BIBDR 北京大数据研究院 | 博雅大数据学院 BOYA BIG DATA COLLEGE | 数据酷客 COOKDATA

北京市大数据人才培训示范基地

第一讲：机器学习介绍

欧高炎，北京大学博士、博士后

数据酷客创始人，博雅大数据学院院长

“机器学习十讲”直播安排

时间	主题	介绍
5/14	机器学习介绍	机器会学习吗？
5/21	回归	回归初心，方得始终
5/28	分类	分门别类，各得其所
6/4	模型提升	三个臭皮匠，顶个诸葛亮
6/11	聚类	物以类聚，人以群分
6/18	降维	取其精华，去其糟粕
6/23	最优化	摸着石头过河，蒙着眼睛爬山
7/2	维度灾难	来自维数的诅咒
7/9	深度学习	深层次学习的艺术
7/16	强化学习	让机器像人类一样学习

We define *machine learning* as a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty (such as planning how to collect more data!).

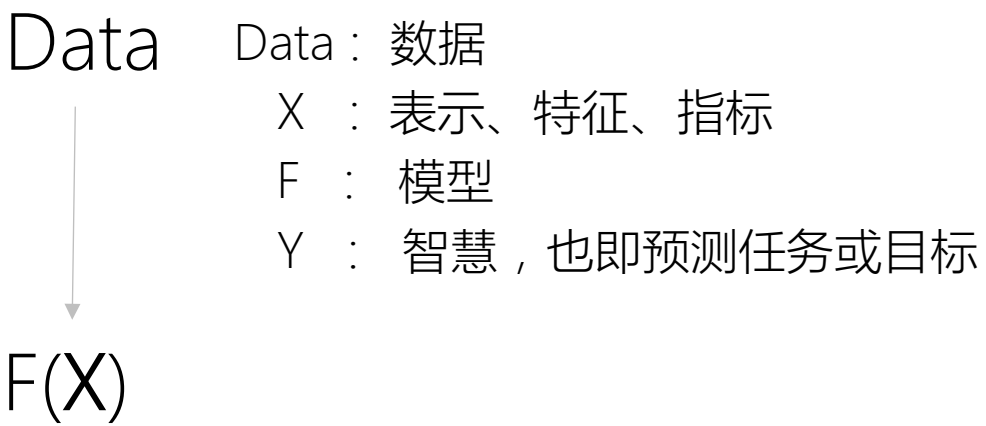
— 《Machine Learning: A probabilistic perspective》

为什么机器学习很重要？

大数据分析和人工智能已经成为整个社会发展最主要的基础推动力，两者的基础都是机器学习。

大数据分析火热的深刻原因：

- **数据源**：非结构化数据（语音、视频、文本、网络数据）
- **模型和计算能力**：深度学习、GPU、分布式系统
- **广泛的应用场景**：营销、广告、金融、交通、医疗等



Data：数据

X ：表示、特征、指标

F ： 模型

Y ： 智慧，也即预测任务或目标

- 大数据是指数据采集、数据清洗、数据分析和数据应用的整个流程中的理论、技术和方法。
- 机器学习是大数据分析的核心内容。机器学习解决的是找到将X和Y关联的模型F，从Data到X的步骤通常是人工完成的（特征工程）。
- 深度学习是机器学习的一部分，其核心是自动找到对特定任务有效的特征，也即自动完成Data到X的转换。
- 如果我们的任务Y是模拟人类（自动驾驶、围棋AlphaGo）的行为，则这类任务称为人工智能。深度学习也是目前AI中的核心技术。

机器学习方法的分类

- 有监督学习（supervised learning）
  - 数据集中的样本带有标签，有明确目标
  - 回归和分类
- 无监督学习（unsupervised learning）
  - 数据集中的样本没有标签，没有明确目标
  - 聚类、降维、排序、密度估计、关联规则挖掘
- 强化学习（reinforcement learning）
  - 智慧决策的过程，通过过程模拟和观察来不断学习、提高决策能力
  - 例如：AlphaGo

有监督学习

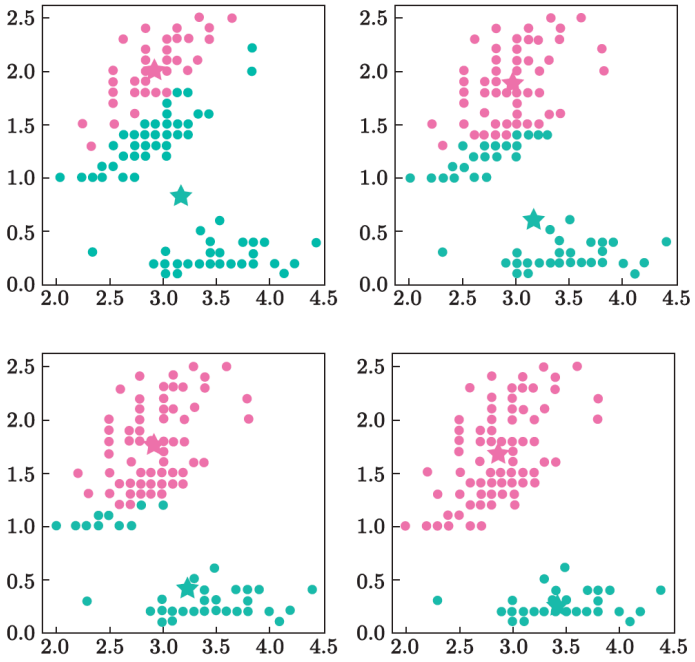
- 数据集中的样本带有标签
- 目标：找到样本到标签的最佳映射
- 应用场景：垃圾邮件分类、病理切片分类、客户流失预警、客户风险评估、房价预测等。
- 典型方法
  - **回归模型**：线性回归、岭回归、LASSO和回归样条等
  - **分类模型**：逻辑回归、K近邻、决策树、支持向量机等

无监督学习：根据数据本身的分布特点，挖掘反映数据的内在特性

- 聚类：将数据集中相似的样本进行分组，使得：
  - 同一组对象之间尽可能相似；
  - 不同组对象之间尽可能不相似。
- 应用场景：
  - 基因表达水平聚类：根据不同基因表达的时序特征进行聚类，得到基因表达处于信号通路上游还是下游的信息
  - 篮球运动员划分：根据球员相关数据，将其划分到不同类型（或者不同等级）的运动员阵营中
  - 客户分析：把客户细分成不同客户群，每个客户群有相似行为，做到精准营销

无监督学习：例子

- 1.选择K个点作为初始质心
- 2. Repeat：
  - 将每个点指派到最近的质心，形成K个簇
  - 重新计算每个簇的质心
- 3. 直到质心不发生变化

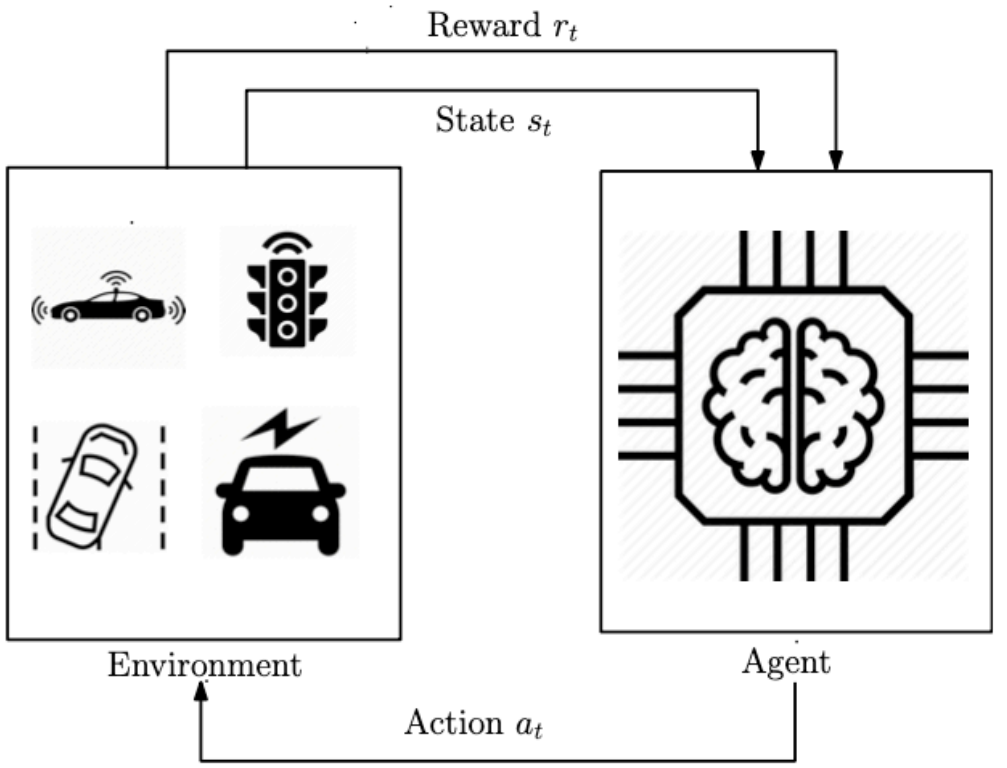


- **电信客户数据集**
- 入网时间、月均流量、月均话费、欠费额、欠费月数等

→ K-means →

- 客群1：忠诚度高，消费能力中等，信用较好
- 客群2：新用户，消费能力高，信用一般
- 客群3：新用户，消费能力一般，信用较差

- 基本概念
  - agent：智能体
  - environment：环境
  - state：状态，s_t
  - action：行动，a_t
  - reward：奖励，r_t



- 策略：$\pi(a|s)$
- 目标：
  - 求解最大化效用 $\mathrm{E}(\sum_t \gamma^t r_t)$ 的最优策略

- 数据集：一组样本的集合。
- 样本：数据集的一行。一个样本包含一个或多个特征，此外还可能包含一个标签。
- 特征：在进行预测时使用的输入变量。

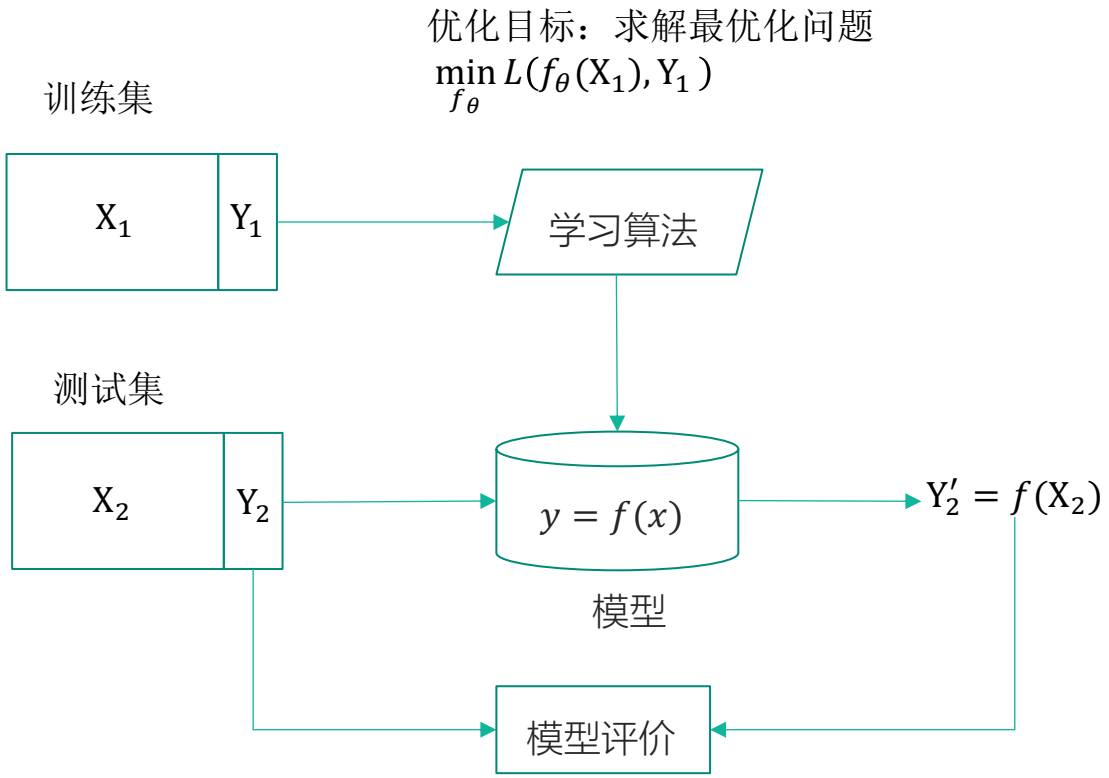
一个样本

建成年代 (特征)	面积 (特征)	布局 (特征)	房价 (标签)
1988	75	3室1厅	780
1988	60	2室1厅	705
1996	210	3室1厅	1400
2004	39	1室1厅	420
2010	90	2室2厅	998

- 训练集：用于训练模型的数据集
- 测试集：用于测试模型的数据集
- 模型：建立数据的输入$\boldsymbol{x}$和输出y之间的映射关系 $y = f(\boldsymbol{x})$
- 损失函数：$L(f(\boldsymbol{x}_i), y_i) = (f(\boldsymbol{x}_i) - y_i)^2$
- 优化目标：

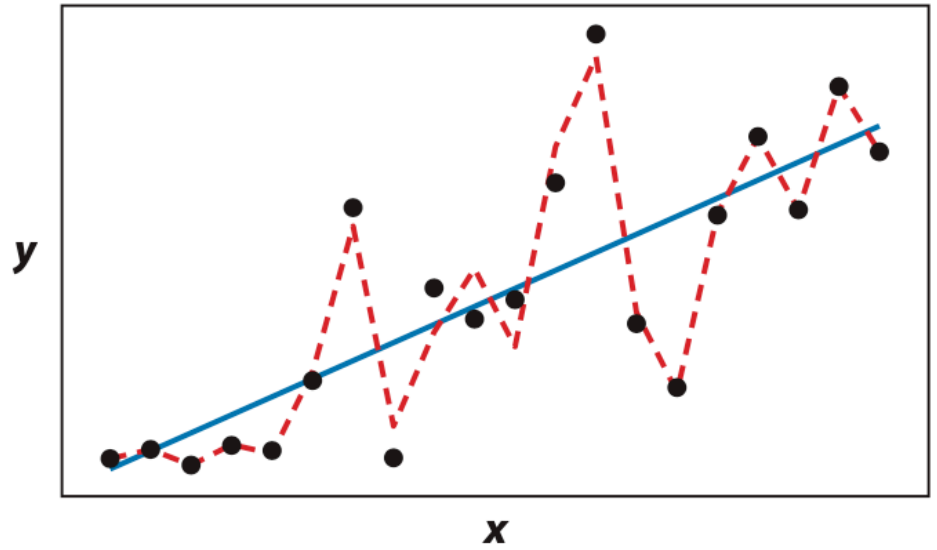
$$\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i))$$

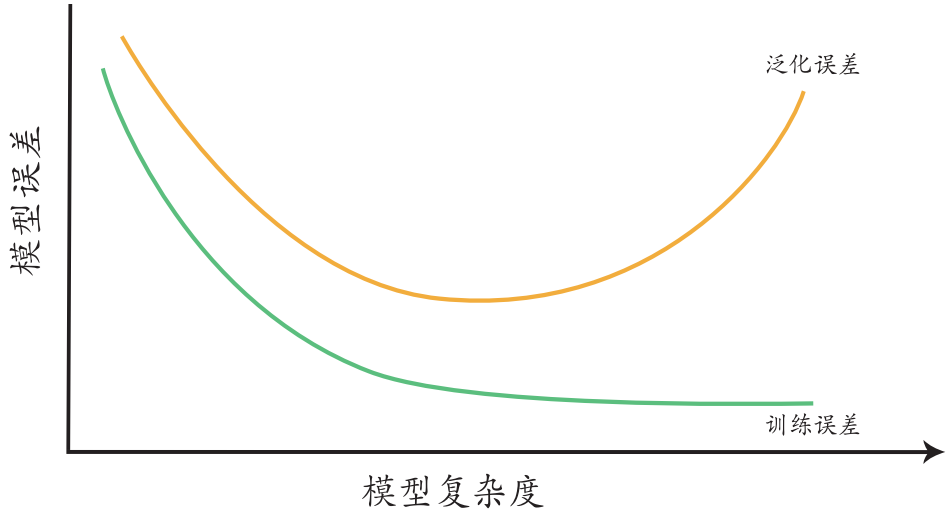
机器学习的一般流程


优化目标：求解最优化问题
$\min_{f_\theta} L(f_\theta(X_1), Y_1)$
训练集
X_1
Y_1
学习算法
测试集
X_2
Y_2
$y = f(x)$
模型
$Y_2' = f(X_2)$
模型评价

过度拟合问题

- 模型过于复杂(例如参数过多)，导致所选模型对已知数据预测得很好，但对未知数据预测很差。



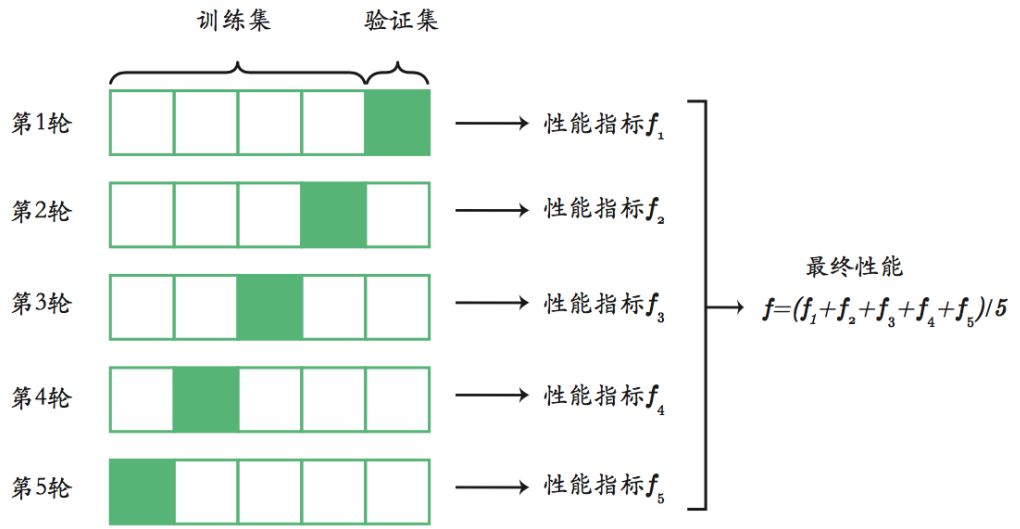


- 正则化：$\min_{f \in F} \frac{1}{n} \sum_{i=1}^{n} L(y_i, f(\boldsymbol{x}_i)) + \lambda J(f)$

- 正则化是模型选择的典型方法
- 在误差函数上加一个正则项，正则项通常为参数向量的范数
- 在训练误差和模型复杂度之间的权衡

- 交叉验证： 基本想法是重复地使用数据。将数据集随机切分，将切分的数据集组合为训练集和测试集，在此基础上反复进行训练，测试和模型选择。

- K折交叉验证
  - 随机地将数据切分为 k 个子集；
  - 每次利用$k-1$个子集的数据训练模型，余下的数据测试模型；
  - 最后选择在k次测评中平均性能最好的模型。



机器学习的数学结构

- 数据也是有数学结构的，没有数学结构我们便无法处理数据。
  - 度量结构：表示数据之间的距离。
  - 网络结构：有些数据本身就有网络结构，如社交网络。如果没有，可以利用度量结构给数据附加一个网络结构。
  - 代数结构：将数据看作向量、矩阵或更高阶的张量。
  - 几何结构：流形、对称性等
  - …

度量结构：以文本处理为例

如何计算两篇文章的距离？

- 以字典上所有的词作为坐标，对应的文章中词频作为坐标值，便可以将文章表示为向量。
- 使用余弦相似度来计算文章的距离：

$$\cos(D_1, D_2) = \frac{D_1 \cdot D_2}{\|D_1\|\|D_2\|}$$

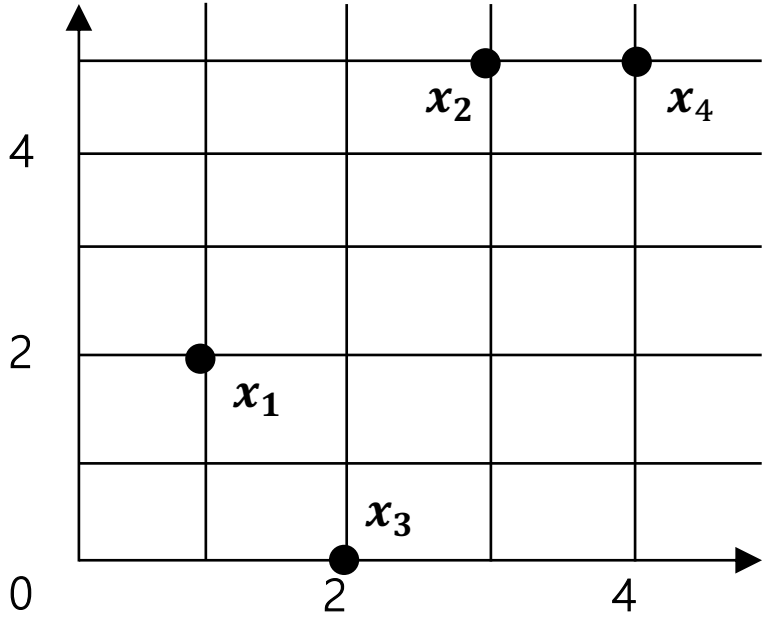
	Team	Coach	Hockey	Baseball	Soccer	penalty	Score	Win	Loss	Season
D1	5	0	3	0	2	0	0	2	0	0
D2	3	0	2	0	1	1	0	1	0	1

上表中实例 1 和示例 2 的余弦相似度为：

$$\cos(D_1, D_2) = \frac{5*3+0*0+3*2+0*0+2*1+0*1+2*1+0*0+0*1}{(25+9+4+4)^{0.5}*(9+4+1+1+1+1)^{0.5}} \approx 0.94$$

其他常用的距离

点集	特征1	特征2
x_1	1	2
x_2	3	5
x_3	2	0
x_4	4	5



- 曼哈顿距离

L_1	x_1	x_2	x_3	x_4
x_1	0			
x_2	5	0		
x_3	3	6	0	
x_4	6	1	7	0

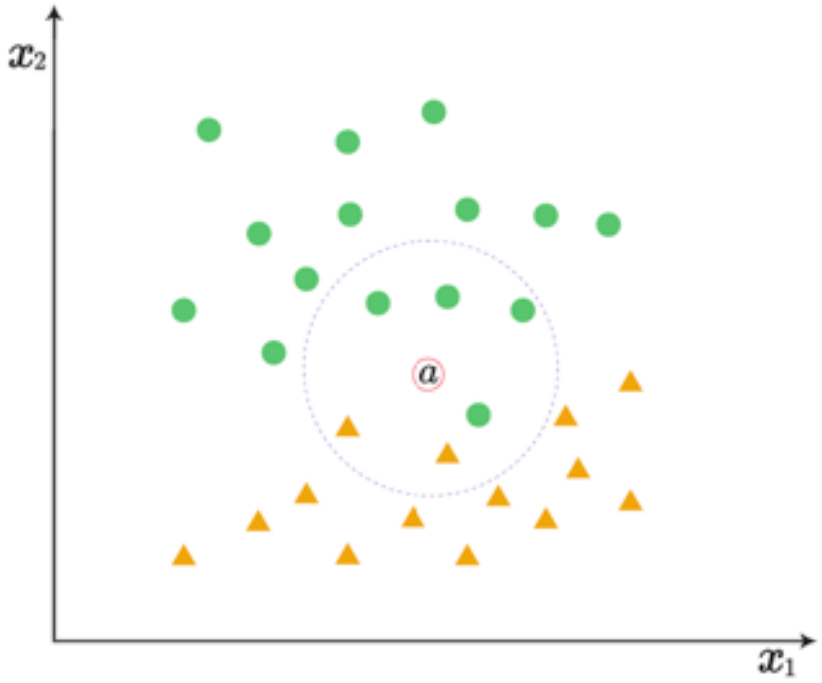
- 欧式距离

L_2	x_1	x_2	x_3	x_4
x_1	0			
x_2	3.61	0		
x_3	2.24	5.1	0	
x_4	4.24	1	5.39	0

- 极大距离

L_∞	x_1	x_2	x_3	x_4
x_1	0			
x_2	3	0		
x_3	2	5	0	
x_4	3	1	5	0

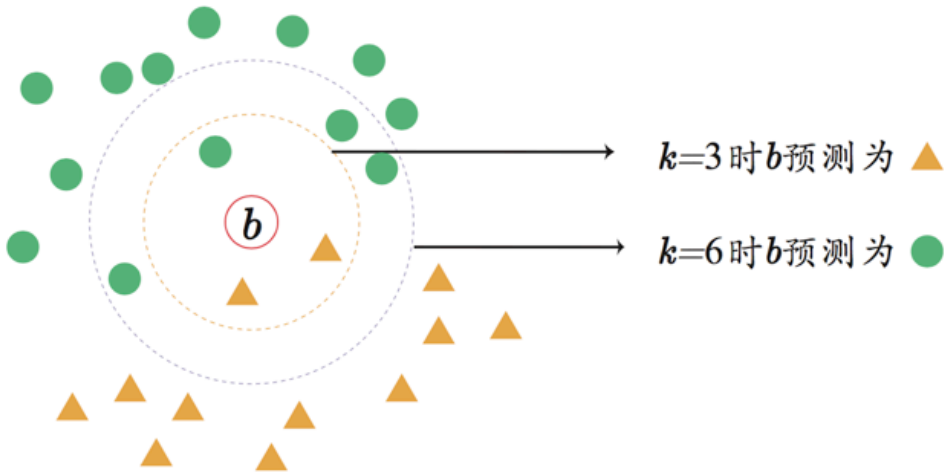
- 当对测试样本进行分类时
- 找到训练集中与该样本集最相似的 k 个样本
- 根据 k 个样本的标签确定测试样本的标签



a 的6个邻居样本中，有4个正类，2个负类，因此 a 的标签为正类

K近邻：K的选择

- 如右图，当 $k=3$ 和 $k=6$ 时，样本 b 会被分到不同的类中

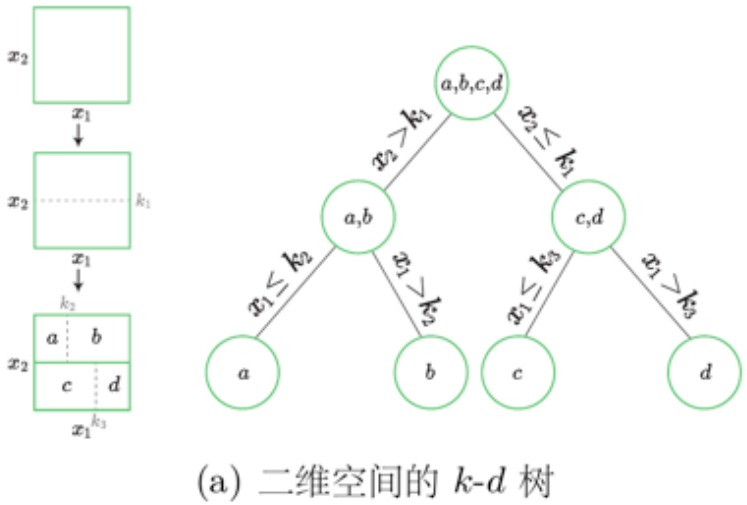


K近邻：提高计算速度

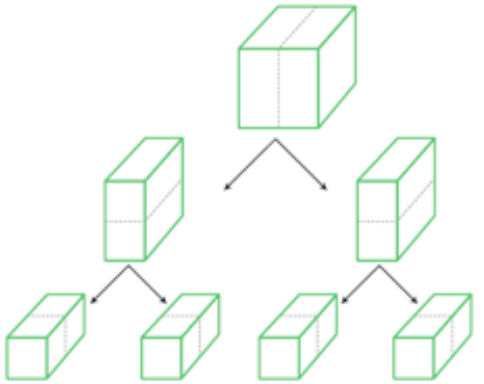
- K近邻算法最常用的数据结构为 k-d树，它是二叉搜索树在多维空间上的扩展

- 当落在某一个节点的超立方体中的样本数少于给定阈值时，节点便不再进一步分裂

- 在K近邻算法中，k-d树的作用是对训练数据集构建索引，从而在预测时，能够快速找到与测试样本近似的样本

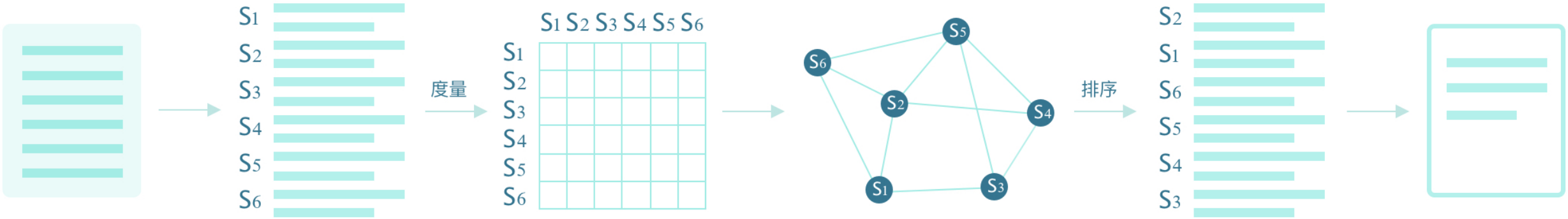


(a) 二维空间的 k-d 树



(b) 三维空间的 k-d 树

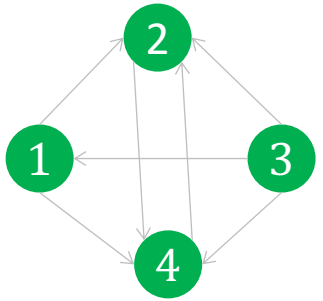
- 有了数据的度量结构，可以很容易定义一个网络结构。网络表示为 $G = \langle V, E \rangle$， V 表示节点， E 表示边。
- 例如两个样本的距离小于某个阈值，就连一条边。
- 也可以进一步将边赋予权重，权重就是两个样本的相似度。



文本摘要

- 在网络结构上定义邻接矩阵 $\boldsymbol{A} = [a_{ij}]$，其中$a_{ij}$定义为：

$$a_{ij} = \begin{cases} 1，\text{节点}i\text{与}j\text{相连} \\ 0，\text{节点}i\text{与}j\text{不相连} \end{cases}$$



$$\pi_4 \leftarrow \frac{1}{2}\pi_1 + \pi_2 + \frac{1}{3}\pi_3$$

- 从邻接矩阵得到概率转移矩阵，$\boldsymbol{T} = [t_{ij}]$，其中$t_{ij} = \frac{a_{ij}}{\sum_j a_{ij}}$

- 如果用 π_i 表示节点 i 的重要性，PageRank 算法主要在求解方程：

$$\boldsymbol{\pi} = \boldsymbol{\pi T}$$

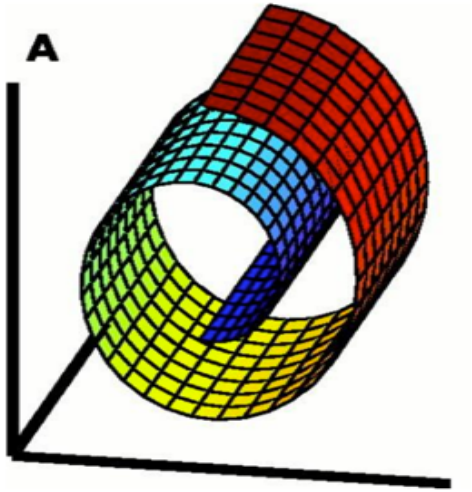
- 可见 PageRank 的解是转移矩阵特征值 1 对应的特征向量

- 如果网络是连通的，则算法有唯一的正的解。（Perron-Frobenius 定理）

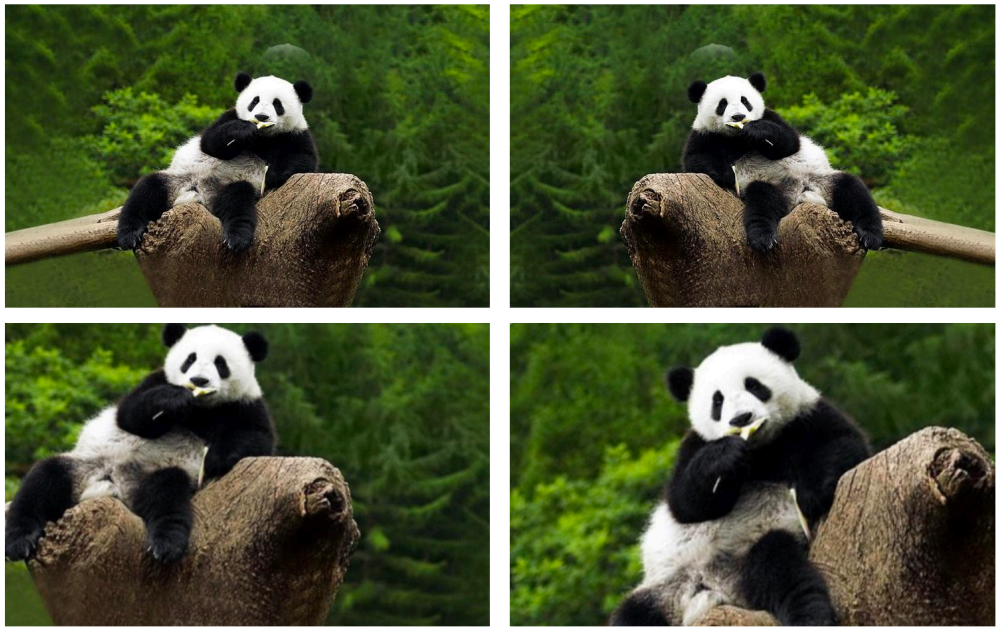
$$\boldsymbol{A} \quad \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

$$\boldsymbol{T} \quad \begin{bmatrix} 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 \\ 1/3 & 1/3 & 0 & 1/3 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

几何结构：流形和对称性等



- 很多看上去是高维空间的数据集，实际上在高维空间的一个低维的流形上
- 深度学习为什么不容易过度拟合？

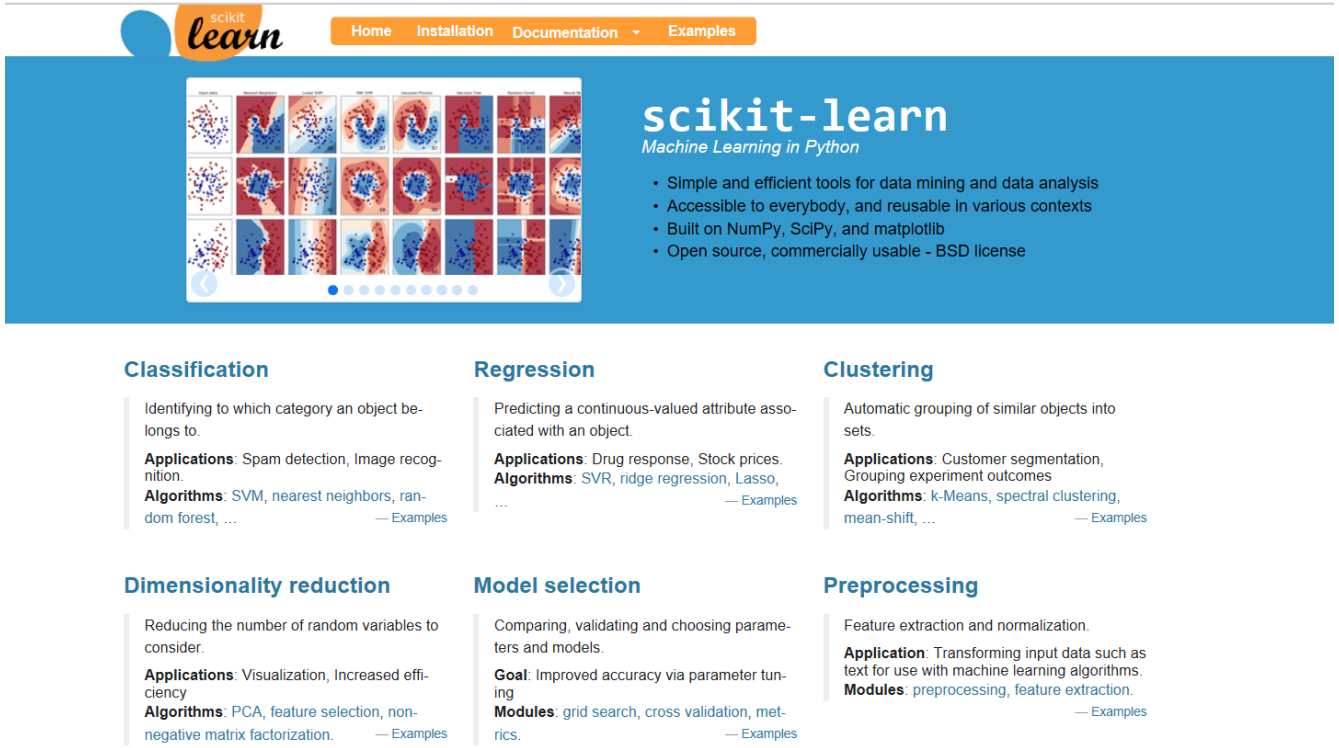


- 图像往往具有旋转不变性和平移不变性等
- 例如卷积神经网络主要考虑了这些性质

Scikit-learn概览

- Scikit-learn
  - Python用于数据建模的第三方库
  - 实现主要的机器学习、数据挖掘算法

- Scikit-learn的主要功能：
  - 数据集预处理
  - 数据集划分
  - 构建模型
  - 模型提升
  - 模型评估



https://scikit-learn.org/

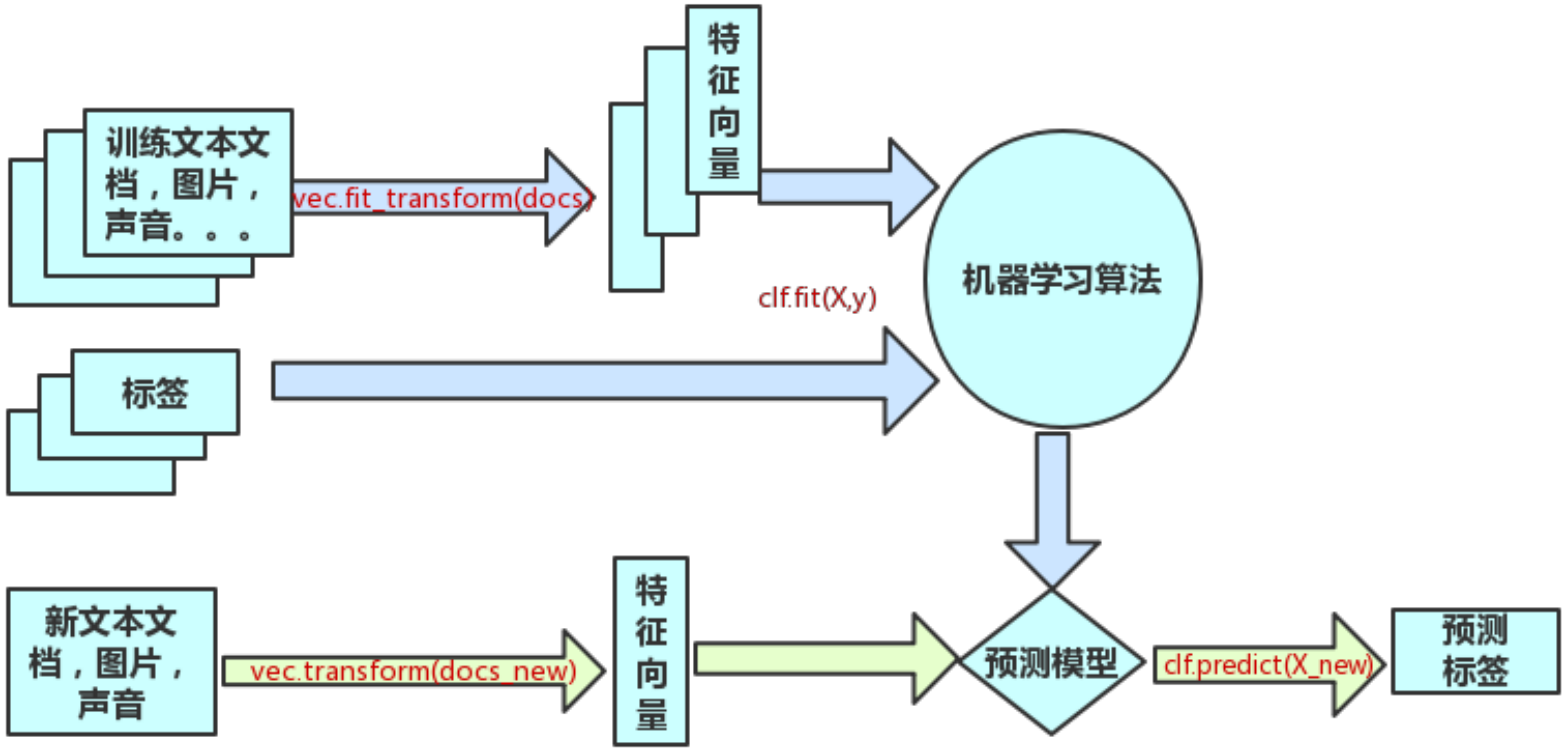
Scikit-learn的安装

- 若电脑中已安装了Anaconda科学计算环境，由于Anaconda中已包含大量常用库，则可直接使用Scikit-learn，无须再安装。

- 使用pip进行Scikit-learn的安装：

```
pip  install  scikit-learn
```

Scikit-learn基本建模流程



Scikit-learn常用函数

transform 函数：数据转换

```
from sklearn import preprocessing
scaler =preprocessing.StandardScaler().fit(X_train)
X_train = scaler.transform(X_train)
X_test = scaler.transform(X_test)
```

fit 函数：模型训练

```
from sklearn.linear_model import LinearRegression
lr = LinearRegression()
lr.fit(X_train, y_train)
```

predict 函数：模型预测

```
y_pred = lr.predict(X_test)
```

Scikit-learn主要模块

模块	说明
preprocessing	数据预处理和标准化
feature_extraction	特征提取
feature_selection	特征选择
linear_model	线性模型
tree	基于树的模型
cluster	无监督聚类算法
discriminant_analysis	线性判别分析
ensemble	集成模型
metrics	模型评价
model_selection	模型选择与参数调优

案例1：使用 KNN 对新闻主题进行自动分类

现有一份中文新闻数据集，每一条新闻有一个主题标签，例如“体育”，“教育”，“文化”等。

- 如何计算新闻的相似度？
- 使用K近邻算法，构建一个自动新闻主题分类器。

案例2：使用 PageRank 对全球机场进行排序

现有一份全球航班航线网络数据集。在网络中，节点代表机场，边则代表机场之间是否有航线开通。

- 有哪些方法可以对机场进行排序？
- 请使用PageRank算法完成对网络中节点（机场）的排序。

BIBDR 北京大数据研究院 | 博雅大数据学院 BOYA BIG DATA COLLEGE | 数据酷客 COOKDATA

数据酷客

数据科学人工智能

总结

- 机器学习的基本内容，常见的机器学习方法
- 机器学习核心概念和一般流程：
  - 样本、特征、模型、模型训练、训练集、测试集、损失函数、过度拟合、正则化、交叉验证
- 机器学习中的数学结构
  - 度量结构与 K 近邻算法
  - 网络结构与 PageRank 算法
- 实践案例：中文新闻主题自动分类、航线网络中的机场排序
- 实践工具：jieba, sklearn, pandas, matplotlib, networkx 等

数据酷客

数据科学人工智能

加入数据酷客交流群