机器学习入门
A short course of machine learning

李德维
ldw@cumtb.edu.cn infhighdim.github.io

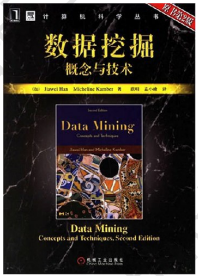
中国矿业大学(北京) 理学院

2023.05.31

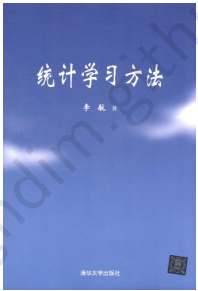
课程教材

a. 机器学习, 周志华著. 北京: 清华大学出版社, 2016年1月.
b. 数据挖掘概念与技术, (加)Jiawei Han, Micheline Kamber. 机械工业出版社, 2007年3月.
c. 统计学习方法, 李航, 清华大学出版社, 2012年3月.
d. Foundations of Machine Learning. Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar MIT Press, Second Edition, 2018.(https://cs.nyu.edu/ mohri/mlbook/)

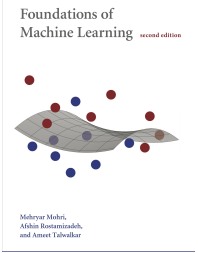
(a)



(b)



(c)



(d)

课程提纲

1. 机器学习概述
2. 逻辑斯蒂回归
3. 贝叶斯分类器
4. 决策树
5. 支持向量机
6. 聚类分析

机器学习概述

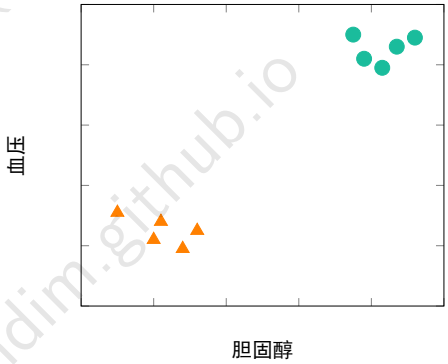
为什么要学机器学习？

色泽	根蒂	敲击声	好瓜
青绿	蜷缩	浊响	是
乌黑	蜷缩	浊响	是
青绿	硬挺	清脆	否
乌黑	稍蜷	沉闷	否

机器学习概述

为什么要学机器学习？

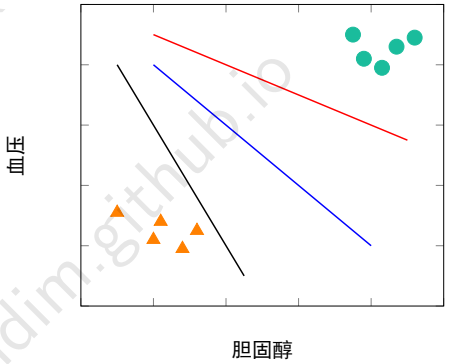
色泽	根蒂	敲击声	好瓜
青绿	蜷缩	浊响	是
乌黑	蜷缩	浊响	是
青绿	硬挺	清脆	否
乌黑	稍蜷	沉闷	否



机器学习概述

为什么要学机器学习？

色泽	根蒂	敲击声	好瓜
青绿	蜷缩	浊响	是
乌黑	蜷缩	浊响	是
青绿	硬挺	清脆	否
乌黑	稍蜷	沉闷	否



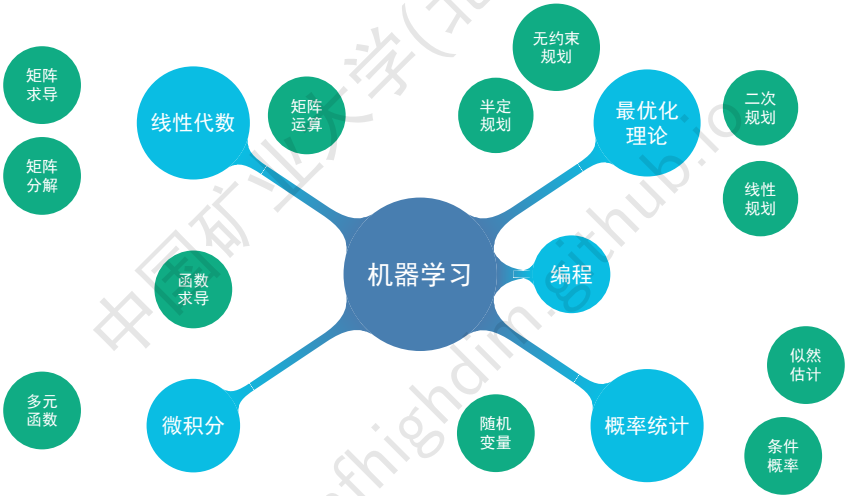
机器学习概述

(a) 人脸识别

(b) 游戏玩家匹配

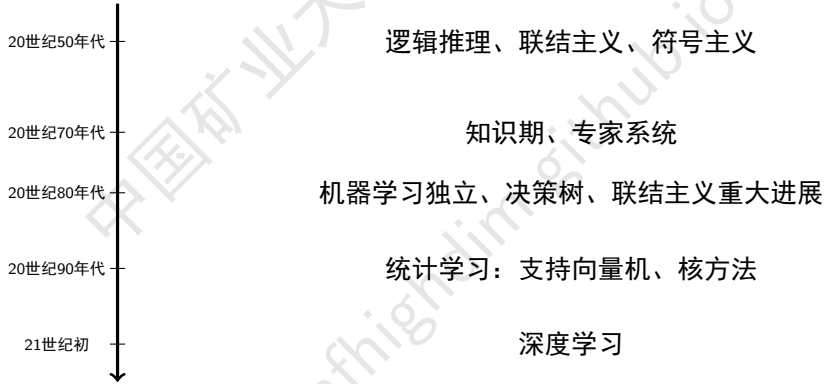
语音识别(微信语音锁)、新闻推荐(今日头条)、无人驾驶(特斯拉)、医学诊断、金融欺诈...

机器学习概述–知识基础

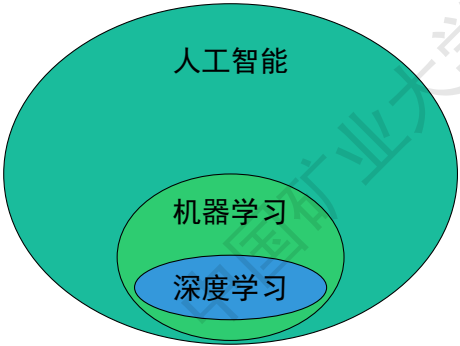


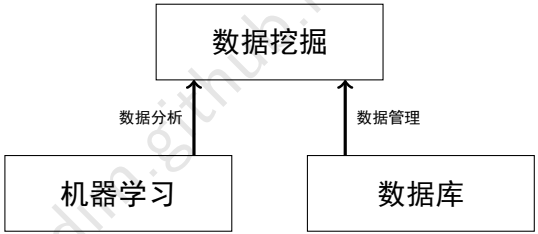
机器学习概述

机器学习 (Machine Learning) 是对研究问题进行模型假设，利用计算机从训练数据中学习得到模型参数，并最终对数据进行预测和分析的一门学科。机器学习涉及多个领域的理论交叉，包括最优化理论、概率论、统计学、逼近论、算法复杂度理论等。



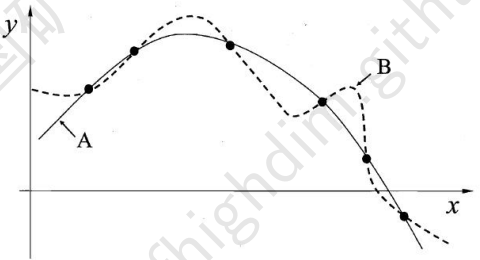
机器学习概述





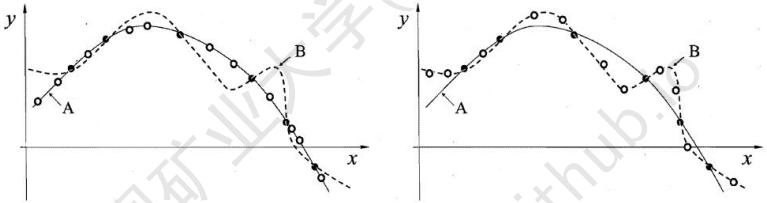
机器学习概述–基本概念

令数据集$D = \{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\} \in R^{m \times n}$，它包含了$m$个数据样本，以及对应的标签信息$y$，每个样本的特征(属性)个数为$n$，也称为样本的维数。标签一般为标量，可以是连续或离散。数据集一般分为两部分，训练集和测试集。从训练数据中学得模型的过程称为"学习"(learning)或"训练"(training)，训练完成后利用测试集对模型进行评估。模型评估的目标是验证学习到的模型能否具有适用新数据的能力，这种能力称为泛化(generalization) 能力。



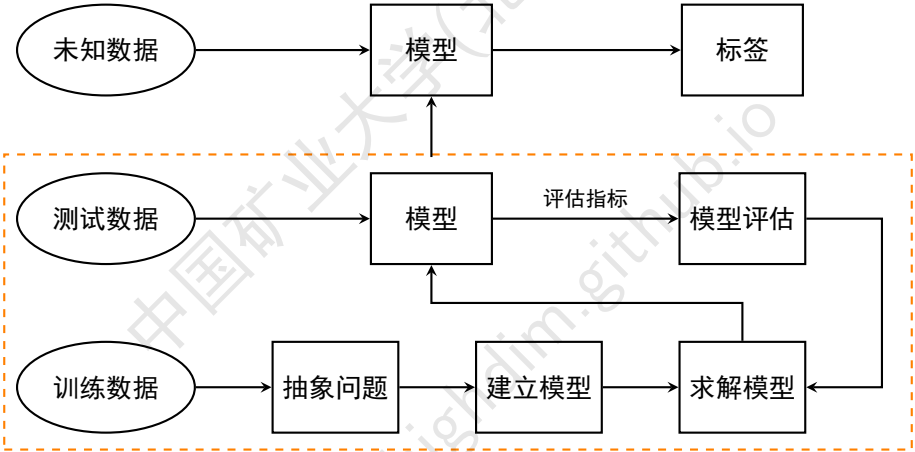
奥卡姆(Occam's razor)剃刀原理: 若有多个假设与观察一致，则选最简单的那个。

“没有免费的午餐”(No Free Lunch Theorem)

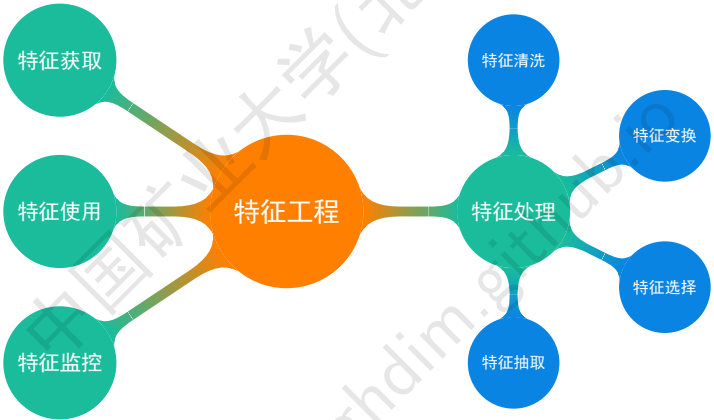


任何模型算法的性能和效果都与具体的问题相关，没有一个算法能在所有任务上都表现最优。

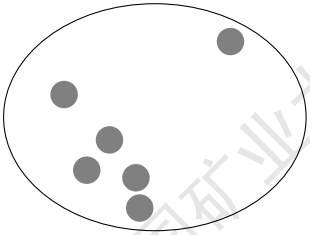
机器学习概述–学习步骤



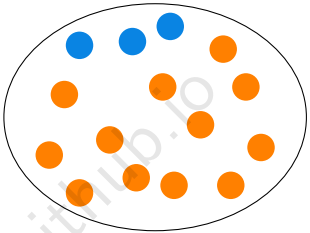
机器学习概述–特征工程



机器学习概述–特征清洗



(a) 离群点：剔除、平滑



(b) 不平衡：上下采样、数据增强

编号	特征1	特征2	特征3
1	5	20.1	3.6
2	4.3	9.5	5.9
3	5	20.1	3.6

(c) 重复值：保留、剔除

编号	特征1	特征2	特征3
1	5	20.1	
2	4.3		5.9
3	5	20.1	3.6

(d) 缺失值：剔除、人工填写、全局常量、属性平均值、同类属性平均值、算法推测

机器学习概述–特征处理

- 聚集：对数据集进行汇总和聚集。例如对于一份销售数据，可以考虑聚集每天的数据，计算月销售额或年销售额；
- 数据泛化: 用高层次概念替换原始数据。数值属性age，可以离散化成young, middle, senior等；
- 规范化：标准化、归一化。将数据按比例缩放，映射到一个标准区间内，比如$[0,1],[-1,1]$；(思考：为什么要做归一化？)

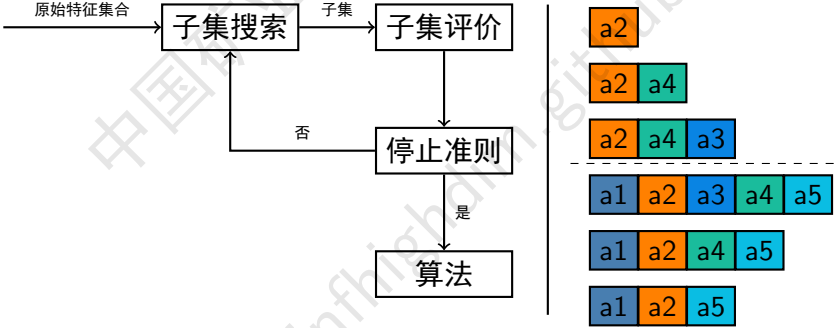
$$\text{max-min}: x' = \frac{x - min}{max - min}; z - score: x' = \frac{x - \hat{X}}{\sigma}; \text{小数定标}: x' = \frac{x}{10^j}$$

非线性归一化、批量归一化(BatchNormalization)

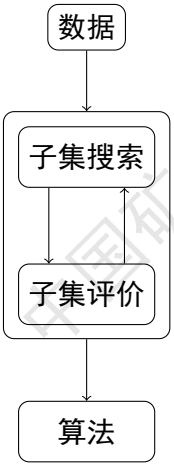
- 特征构造：例如在电商领域，用户行为数据表中每条记录为某个用户的一次浏览行为或一次点击行为，可以由此构造出用户最近一次浏览的时长、用户最近一次登录的点击次数等特征。

机器学习概述–特征选择

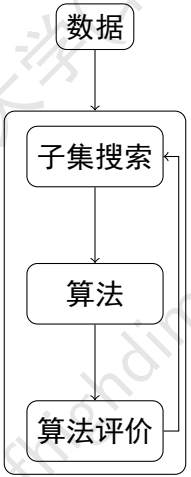
- **动机**：相关特征、无关特征
- **优点**：维数灾难–随着维数的增加，计算量呈指数倍增长的一种现象；降低学习任务的难度–更容易学习到好的模型
- **基本思路**：子集搜索–前向、后向、双向；子集评价



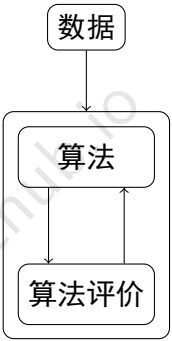
机器学习概述–特征选择



(a) 过滤式选择



(b) 包裹式选择



(c) 嵌入式选择

机器学习概述–特征选择

Relief (Relevant Features): 设计了一个"相关统计量"来度量特征的重要性
给定训练集，对于每个样本x_i，分别寻找空间中的同类近邻$x_{i,nh}$和异类近邻$x_{i,nm}$，那么使用如下方式计算特征j的重要性，

$$\delta^j = \sum_i -\text{diff}(x_i^j, x_{i,nh}^j)^2 + \text{diff}(x_i^j, x_{i,nm}^j)^2 \tag{1}$$

其中

$$\text{diff}(a,b) = \begin{cases} \mathbf{1}_{a=b}, & \text{离散} \\ |a-b|, & \text{连续} \end{cases} \tag{2}$$

对基于不同样本得到的估计结果进行平均, 就得到各属性的相关统计量分量, 分量值越大，则对应属性的分类能力就越强。(思考：为什么有效？)

机器学习概述–特征选择

LVW (Las Vegas Wrapper)：使用随机策略来进行子集搜索，并以最终分类器的误差为特征子集评价准则。

输入： 数据集 D;
特征集 A;
学习算法 $\mathfrak{L}$;
停止条件控制参数 T.

过程：

$E = \infty$;
$d = |A|$;
$A^* = A$;
$t = 0$;
while $t < T$ **do**
  随机产生特征子集 A';
  $d' = |A'|$;
  $E' = \text{CrossValidation}(\mathfrak{L}(D^{A'}))$;
  if $(E' < E) \vee ((E' = E) \wedge (d' < d))$ **then**
    $t = 0$;
    $E = E'$;
    $d = d'$;
    $A^* = A'$
  else
    $t = t + 1$
  end if
end while

输出： 特征子集 A^*

机器学习概述–特征选择

给定数据集D, 考虑线性回归模型

岭回归

$$\min_w \sum_{i=1}^{m}(y_i - w^\top x_i)^2 + \lambda||w||^2$$

LASSO

$$\min_w \sum_{i=1}^{m}(y_i - w^\top x_i)^2 + \lambda||w||$$

哪一种可以做特征选择？

机器学习概述–特征选择

给定数据集D, 考虑线性回归模型

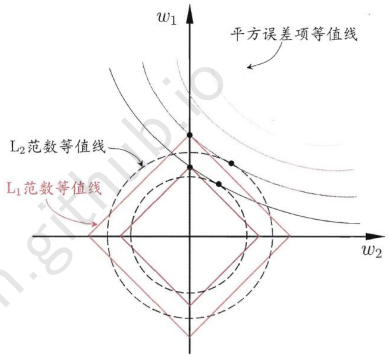
岭回归

$$\min_w \sum_{i=1}^{m}(y_i - w^\top x_i)^2 + \lambda||w||^2$$

LASSO

$$\min_w \sum_{i=1}^{m}(y_i - w^\top x_i)^2 + \lambda||w||$$

哪一种可以做特征选择？

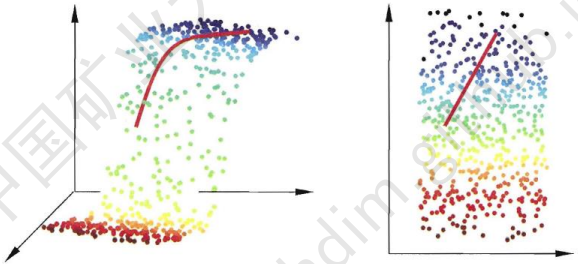


机器学习概述–特征抽取

降维：高维空间，低维映射，学习难度降低，学习效果不变甚至更好

学习映射矩阵W，获得

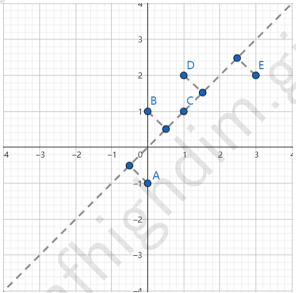
$$z = Wx$$



机器学习概述–特征抽取

主成分分析(Principal Component Analysis，简称PCA)：将n维原始特征映射到k维($k < n$)上，称这k维特征为主成分。其主要目标是将特征维度变小，同时尽量减少信息损失。
PCA将原始样本点投影到理想的超平面上：

- **最大可分性**: 样本点在这个超平面上的投影能尽可能分开;
- 最近重构性: 样本点到这个超平面的距离都足够近.



机器学习概述–特征抽取

下面从最大可分性角度推导。给定m个样本点$x'_1, x'_2, \cdots, x'_m$，首先进行中心化，令$x_i = x'_i - \frac{1}{m}\sum_{i=1}^{m} x'_i$，那么有

$$\mu_x = \sum_{i=1}^{m} x_i = 0 \tag{3}$$

为了将原始样本投影到低维空间，计算样本x_i与单位方向向量w的内积$y_i = x_i^\top w$，投影后的方差为

$$Var(y) = \frac{1}{m}\sum_{i=1}^{m}(y_i - \mu_y)^2 = \frac{1}{m}\sum_{i=1}^{m}(y_i - \frac{1}{m}\sum_{i=1}^{m} x_i w)^2 = \frac{1}{m}\sum_{i=1}^{m} y_i^2 \tag{4}$$

$$Var(y) = \frac{1}{m}\sum_{i=1}^{m}(x_i^\top w)^2 = w^\top(\frac{1}{m}\sum_{i=1}^{m}x_i x_i^\top)w = w^\top \Sigma w \tag{5}$$

基于最大可分性，我们构建如下最优化问题

$$\max_{w} \quad w^\top \Sigma w \tag{6}$$

$$s.t. \quad w^\top w = 1 \tag{7}$$

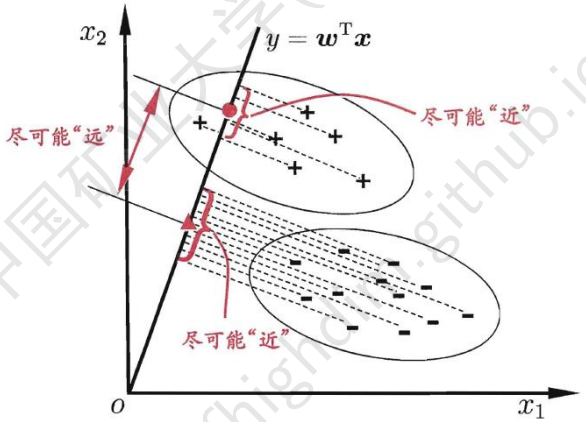
对上式构建拉格朗日函数$L(w,\lambda) = w^\top \Sigma w + \lambda(1 - w^\top w)$，令$\nabla_w L = 0$可得

$$\Sigma w = \lambda w \tag{8}$$

将上式带入到方差可得$Var(y) = \lambda$.

机器学习概述–特征抽取

线性判别分析(Linear Discriminant Analysis，简称LDA)



定义类内散度矩阵

$$S_w = \Sigma_0 + \Sigma_1 = \sum_{x \in X_0} (x - \mu_0)(x - \mu_0)^\top + \sum_{x \in X_1} (x - \mu_1)(x - \mu_1)^\top$$

和类间散度矩阵

$$S_b = (\mu_0 - \mu_1)(\mu_0 - \mu_1)^\top$$

可以构建如下最优化问题

$$\max \frac{w^\top S_b w}{w^\top S_w w}$$

可以等价转化为如下最优化问题

$$\min_w \quad -w^\top S_b w \tag{9}$$

$$s.t. \quad w^\top S_w w = 1 \tag{10}$$

机器学习概述–特征抽取

利用拉格朗日乘子法，有

$$S_b w = \lambda S_w w$$

由于$S_b w$方向恒为$\mu_0 - \mu_1$，不妨令$S_b w = \lambda(\mu_0 - \mu_1)$, 那么有

$$w = S_w^{-1}(\mu_0 - \mu_1)$$

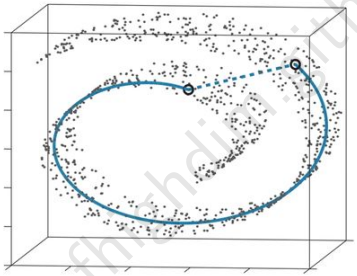
考虑到数值解的稳定性，一般会先对S_w进行奇异值分解，然后再求逆矩阵。
LDA也可以推广至多分类情形。

机器学习概述–特征抽取

- 多维标度分析(MDS): 降维前后能够保持距离关系不变

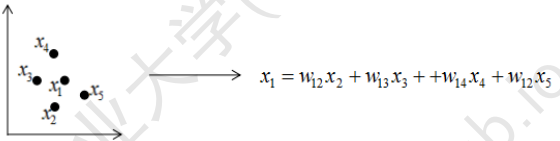
$$\min_{w} \sum_{ij} (||Wx_i - Wx_j|| - d(x_i, x_j))^2$$

- 等距特征映射ISOMAP：引入测地距离



机器学习概述–特征抽取

- 局部线性嵌入LLE：保持局部线性关系



- 拉普拉斯特征映射LE：基于图构建邻接矩阵，降维后仍能保持原有的数据结构信息

$$\min \sum_{ij} ||z_i - z_j||^2 w_{ij}$$

其中$w_{ij} = \exp(-\frac{||x_i - x_j||^2}{\sigma^2})$

机器学习概述–特征抽取

- t-分布随机近邻嵌入tsne：利用概率分布定义距离关系

$$p_{ij} = \frac{exp(-||x_i - x_j||^2)/2\sigma^2}{\sum_{k \neq l} exp(-||x_k - x_l||^2)/2\sigma^2}$$

$$q_{ij} = \frac{exp(-||z_i - z_j||^2)}{\sum_{k \neq l} exp(-||z_k - z_l||^2)}$$

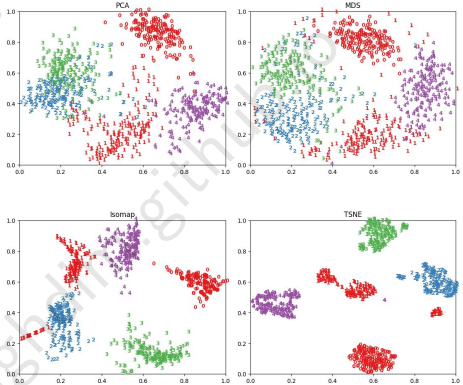
目标函数

$$C = KL(P||Q) = \sum_{ij} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

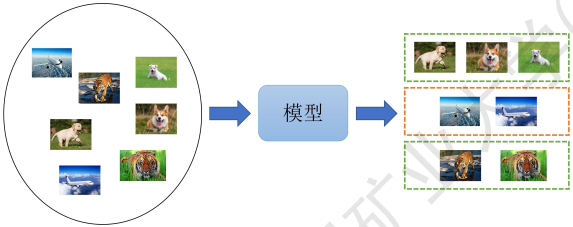
机器学习概述–特征抽取

手写字体集合

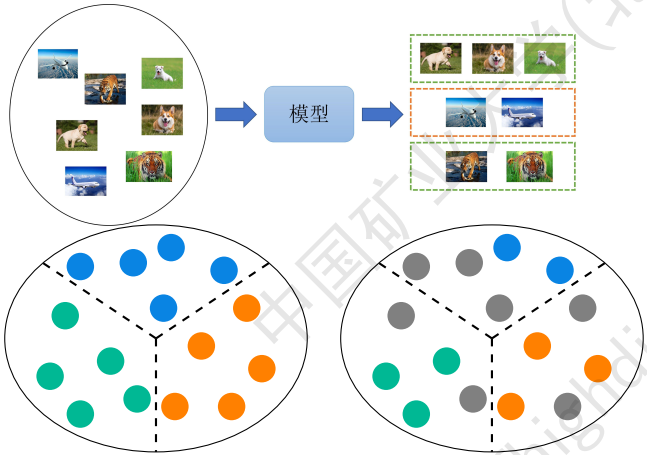
降维方法对比



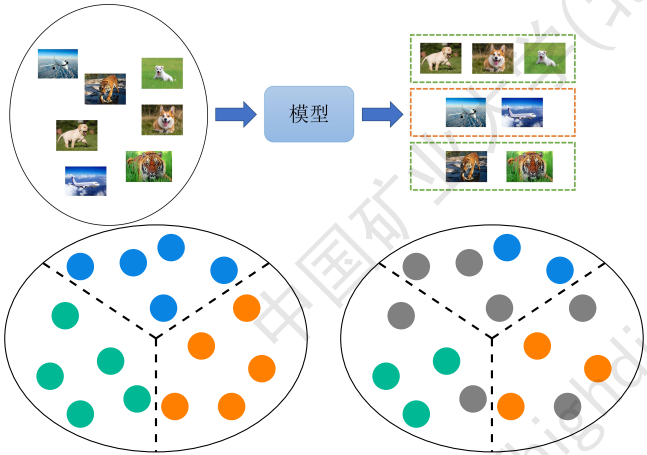
机器学习概述–分类问题



机器学习概述–分类问题



机器学习概述–分类问题



- 医疗诊断：患者生理指标，是否患病
- 金融欺诈：客户信息、还款历史，是否违约
- 人脸识别：面部信息，谁
- 市场营销：消费者行为，是否目标客户

机器学习概述–分类问题评估指标

对于一个二分类问题，定义如下指标

- 真正(True Positive, TP)：被模型预测为正的正样本；
- 假正(False Positive, FP)：被模型预测为正的负样本；
- 假负(False Negative, FN)：被模型预测为负的正样本；
- 真负(True Negative, TN)：被模型预测为负的负样本；

计算比例

- 真正率$=\frac{TP}{TP+FN}$
- 假正率$=\frac{FP}{TN+FP}$
- 假负率$=\frac{FN}{TP+FN}$
- 真负率$=\frac{TN}{TN+FP}$

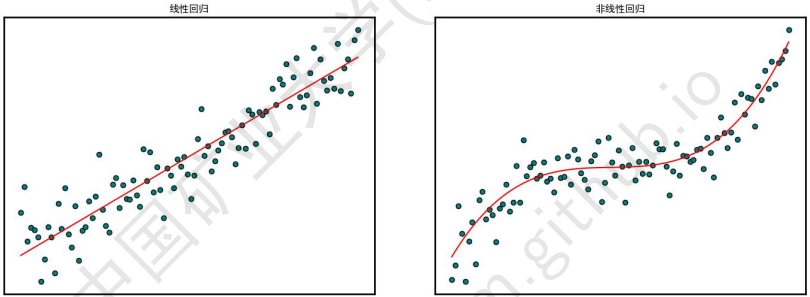
机器学习概述–分类问题评估指标

常见的指标

- 准确率(Accuracy)=$\frac{TP+TN}{TP+FP+FN+TN}$
- 精度(Precision)=$\frac{TP}{TP+FP}$
- 召回率(Recall)=$\frac{TP}{TP+FN}$
- $F1 = \frac{2\times Precision\times Recall}{Precision+Recall}$
- $F_\beta = \frac{(1+\beta^2)\times Precision\times Recall}{\beta^2\times Precision+Recall}$

ROC曲线, AUC, P-R曲线, 混淆矩阵

机器学习概述–回归



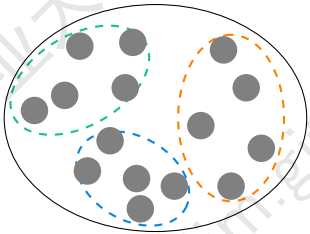
- 金融领域：历史股票价格，未来股票价格
- 医疗领域：患者信息、用药情况，血药浓度
- 人脸识别：面部信息，用户年龄

机器学习概述–回归问题评估指标

- MAE
- MSE
- RMSE
- MAPE
- R2 score
- Adjusted R2 score

机器学习概述–聚类

聚类学习是按照某种特定标准(如距离等)把一个数据集划分为不同的类或簇（子集），使得同一个簇内的数据对象的相似性尽可能大，不在同一个簇中的数据对象的差异性也尽可能地大（即聚类后同一类的数据尽可能聚集到一起，不同类数据尽量分离）。



- 社交网络：用户特征，社交圈子
- 电子商务：客户信息，分群
- 新闻分类：新闻文本，新闻类型

机器学习概述–聚类问题评估指标

评估标准：类内相似度高，类间相似度低

外部评价指标

- 纯度
- 归一化互信息(Normalized Mutual Information, NMI)
- 兰德指数(Rand index, RI)
- 调整兰德系数(Adjusted Rand index, ARI)
- R2 score
- Adjusted R2 score

内部评价指标

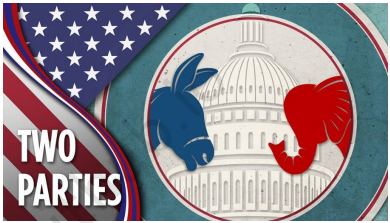
- 轮廓系数(Silhouette Coefficient)
- Calinski-Harabaz指数

机器学习概述–综合案例

2004年3月，在美国的自动驾驶车比赛，斯坦福大学机器学习专家S. Thrun的小组研制的参赛车用6小时53分钟成功走完了132英里赛程获得冠军。感知：语音识别、目标识别、物体追踪；预测：车辆行人的行为预测。

机器学习概述–综合案例

2004年3月，在美国的自动驾驶车比赛，斯坦福大学机器学习专家S. Thrun的小组研制的参赛车用6小时53分钟成功走完了132英里赛程获得冠军。感知：语音识别、目标识别、物体追踪；预测：车辆行人的行为预测。



2012美国大选，奥巴马麾下有一支机器学习团队，他们对各类选情数据进行分析，为奥巴马成功竞选提供了有力支持。涉及选民分类、选民的偏好画像。

谢谢观看!