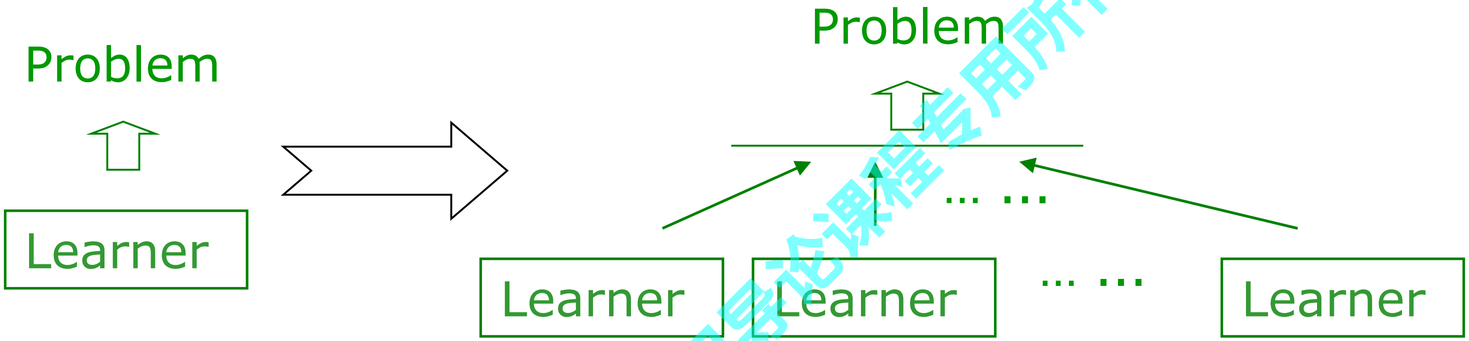
机器学习导论（2017 春季学期）

八、集成学习

主讲教师：周志华

集成学习 (Ensemble learning)

集成学习通过构建并结合多个学习器来完成学习任务



- □ 同质(homogeneous)集成：集成中只包含同种类型的“个体学习器”

  相应的学习算法称为“基学习算法” (base learning algorithm)

  个体学习器亦称“基学习器” (base learner)

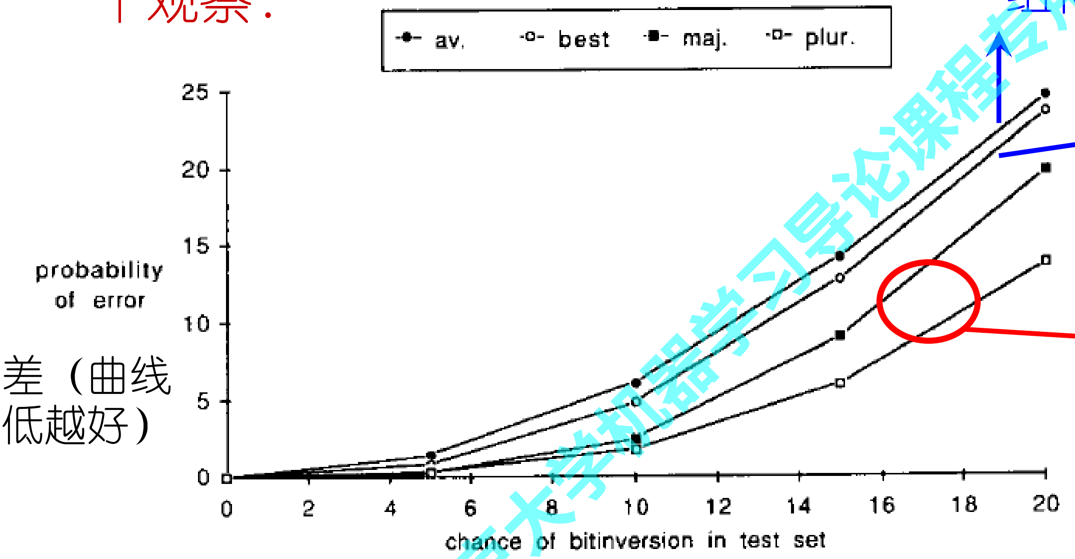
- □ 异质(heterogeneous)集成：个体学习器由不同的学习算法生成

  不存在“基学习算法”

Why Ensemble?

集成的泛化性能通常显著优于单个学习器的泛化性能

一个观察：



一组神经网络的平均性能

选择最优神经网络

两种简单的神经网络集成方法

误差（曲线越低越好）

Fig. 4. Performance versus noise level in the test set is shown for individual and for consensus decisions. Data displayed shows the average and the best network, as well as collective decisions using majority and plurality for seven networks trained on individual training sets.

[Hansen & Salamon, TPAMI90]

集成学习的巨大成功

例如，著名的数据挖掘竞赛 KDDCup：

- □ 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 … ”

集成学习的巨大成功 (con’t)

- □ 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…”

集成学习的巨大成功 (con’t)

□ 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”

□ Kaggle competitions

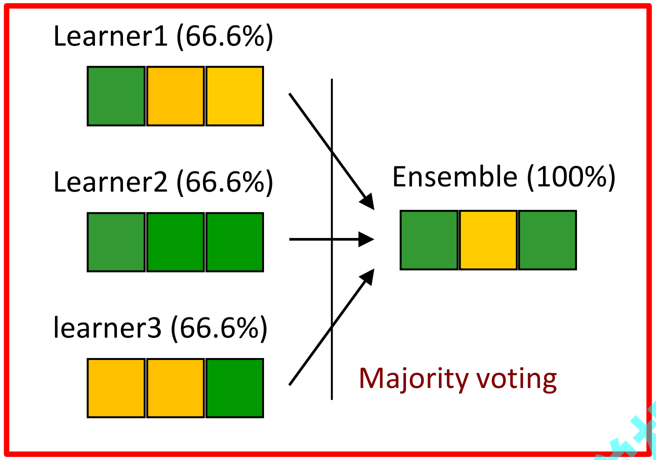
想获胜，用集成

现实各类机器学习、数据挖掘应用中，广泛使用集成学习技术

如何得到好的集成？

Some intuitions:

🟩🟨🟩 Ground-truth


Learner1 (66.6%) 🟩🟨🟨

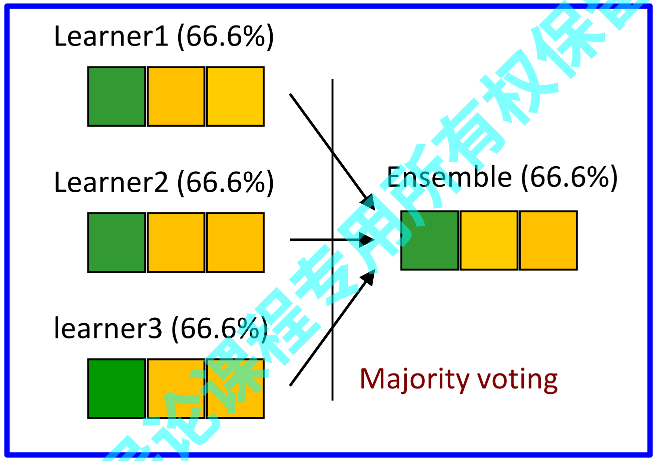
Learner2 (66.6%) 🟩🟩🟩

learner3 (66.6%) 🟨🟨🟩

Ensemble (100%) 🟩🟨🟩

Majority voting

Ensemble really helps


Learner1 (66.6%) 🟩🟨🟨

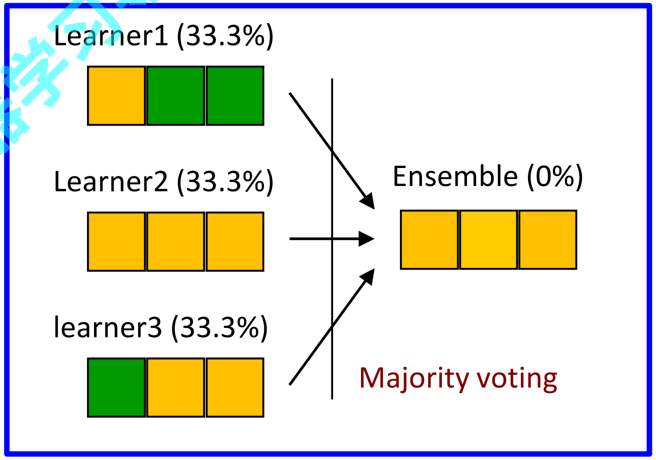
Learner2 (66.6%) 🟩🟨🟨

learner3 (66.6%) 🟩🟨🟨

Ensemble (66.6%) 🟩🟨🟨

Majority voting

Individuals must be different


Learner1 (33.3%) 🟨🟩🟩

Learner2 (33.3%) 🟨🟨🟨

learner3 (33.3%) 🟩🟨🟨

Ensemble (0%) 🟨🟨🟨

Majority voting

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

However,

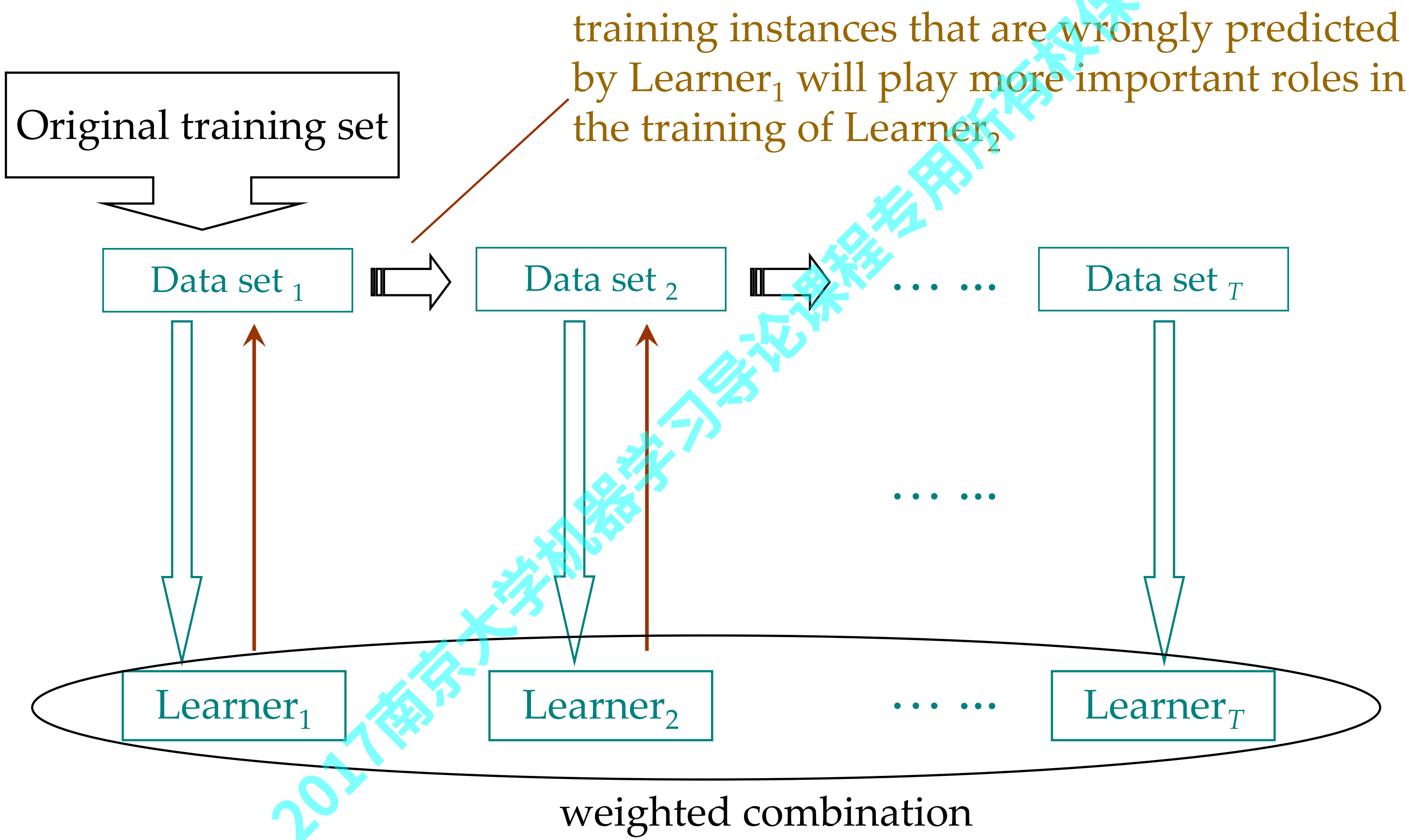
- the “ambiguity” does not have an operable definition
- The error-ambiguity decomposition is derivable only for regression setting with squared loss

很多成功的集成学习方法

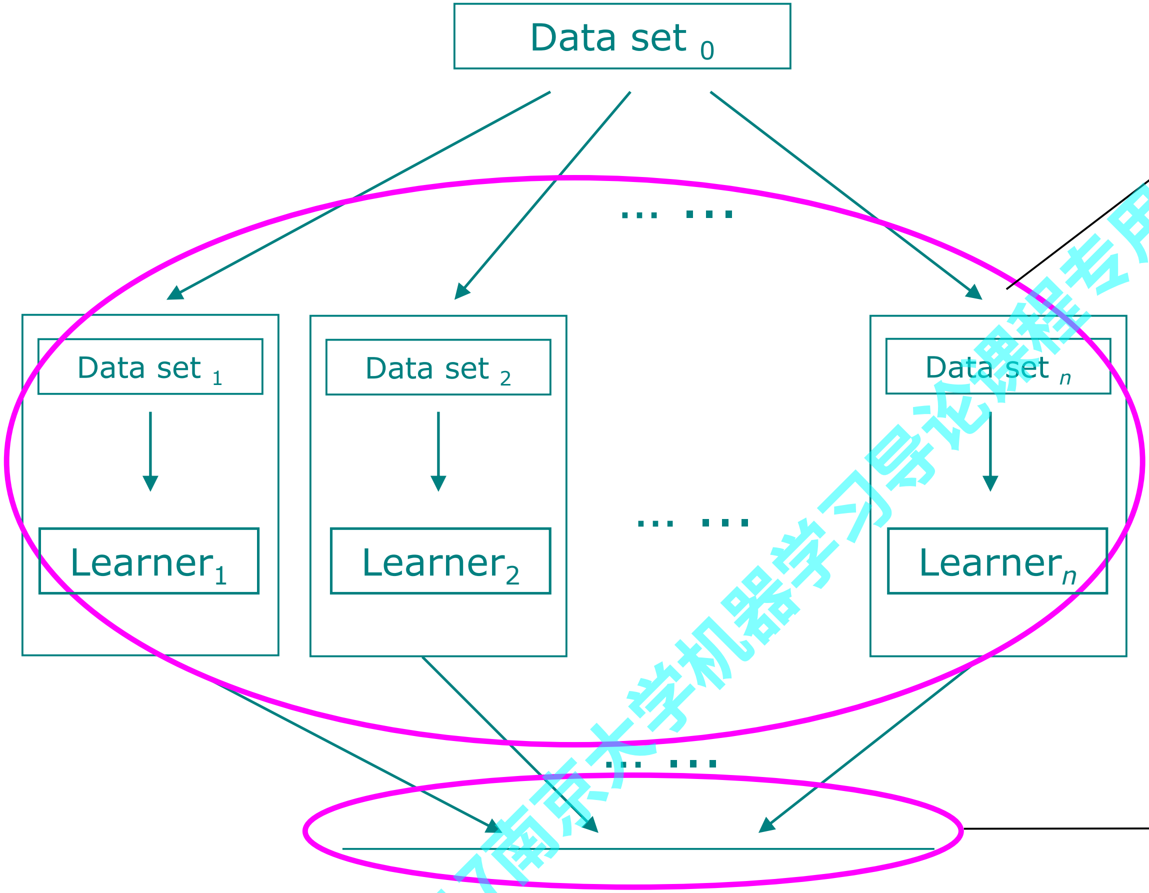
- 序列化方法
  - **AdaBoost** [Freund & Schapire, JCSS97]
  - GradientBoost [Friedman, AnnStat01]
  - LPBoost [Demiriz, Bennett, Shawe-Taylor, MLJ06]
  - … …

- 并行化方法
  - **Bagging** [Breiman, MLJ96]
  - Random Forest [Breiman, MLJ01]
  - Random Subspace [Ho, TPAMI98]
  - … …

Boosting: A flowchart illustration


training instances that are wrongly predicted by Learner$_1$ will play more important roles in the training of Learner$_2$
Original training set
Data set $_1$
Data set $_2$
… …
Data set $_T$
… …
Learner$_1$
Learner$_2$
… …
Learner$_T$
weighted combination

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

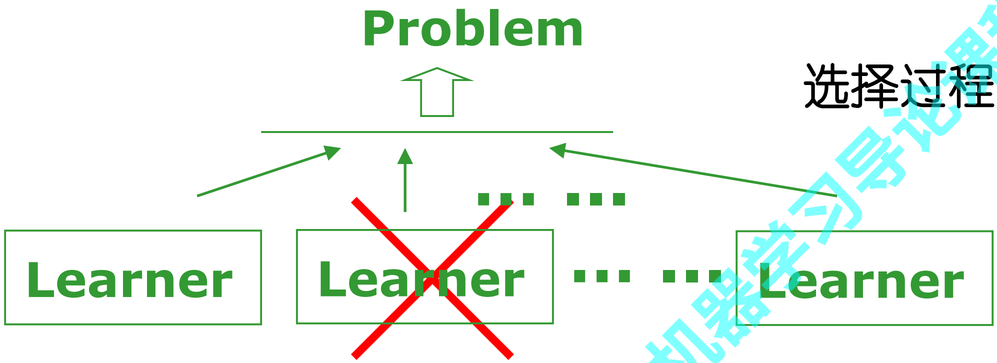
averaging for regression

the output is the average output of the individual learners

“越多越好”？

选择性集成 (selective ensemble):

给定一组个体学习器，从中选择一部分来构建集成，经常会比使用所有个体学习器更好 (更小的存储/时间开销，更强的泛化性能)



选择过程需考虑个体 性能 与 多样性/互补性

仅选出“精度最高的”通常不好！

集成修剪 (ensemble pruning) [Margineantu & Dietterich, ICML’97] 较早出现，针对序列型集成
减小集成规模、降低泛化性能

选择性集成 [Zhou, et al, AIJ 02] 稍晚，针对并行型集成， MCBTA (Many could be better than all)定理
减小集成规模、增强泛化性能

目前“集成修剪”与“选择性集成”基本被视为同义词

更多关于集成学习的内容，可参考：

Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series
Ensemble Methods
Foundations and Algorithms
Zhi-Hua Zhou
CRC Press

Z.-H. Zhou.
Ensemble Methods:
Foundations and Algorithms,
Boca Raton, FL: Chapman &
Hall/CRC, Jun. 2012.
(ISBN 978-1-439-830031)

集成学习常用软件/工具包

- □Random Forest

  https://cran.r-project.org/web/packages/randomForest/index.html

- □Boosting (LightGBM, Light Gradient Boosting Machine including GBDT, GBRT, GBM/MART)

  https://github.com/Microsoft/LightGBM

- □Boosting (multi-class / multi-label / multi-task classification)

  http://www.multiboost.org/

- □… …

前往……