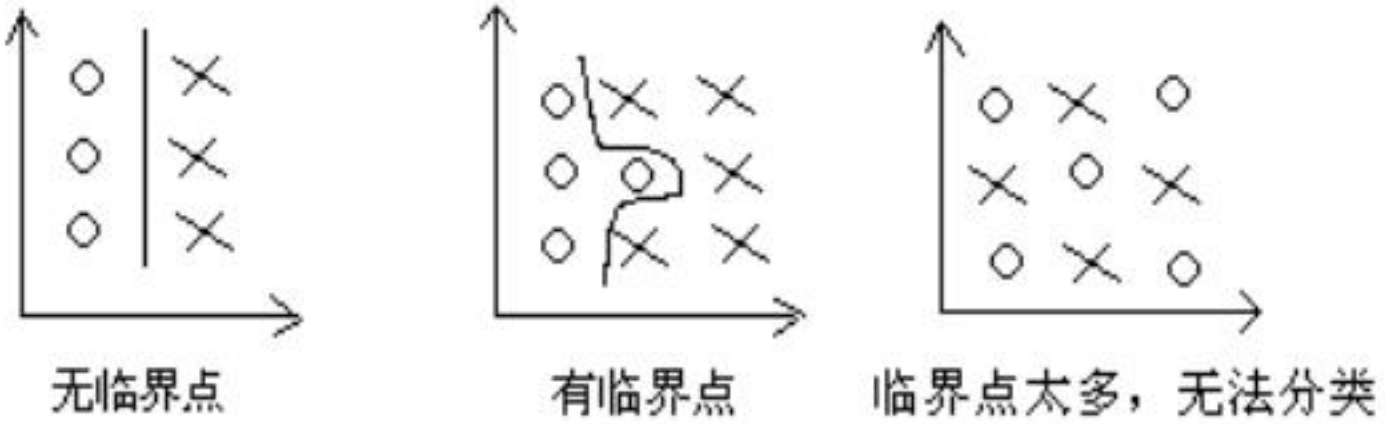


该方法在句法模式识别中会用到。

2. 模式类的紧致性

(1)紧致集

同一类模式类样本的分布比较集中，没有或临界样本很少，这样的模式类称紧致集(compact set)。



(2) 临界点(样本)：在多类样本中，某些样本的值有微小变化时就变成另一类样本称为临界样本(点)。

(3) 紧致集的性质

① 要求临界点很少

② 集合内的任意两点的连线，在线上的点属于同一集合

③ 集合内的每一个点都有足够大的邻域，在邻域内只包含同一集合的点

(4) 模式识别的要求：满足紧致集，才能很好的分类；如果不满足紧致集，就要采取变换的方法，以满足紧致集。

3. 相似与分类

(1)两个样本x_i，x_j之间的相似度量满足以下要求：

①应为非负值

②样本本身相似性度量应最大

③度量应满足对称性

④在满足紧致性的条件下，相似性应该是点间距离的单调函数

(2) 常用各种距离表示样本间的相似性：

①绝对值距离

已知两个样本，每个样本有n个特征：

$$x_i=(x_{i1}, x_{i2}, x_{i3},\ldots,x_{in})^T$$

$$x_j=(x_{j1}, x_{j2}, x_{j3},\ldots,x_{jn})^T$$

定义：$$d_{ij}=\sum_{k=1}^{n}|X_{ik}-X_{jk}|$$

②欧几里德距离

$$d_{ij}=\sqrt{\sum_{k=1}^{n}\left(X_{ik}-X_{jk}\right)^{2}}$$

③明考夫斯基距离

$$d_{ij}(q)=\left(\sum_{k=1}^{n}\left|X_{ik}-X_{jk}\right|^{q}\right)^{1/q}$$

其中当q=1时为绝对值距离，当q=2时为欧氏距离

④切比雪夫距离

$$d_{ij}(\infty)=\max_{1\le k\le n}|X_{ik}-X_{jk}|$$

明氏距离，q趋向无穷大时的极限情况

⑤马氏距离(Mahalanobis：马哈拉诺比斯)：样本X_i和X_j间的马氏距离为

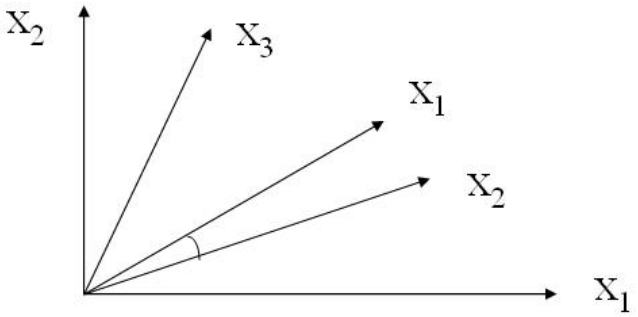
$$d_{ij}^2(M)=\left(X_i-X_j\right)^T\Sigma^{-1}(X_i-X_j)$$

其中X_i、X_j为特征向量集$\{X_1, X_2, \ldots, X_m\}$中的两个n维特征向量，$\Sigma$为模式总体样本的协方差矩阵。马氏距离的使用条件是样本符合正态分布。

⑥夹角余弦 $$C_{ij}=\cos\theta=\frac{X_i^T\cdot X_j}{|X_i|\cdot|X_j|}=\frac{\sum_{k=1}^{n}X_{ik}X_{jk}}{\sqrt{\left(\sum_{k=1}^{n}X_{ik}^{\ 2}\right)\left(\sum_{k=1}^{n}X_{jk}^{\ 2}\right)}}$$

即两样本间夹角小的为一类，具有相似性。

例： x_1 , x_2 , x_3的夹角如右图。



因为x_1 , x_2 的夹角小，所以x_1 , x_2 最相似。

⑦ 相关系数

$$r_{ij}=\frac{s_{ij}^2}{\sqrt{s_{ii}^2 s_{jj}^2}}=\frac{\sum_{k=1}^{n}\left(X_{ki}-\overline{X_i}\right)\left(X_{kj}-\overline{X_j}\right)}{\sqrt{\sum_{k=1}^{n}\left(X_{ki}-\overline{X_i}\right)^2\sum_{k=1}^{n}\left(X_{kj}-\overline{X_j}\right)^2}}$$

$\overline{X_i},\overline{X_j}$ 为x_i 、 x_j的均值

注意：在求相关系数之前，要将数据标准化。

3. 分类的主观性和客观性

(1)分类带有主观性：目的不同，分类不同。如：鲸鱼、牛、马从生物学的角度来讲都属于哺乳类，但是从产业角度来讲鲸鱼属于水产业，牛和马属于畜牧业。

(2)分类的客观性：科学性

判断分类必须有客观标准，因此分类是追求客观性的，但主观性也很难避免，这就是分类的复杂性。

4. 特征的生成

(1)低层特征：

①无序尺度：有明确的数量和数值。

②有序尺度：有先后、好坏的次序关系， 如酒分为上、中、下三个等级。

③名义尺度：无数量、无次序关系，如有红、黄两种颜色。

(2) 中层特征：经过计算、变换得到的特征

(3) 高层特征：在中层特征的基础上有目的的经过运算形成

例如：椅子的重量=体积*比重

体积与长、宽、高有关；比重与材料、纹理、颜色有关。这里低、中、高三层特征都具备了。

5. 数据的标准化

(1) 极差标准化

一批样本中，每个特征的最大值与最小值之差，称为极差。

极差/全距 $R_i = \max X_{ij} - \min X_{ij}$

极差标准化 $Xij = \left(X_{ij} - \overline{X_i}\right) \Big/ R_i$

(2) 方差标准化

$Xij = \left(X_{ij} - \overline{X_i}\right) \Big/ S_i$

S_i为样本方差

标准化的方法很多，原始数据是否应该标准化，应采用什么方法标准化，都要根据具体情况来定。

●48学时课程学习说明

第01讲：概论

第02讲：机器学习基本方法入门

第03讲：线性判别分析

第04、05讲：概率分类

第06、07讲：支持向量机

第08讲：决策树

第09讲：集成学习与随机森林

第10讲：模型评估与选择

第11讲：特征提取

第12讲：聚类分析

第13讲：神经网络

第14讲：华为AI平台及应用（ModelArts&MindSpore）

第15讲：深度学习（卷积神经网络CNN）

第16讲：华为MindSpore应用案例

（32~36少学时：建议不讲第05讲、第07讲、第09讲、第10讲、第15讲、第16讲）

理论学习

- 补足**矩阵计算、概率统计、最优化**相关知识；
- 牢固掌握**基本概念**、**重要原理**及其**算法步骤**；
- 运用重要算法对简单实例进行**手工演算求解**；
- 熟记重点算法与指标的**计算公式和推导过程**；
- 熟记重点概念与算法的**原理图**及其丰富含义；
- 能为后续科研工作打下必要的**专业理论基础**。
- 数学基础：高等数学（Tayler展开）、线性代数与矩阵计算，概率论与数理统计，数值计算（数据拟合，非线性最优化）

机器学习与模式式识别是一门综合性很强、但是对数学和统计学要求较高的技术学科，每一方面内容都涉及较多数学、统计与应用背景，因此不同基础的同学，学习难度也不同。

课堂学习主要进行概念讲解，原理与算法过程简介，想要切实掌握机器学习与模式识别方法，须在课后多读、多算、多实验，熟记概念、弄懂原理、熟悉算法，通过开源工具解决实际问题。

编程实验要求

- 编程环境：使用Python语言以及机器学习相关的软件工具。熟悉Python开发环境（Anaconda，Jupyter Notebook，JupyterLab，Spyder，etc.），结合示例代码，读懂用NumPy，Pandas，matplotlib，scikit-learn等写成的机器学习核心算法，了解主要函数的使用方式和参数含义
- 编程要求：熟悉面向对象编程概念，能看懂Python代码，能进行数据处理类编程，了解主要工具包（scikit-learn）的库、模块、类、函数的依赖关系，掌握华为MindSpore深度学习框架基本应用，以及机器学习算法参数的取值与调优
- 实验作业：课程实验分为两部分。第一部分主要以运行示例代码、对示例代码中部分处理过程进行参数调整或修改，观察并记录运行结果，结合所学知识写出实验分析报告。第二部分则属于综合性实验，能完成一个完整的模式识别或数据分析任务（具体在课程进行中安排）

Python+Numpy基本编程举例

例：设X=[1.70 1.75 1.65 1.80 1.78]，编程求和、均值、方差、标准差。

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$$

方差(s^2：修正样本方差)是描述数据取值分散性的一个度量，它是数据相对于均值的偏差平方的平均值。

采用Python编写脚本：

```
#Filename: python_statistics_1.ipynb
import numpy as np
x=[1.7,1.75,1.65,1.80,1.78] #height data
su=np.sum(x)
m=np.mean(x)
vr=np.var(x)  #Compute the variance
sd=np.std(x)  #Compute the standard deviation
print('su=',su,'\nm=',m,'\nvr=',vr,'\nsd=',sd)
```

以上*.ipynb格式Python脚本可在IPython notebook/ Anaconda Jupyter notebook等环境下运行。

Jupyter notebook环境下
程序运行情况：

localhost:8888/notebooks/python_statistics_1.ipynb

File Edit View Insert Cell Kernel Widgets Help

In [1]:

```
#Filename:python_statistics_1.ipynb
import numpy as np
x=[1.7,1.75,1.65,1.80,1.78] #height data
su=np.sum(x)
m=np.mean(x)
vr=np.var(x)   #Compute the variance
sd=np.std(x)   #Compute the standard deviation
print('su=',su,'\nm=',m,'\nvr=',vr,'\nsd=',sd)
```

```
su= 8.68
m= 1.736
vr= 0.002984
sd= 0.054626001135
```

数学基础提要

- 参见文件：第0讲 数学基础提要

●本讲小结

机器学习~模式识别

Definition: A computer program is 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. ---机器学习大牛Mitchell (CMU, Carnegie Mellon University, 美国卡内基梅隆大学)从学术的角度定义

翻译过来用大白话说就是：针对某件事情，计算机会从经验中学习，并且越做越好。从机器学习和模式识别领域先驱们的定义来看，我们可自己总结出对机器学习和模式识别的理解：

机器学习是一种程序，针对某个特定的任务，从经验中学习，并且越做越好。

针对这个理解，我们可以得出以下针对机器学习最重要的内容：

数据：经验最终要转换为计算机能理解的数据，这样计算机才能从经验中学习。谁掌握的数据量大、质量高，谁就占据了机器学习和人工智能领域最有利的资本。用人类来类比，数据就像我们的教育环境，一个人要变得聪明，一个很重要的方面就是要享受到优质的教育。所以，从这个意义上说，就能理解类似Google这种机器学习公司开发出来的机器学习程序性能为什么那么好了，因为他们能获得海量的数据。

模型：也就是机器学习或模式识别算法。有了数据后，可以设计一个模型，让数据作为输入来训练这个模型。经过训练的模型，最终就成了机器学习的核心，使得模型成为了产生决策的中枢。一个经过良好训练的模型，当输入一个新样本(新模式) 时，会做出适当的反应，产生优质的输出。

It is not knowledge, but the act of learning, which grants the greatest enjoyment.

Karl Friedrich Gauss, 1808

End of this lecture.

Thanks!