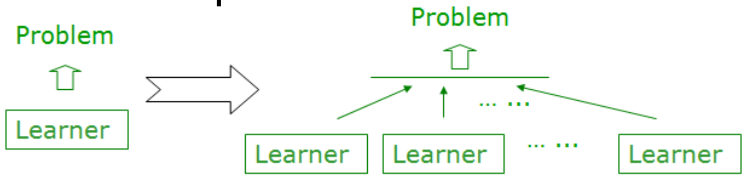
八、集成学习

主讲教师：周志华

集成学习

Ensemble Learning (集成学习): Using multiple learners to solve the problem



Demonstrated great performance in real practice

- □ KDDCup'07: 1st place for "… Decision Forests and …"
- □ KDDCup'08: 1st place of Challenge1 for a method using Bagging; 1st place of Challenge2 for "… Using an Ensemble Method "
- □ KDDCup'09: 1st place of Fast Track for "Ensemble … "; 2nd place of Fast Track for "… bagging … boosting tree models …", 1st place of Slow Track for "Boosting … "; 2nd place of Slow Track for "Stochastic Gradient Boosting"
- □ KDDCup'10: 1st place for "… Classifier ensembling"; 2nd place for "… Gradient Boosting machines … "
- □ KDDCup'11: 1st place of Track 1 for "A Linear Ensemble … "; 2nd place of Track 1 for "Collaborative filtering Ensemble", 1st place of Track 2 for "Ensemble …"; 2nd place of Track 2 for "Linear combination of …"
- □ KDDCup'12: 1st place of Track 1 for "Combining… Additive Forest…"; 1st place of Track 2 for "A Two-stage Ensemble of…"
- □ KDDCup'13: 1st place of Track 1 for "Weighted Average Ensemble" ; 2nd place of Track 1 for "Gradient Boosting Machine"; 1st place of Track 2 for "Ensemble the Predictions"
- □ KDDCup'14: 1st place for "ensemble of GBM, ExtraTrees, Random Forest…" and "the weighted average" ; 2nd place for "use both R and Python GBMs"; 3rd place for "gradient boosting machines… random forests" and "the weighted average of…"
- □ KDDCup'15: 1st place for "Three-Stage Ensemble and Feature Engineering for MOOC Dropout Prediction"
- □ KDDCup'16: 1st place for "Gradient Boosting Decision Tree"; 2nd place for "Ensemble of Different Models for Final Prediction"
- □ KDDCup'17: 1st and 2nd place of Task 1 for "XGBoost"; 1st place of Task 2 for "XGBoost", 2nd place of Task 2 for "Weighted Average of Multiple Models"
- □ KDDCup'18: 1st place for "Gradient Boosting"; 2nd place for "Two-stage stacking"; 3rd place for "Weighted Average of Multiple Models"

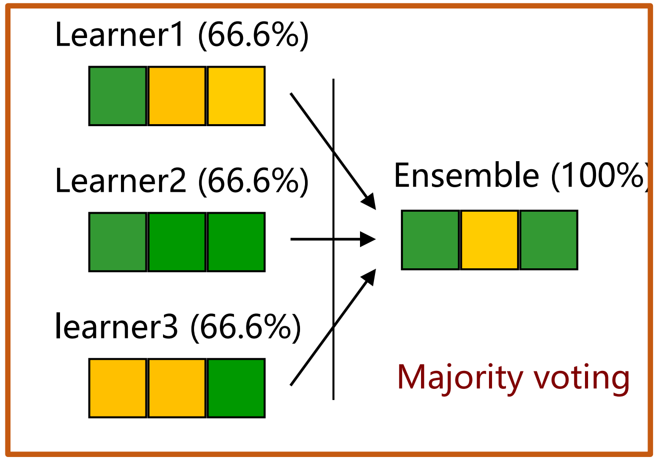
During the past decade, almost all winners of KDDCup, Netflix competition, Kaggle competitions, etc., utilized ensemble techniques in their solutions

To win? Ensemble !

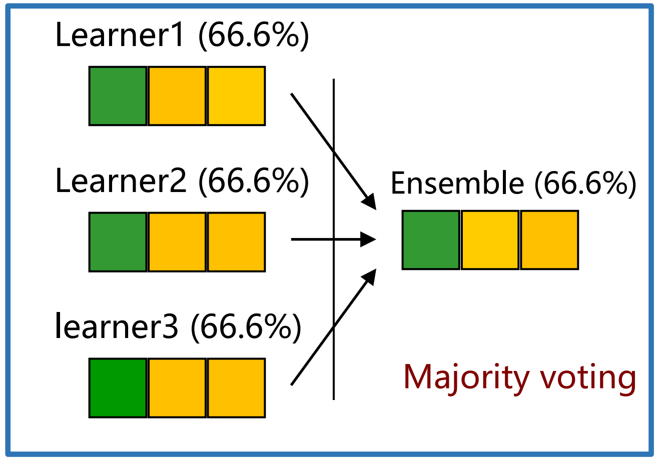
如何得到好的集成

Some intuitions:

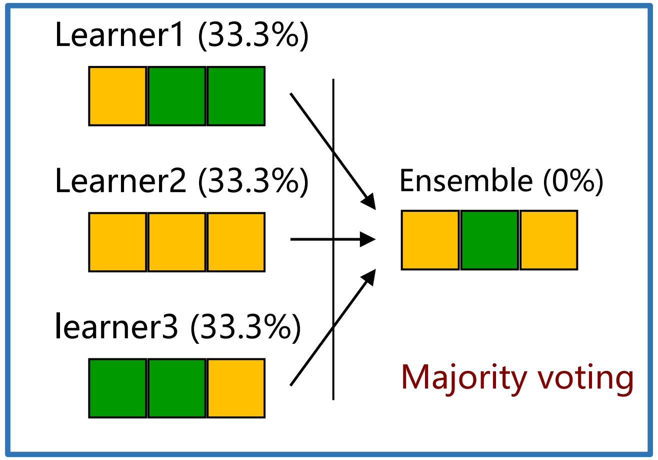
🟩🟨🟩 Ground-truth



Ensemble really helps



Individuals must be different

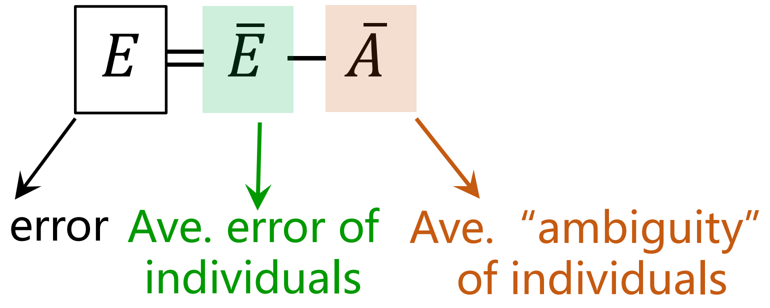


Individuals must be not-bad

令个体学习器“好而不同”

“多样性” （diversity）是关键

误差-分歧分解 (error-ambiguity decomposition):



$$\boxed{E} = \bar{E} - \bar{A}$$

Ensemble error

Ave. error of individuals

Ave. “ambiguity” of individuals

（“ambiguity” later called “diversity”）

The more **accurate** and **diverse** the individual learners, the better the ensemble

[Krogh and Vedelsby, NIPS95]

However,

- The “ambiguity” does not have an operable definition
- The error-ambiguity decomposition is derivable only for regression setting with squared loss

很多成功的集成学习方法

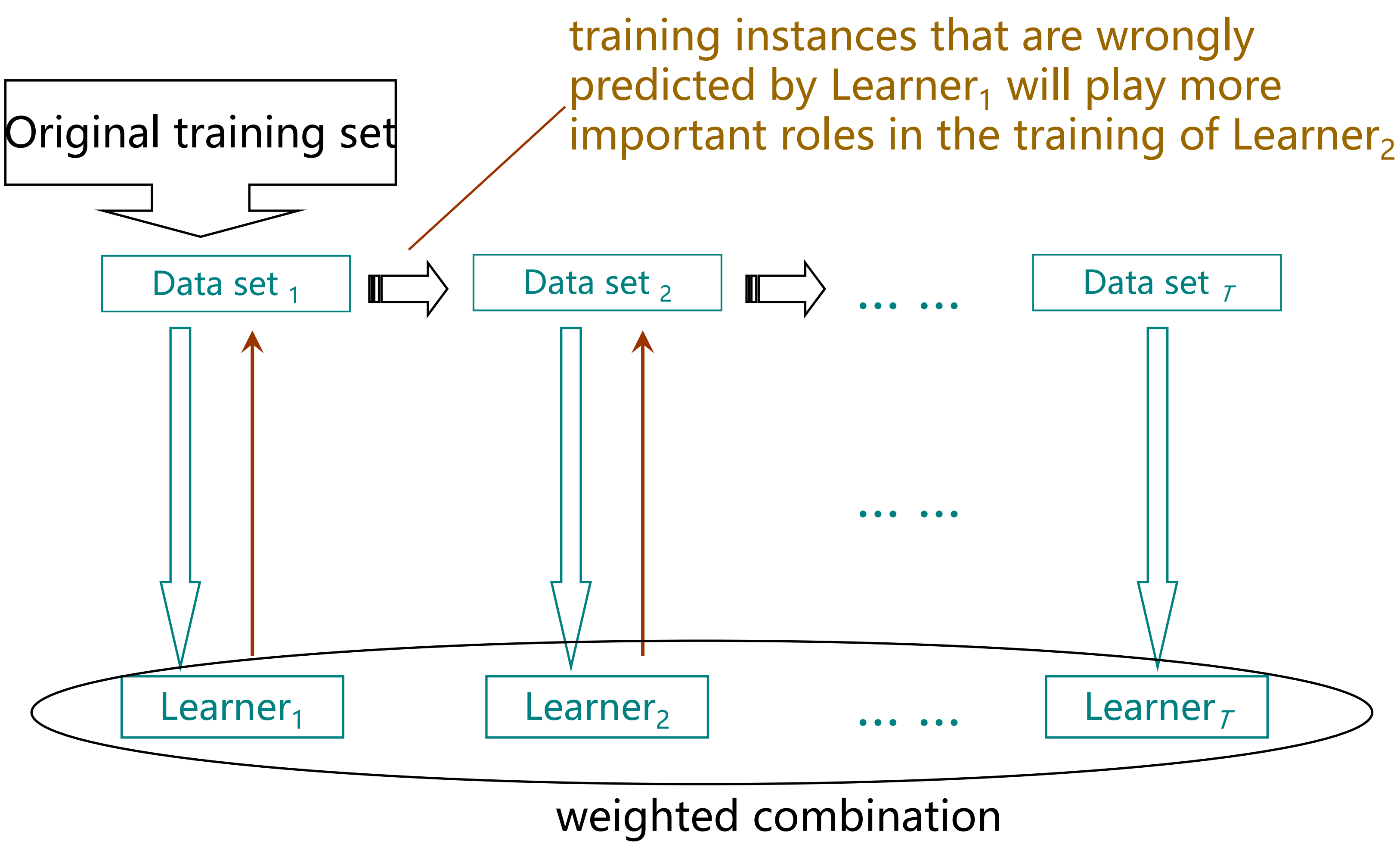
序列化方法

• **AdaBoost**	[Freund & Schapire, JCSS97]
• GradientBoost	[Friedman, AnnStat01]
• LPBoost	[Demiriz, Bennett, Shawe-Taylor, MLJ06]
• … …	

并行化方法

• **Bagging**	[Breiman, MLJ96]
• Random Forest	[Breiman, MLJ01]
• Random Subspace	[Ho, TPAMI98]
• … …	

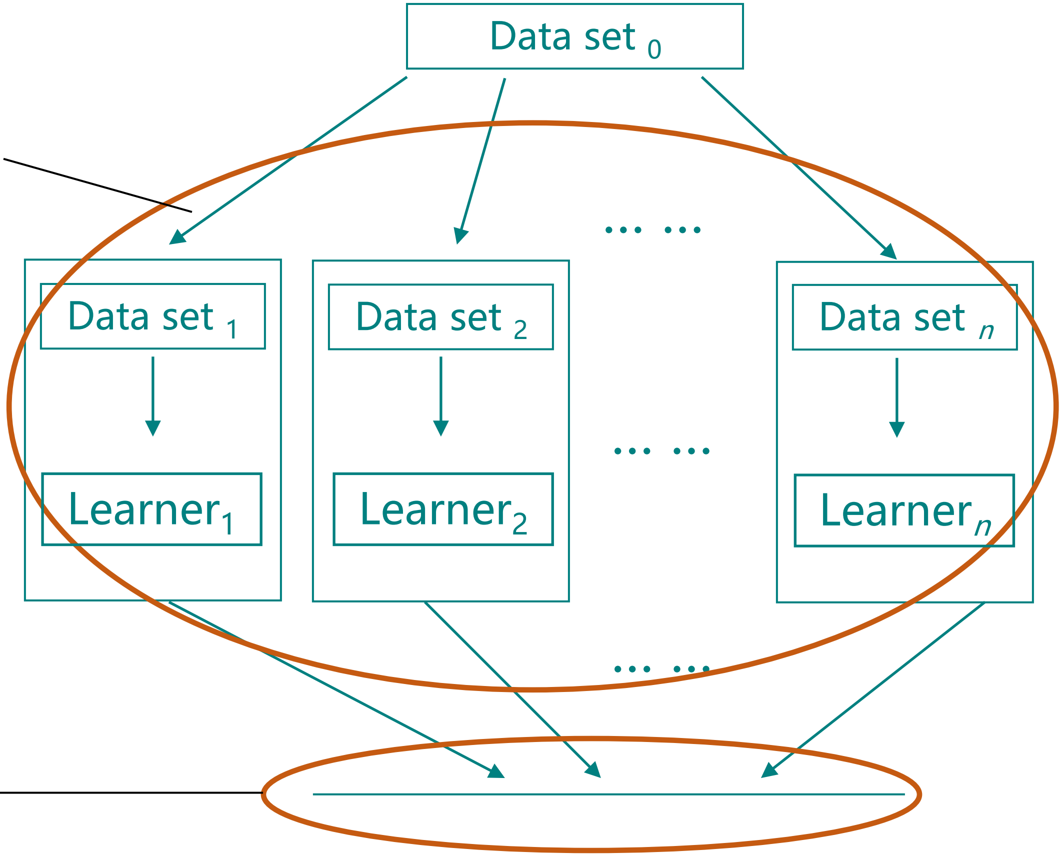
Boosting: A flowchart illustration



Bagging

bootstrap a set of learners

generate many data sets from the original data set through bootstrap sampling (random sampling with replacement), then train an individual learner per data set



voting for classification

the output is the class label receiving the most number of votes

averaging for regression

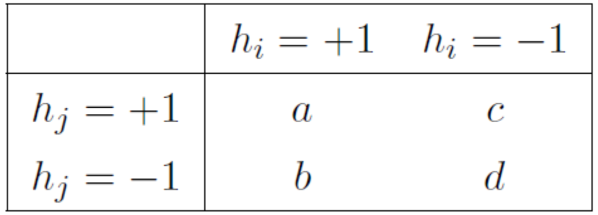
the output is the average output of the individual learners

多样性

“多样性”(diversity) 是集成学习的关键

多样性度量

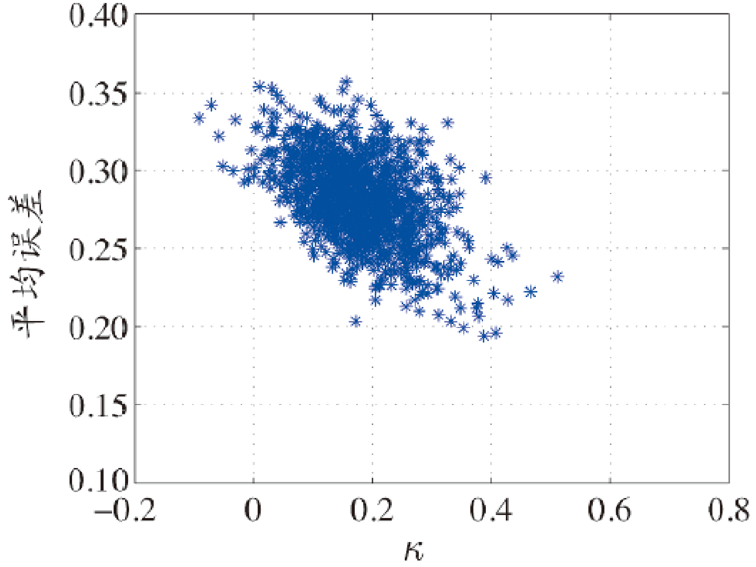
一般通过两分类器的预测结果列联表定义



	$h_i = +1$	$h_i = -1$
$h_j = +1$	a	c
$h_j = -1$	b	d

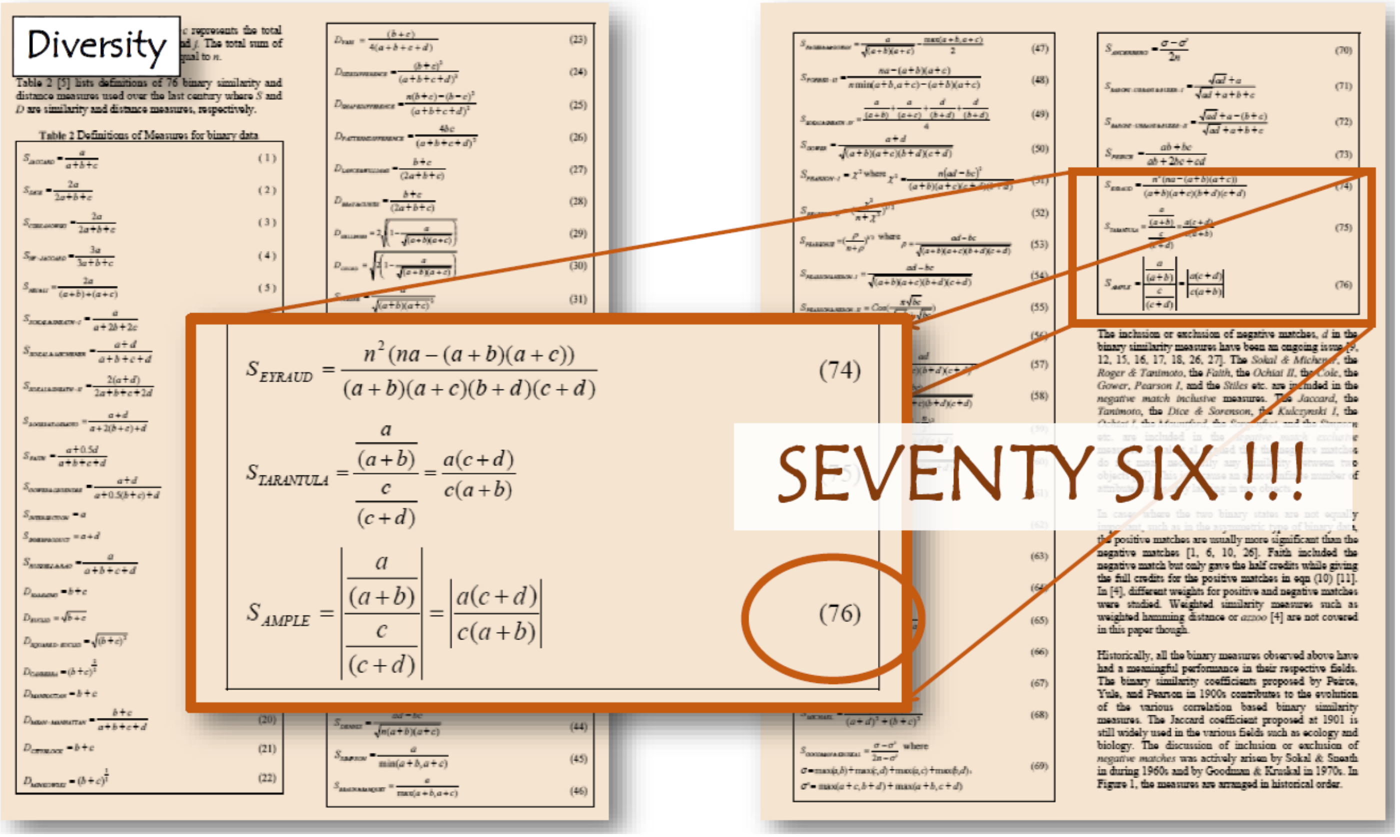
- 不合度量 (disagreement measure)
- 相关系数 (correlation coefficient)
- Q-统计量 (Q-statistic)
- κ-统计量 (κ-statistic)
- … …

κ-误差图



每一对分类器作为图中的一个点

研究者提出了很多 Diversity measure



Diversity

Table 2 [5] lists definitions of 76 binary similarity and distance measures used over the last century where S and D are similarity and distance measures, respectively.

Table 2 Definitions of Measures for binary data

$$S_{EYRAUD} = \frac{n^2(na-(a+b)(a+c))}{(a+b)(a+c)(b+d)(c+d)} \quad (74)$$

$$S_{TARANTULA} = \frac{\dfrac{a}{(a+b)}}{\dfrac{c}{(c+d)}} = \frac{a(c+d)}{c(a+b)} \quad (75)$$

$$S_{AMPLE} = \left|\frac{\dfrac{a}{(a+b)}}{\dfrac{c}{(c+d)}}\right| = \left|\frac{a(c+d)}{c(a+b)}\right| \quad (76)$$

SEVENTY SIX !!!

From [L. Kuncheva, ICPRAM’ 16 keynote]

However,

…

- □ [Kuncheva & Whitaker, MLJ 2003]: Empirical study shows that there seems no clear relation between many diversity measures and the ensemble performance
- □ [Tang, Suganthan, Yao, MLJ 2006]: Exploiting many diversity measures explicitly is ineffective in constructing consistently stronger ensembles

There is no well-accepted definition/formulation of diversity

" What is diversity " remains the holy grail problem of ensemble learning

九、聚类

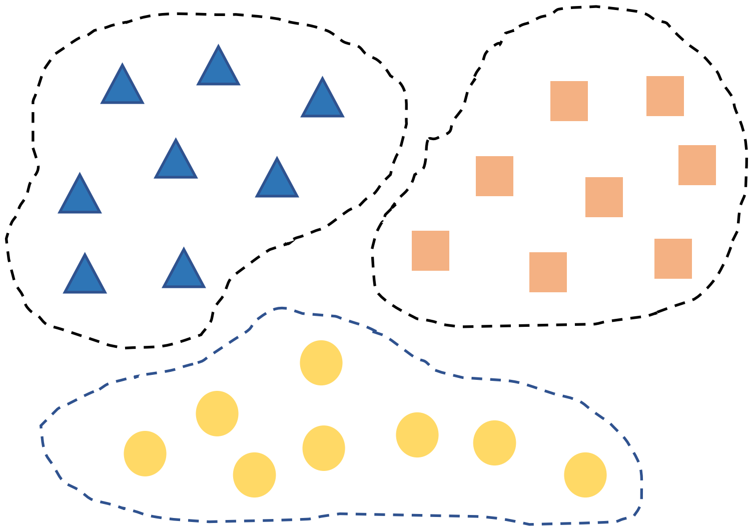
主讲教师：周志华

聚类 (Clustering)

在“无监督学习”任务中研究最多、应用最广

目标：将数据样本划分为若干个通常不相交的“簇”(cluster)

既可以作为一个单独过程（用于找寻数据内在的分布结构）
也可作为分类等其他学习任务的前驱过程



必须记住

聚类的“好坏”不存在绝对标准

The goodness of clustering depends on the opinion of the user.

故事一则

聚类的故事：

老师拿来苹果和梨，让小朋友分成两份。

小明把大苹果大梨放一起，小个头的放一起，老师点头，恩，体量感。

小芳把红苹果挑出来，剩下的放一起，老师点头，颜色感。

小武的结果？不明白。小武掏出眼镜：最新款，能看到水果里有几个籽，左边这堆单数，右边双数。

老师很高兴：新的聚类算法诞生了。

聚类也许是机器学习中“新算法”出现最多、最快的领域
总能找到一个新的“标准”，使以往算法对它无能为力

常见聚类方法

□ 原型聚类

- 亦称“基于原型的聚类”(prototype-based clustering)
- 假设：聚类结构能通过一组原型刻画
- 过程：先对原型初始化，然后对原型进行迭代更新求解
- 代表：k均值聚类，学习向量量化(LVQ)，高斯混合聚类

□ 密度聚类

- 亦称“基于密度的聚类”(density-based clustering)
- 假设：聚类结构能通过样本分布的紧密程度确定
- 过程：从样本密度的角度来考察样本之间的可连接性，并基于可连接样本不断扩展聚类簇
- 代表：DBSCAN, OPTICS, DENCLUE

□ 层次聚类 (hierarchical clustering)

- 假设：能够产生不同粒度的聚类结果
- 过程：在不同层次对数据集进行划分，从而形成树形的聚类结构
- 代表：AGNES (自底向上)，DIANA (自顶向下)