

机器学习先导课



目录



A vertical line on the left side of the page contains four circular nodes, numbered 1 to 4 from top to bottom. Node 1 is orange, while nodes 2, 3, and 4 are blue. To the right of each node is a horizontal rectangular box. The first box is orange and contains the text '机器学习定义'. The subsequent three boxes are blue and contain the text '机器学习基本术语', '机器学习算法分类', and '性能度量' respectively. A horizontal line extends from the right side of the first orange box.

1	机器学习定义
2	机器学习基本术语
3	机器学习算法分类
4	性能度量

如何分辨花朵?

显著式编程

if 花朵大&颜色金黄&花蕊大而突出:

 return 向日葵

else:

 return 栀子花

非显著式编程

向日葵: 花瓣黄色?

 花瓣大而长?

 叶片宽而长?

栀子花: 花瓣洁白?

 花瓣短而圆?

 叶片深绿?

 花瓣交错?



生活中的其他案例



机器学习定义



Tom M. Michell

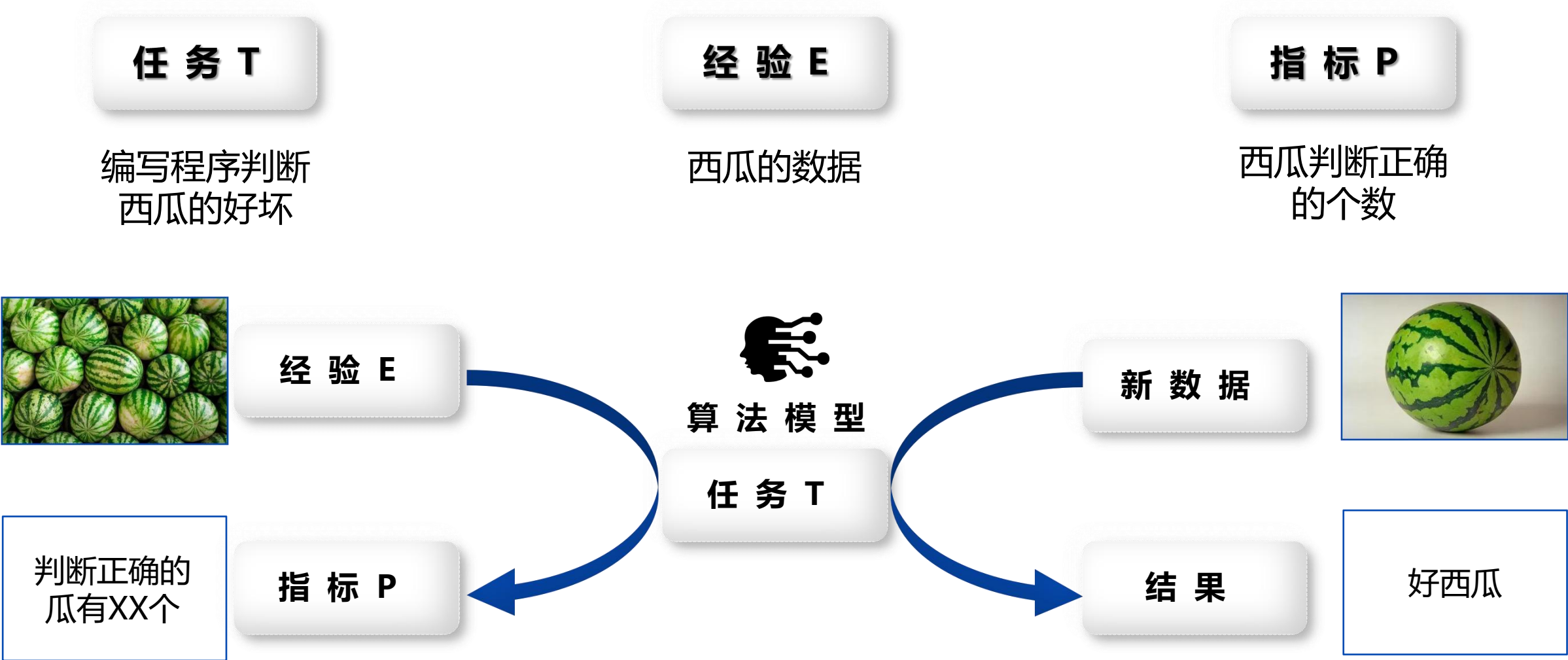
一个计算机程序被称为可以学习，是指它能够针对某个任务 T 和某个性能指标 P ，从经验 E 中学习。这种学习的特点是，它在 T 上的被 P 所衡量的性能，会随着经验 E 的增加而提高。



Wikipedia

机器学习是一门系统的学科，它关注设计和开发算法，使得机器的行为随着经验数据的积累而进化，经验数据通常是传感器数据或数据库记录。

定义实例






目录



基本术语一



	色泽	肚脐	敲声	分类
	青绿	凹陷	浊响	坏瓜
	乌黑	突出	浊响	好瓜
	微黄	突出	清脆	好瓜

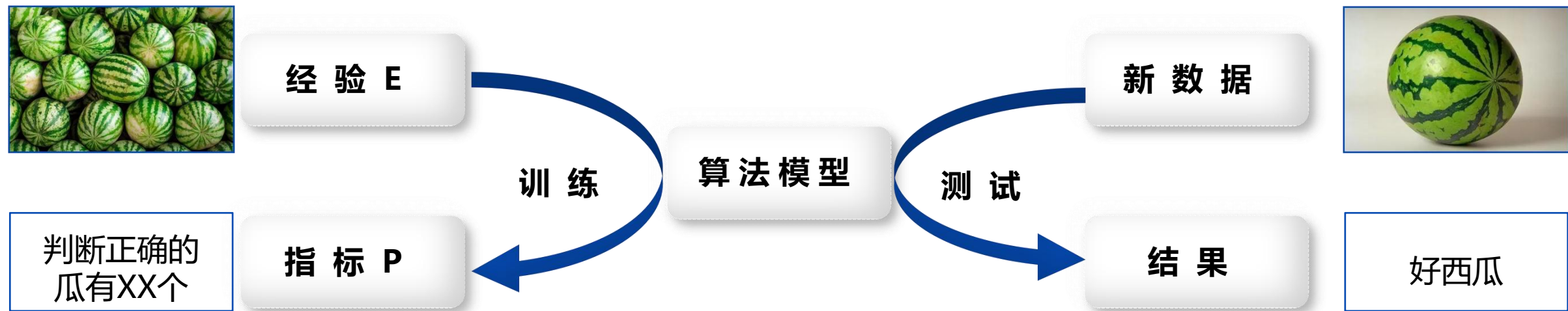
术语	描述
数据集	记录的集合称为一个‘数据集’
样本	关于一个时间或对象的描述，称为‘样本’
属性	反应事件或对象在某方面的表现或性质的事项，如：色泽、肚脐
属性值	属性上的取值，如：青绿、凹陷、浊响
属性空间	属性张成的空间称为‘属性空间’
标记	关于样本结果的信息，例如：好瓜、坏瓜

基本术语二



术语	描述
训练过程	从数据中学得模型的过程称为‘训练’
训练样本	训练过程中的每一个样本称为‘训练样本’
训练集	训练样本组成的集合称为‘训练集’
测试过程	检验模型泛化能力的过程称为‘测试’
测试集	测试过程需要的样本构成的集合
预测结果	算法对测试样本做出的判断结果

基本术语三



术语	描述
误差	模型在实际预测输出与样本的真实输出之间的差异称为‘误差’
训练误差	模型在训练集上的误差称为‘训练误差’或者‘经验误差’
泛化误差	模型在新数据上的误差成为泛化误差

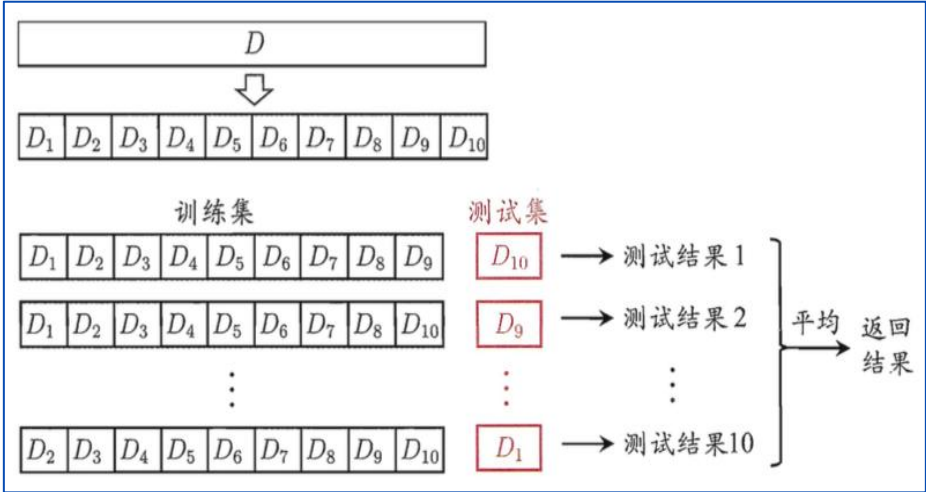
基本术语四



术语	描述
过拟合	把训练样本自身的一些特点当作了所有潜在样本都会具有的一般性质
欠拟合	对训练样本的一般性质尚未学好
归纳偏好	机器学习算法在学习过程中对某类假设的偏好

数据集划分方法

方法	留出法	交叉验证（K折交叉验证）	自助法
实现过程	划分成两个互斥的集合，测试集比例一般在2/3-4/5，若干次随即划分、重复实验后取平均值	划分成k个大小相似的互斥子集，然后进行p次实验。常见的有10次10折交叉验证	通过放回重采样重新构成一个大小相同的集合，训练后对没有参与训练的数据做验证
注意点	集合划分需要分层采样		重采样一定要放回
缺点	单次结果不稳定	稳定性和保真性取决于k的取值	改变数据集分布，引入估计偏差

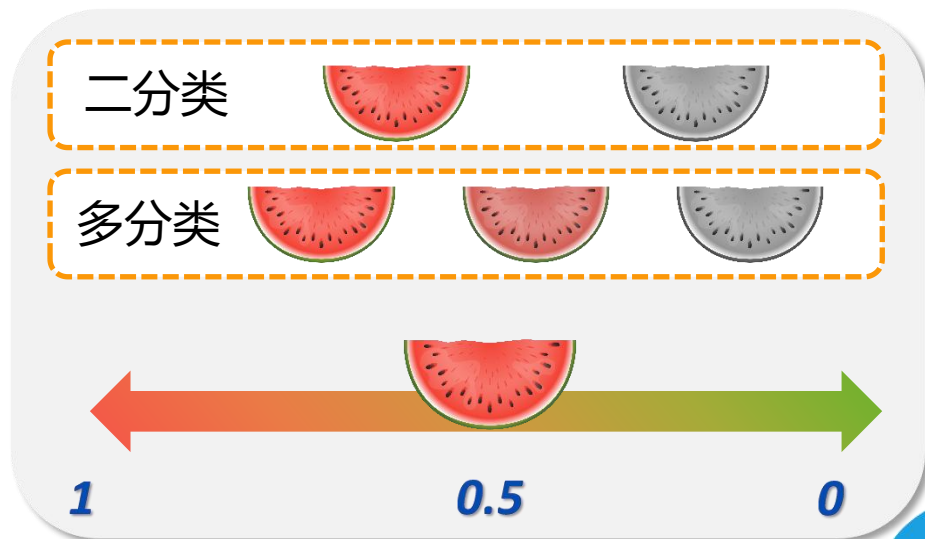


$$\lim_{m \rightarrow \infty} \left(1 - \frac{1}{m}\right)^m \mapsto \frac{1}{e} \approx 0.368$$

目录



算法分类



机器学习

RNN神经网络

CNN神经网络

BP神经网络

深度学习

有监督学习

回归

线性回归

岭回归

分类

决策树

支持向量机

半监督学习

上述算法

无监督学习

降维

主成分分析

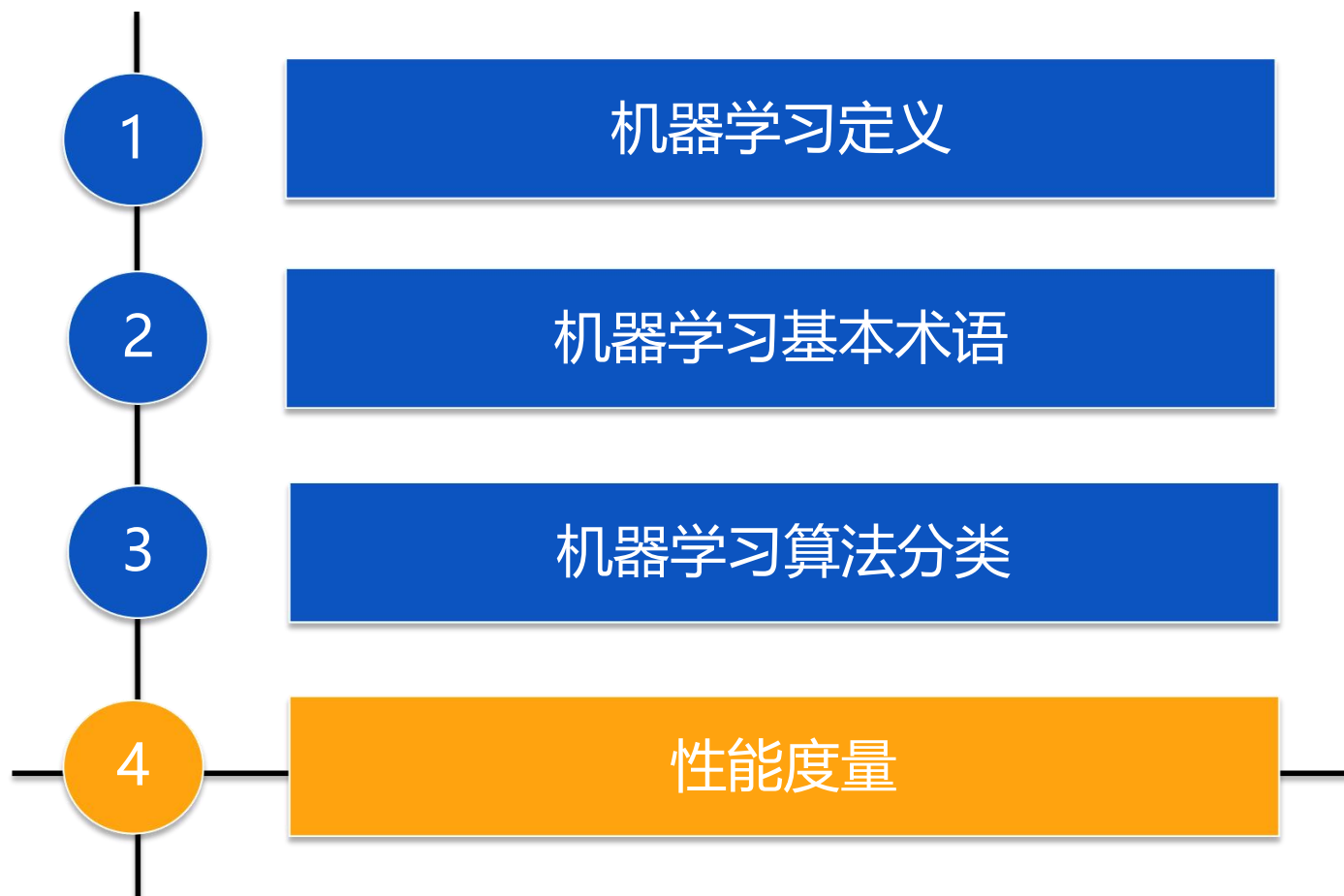
PCA

聚类

层次聚类

K-means

目录



分类算法性能度量

错误率和精度是分类任务中最常见的两种性能度量，既适应二分类任务又适应多分类任务。错误率是分类错误的样本数占样本总数的比例，精度是分类争取的样本数占总数的比例。对样本D的错误率和精度的定义如下：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^m\mathbb{I}(f(x_i)\neq y_i)$$

$$acc(f;D)=\frac{1}{m}\sum_{i=1}^m\mathbb{I}(f(x_i)=y_i)=1-E(f;D)$$

对二分类问题，可以将样根据其正是类别与算法预测类别的组合划分为真正例(True Positive)、假正例(False Positive)、真反例(True Negative)、假反例(False Negative)四种情况，令TP、FP、TN、FN分别表示其对应的样例数，显然TP+FP+TN+FN=样例总数。分类结果对应的“混淆矩阵如下”

真实情况	预测结果	
	正例	反例
正例	TP	FN
反例	FP	TN

查全率 $P=\frac{TP}{TP+FP}$

查准率 $R=\frac{TP}{TP+FN}$

F1度量 $F1=\frac{2\times P\times R}{P+R}$

回归算法性能度量

回归算法最常用的性能度量是“均方误差” (Mean Squard Error)

$$MAE(f; D) = \frac{1}{m} \sum_{i=1}^m (f(x_i) - y_i)^2$$

除了上述的均方误差，均方误差(Mean Squared Error)、均方误差根或均方根误差(Root Mean Squared Error)也是常见的指标信息。

$$MSE(f; D) = \frac{1}{m} \sum_{i=1}^m \|f(x_i) - y_i\|_2^2$$

$$RMSE(f; D) = \sqrt{\frac{1}{m} \sum_{i=1}^m \|f(x_i) - y_i\|_2^2}$$

聚类算法评价指标

好的聚类算法,一般要求类簇具有, 高的类内 (intra-cluster) 相似度, 低的类间 (inter-cluster) 相似度, 因此对于聚类算法大致可分为2类度量标准。

当一个聚类结果是基于数据聚类自身进行评估的, 这一类叫做内部评估方法。如SSE, 该统计参数计算的是拟合数据和原始数据对应点的误差的平方和, 计算公式如下:

$$SSE(f; D) = \sum_{i=1}^m \left(f(x_i) - \hat{f}(x_i) \right)^2$$

在外部评估方法中, 聚类结果是通过使用没被用来做训练集的数据进行评估。例如已知样本点的类别信息和一些外部的基准。常见外部评价指标有纯度 (Purity)、标准化互信息 (NMI) 等。



Thank you!