

Lecture 8: 集成学习

什么是集成模型

引入: 弱可学习和强学习是否可以等价?

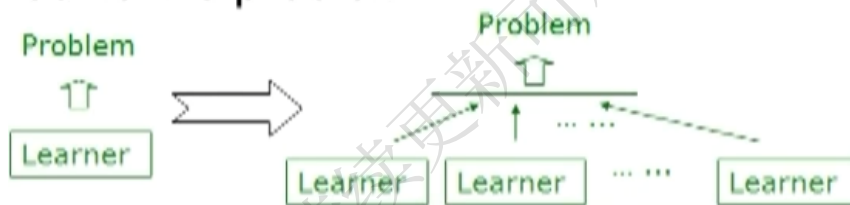
- 强可学习:
 - 至少存在一个算法, 它在多项式时间里面可以为我们训练出一个模型, 这个模型的输出和实际输出的不一致的可能性十分的小;
- 弱可学习:
 - 学习的效果只比随机猜好一点;
- 这个问题的含义:
 - 是否可以使用多个弱的学习的算法, 构成一个强学习算法+;

概念: 集成模型

- 概念:
 - 原本使用一个模型解决的问题, 现在可以用多个模型来解决;

Ensemble Learning (集成学习):

Using multiple learners to solve the problem

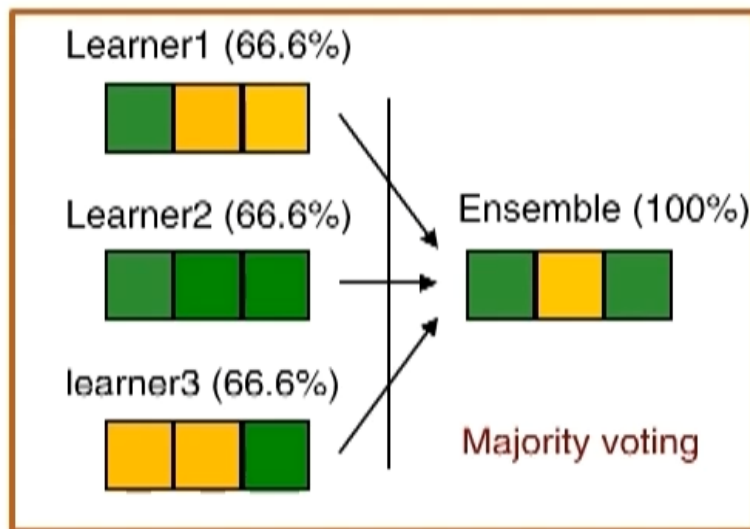


- 也可以每一个都是相同的模型, 此时称之为同质模型 (homo)
- 如果模型不一样, 则称之为异质模型;
- 集成学习是将模型内部多个弱监督模型组成, 每个模型只在某个方向表现比较好;
- 举例:
 - 目标:

Some intuitions:

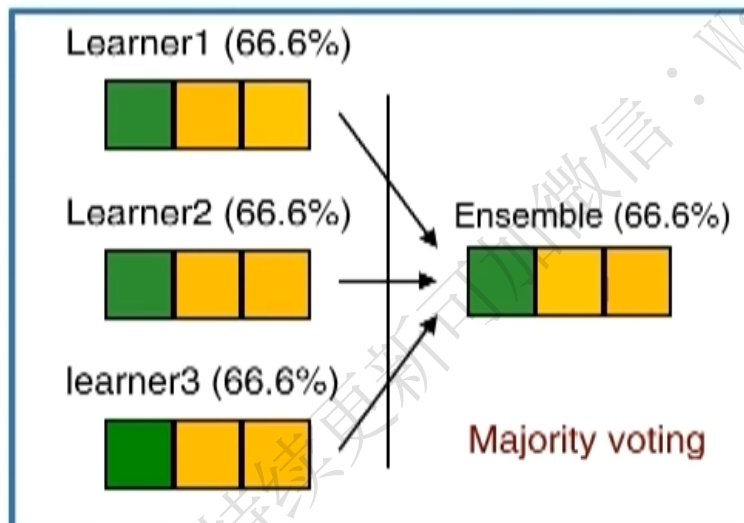


- 三个学习器, 集成学习:
 - 每一个个体都是 66.6%
 - 进行多数投票;



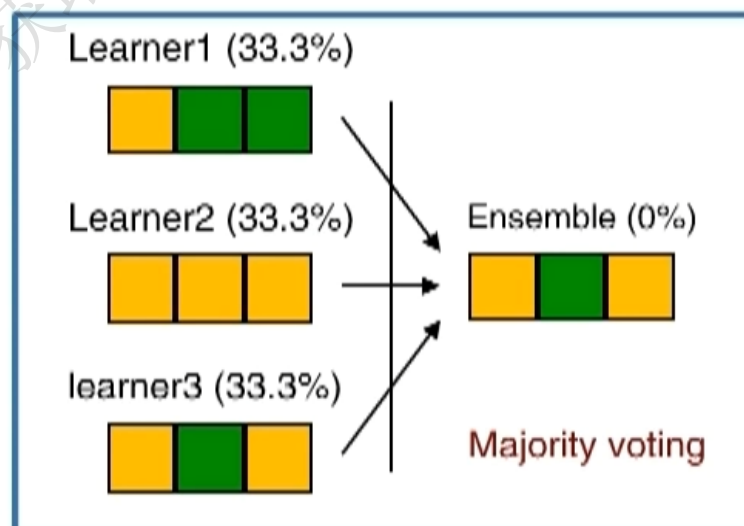
Ensemble really helps

- 也是集成学习:



Individuals
must be
different

- 如果集成的方法要其效果，需要每一个继承的个体不一样；
- 也是集成学习:
- 继承的个体必须得不能太差；



Individuals
must be
not-bad

- 总结:

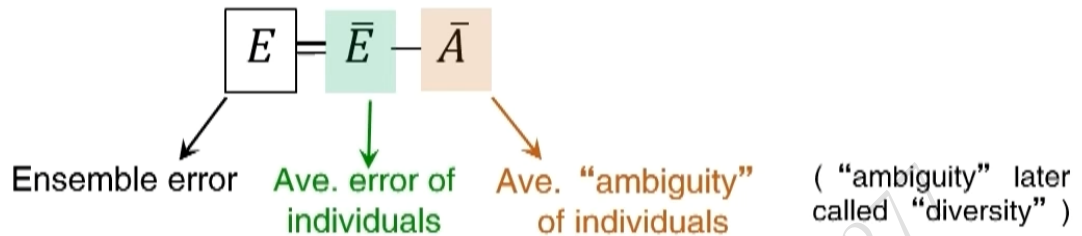
- 需要另每一个个体学习器好而不同；

概念：误差-分歧分解

概念：

- Ensemble error
- Ave. error of individuals

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



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

[Krogh and Vedelsby, NIPS95]

- 做了一个集成模型，首先希望每一个个体的误差小（E 拔），并且希望差异比较大（A 拔）
- 越精确，个体差异越大，效果越好；
- 然而，误差并不能精确的定义出来；

两类集成学习方法

概念：分类

图示：

■ 序列化方法

- AdaBoost
- GradientBoost
- LPBoost
-

[Freund & Schapire, JCSS97]

[Friedman, AnnStat01]

[Demiriz, Bennett, Shawe-Taylor, MLJ06]

■ 并行化方法

- Bagging
- Random Forest
- Random Subspace
-

[Breiman, MLJ96]

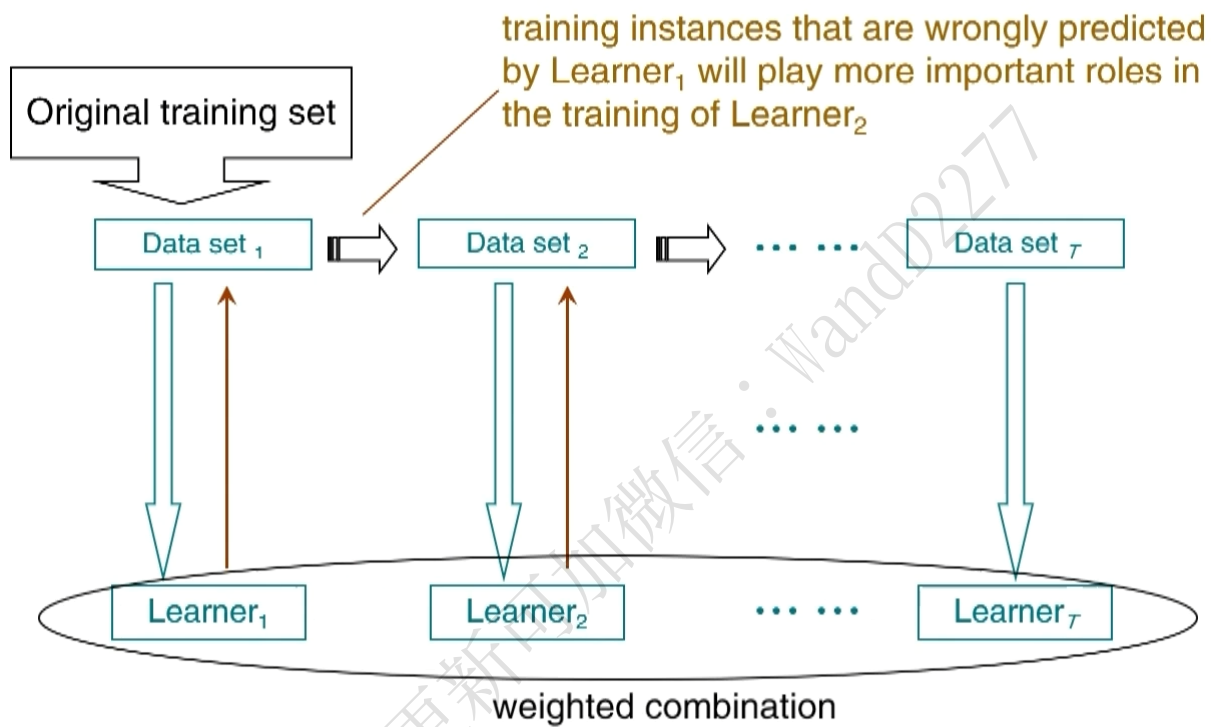
[Breiman, MLJ01]

[Ho, TPAMI98]

序列化方法 - Boosting

概念：Boosting

- 概念: **base** - 基
 - 同质的集成, 让它好很简单, 使用模型选择即可, 但是让它们不同很难;
 - 异质的集成, 让它们不同很简单, 但是让不同的模型直接进行比较是很困难的, 两个不同算法输出不能直接比较, 不同模型进行比较时、必须要进行一个配准 (**Alignment**) 的过程, 而这件事情十分困难;
 - 集成学习里面主要关注同质的;
 - 假设一个算法输入很多数据, 产生很多个模型, 这些模型都是同一种算法生成的, 此时就称这个算法为基学习算法;
- 概念:

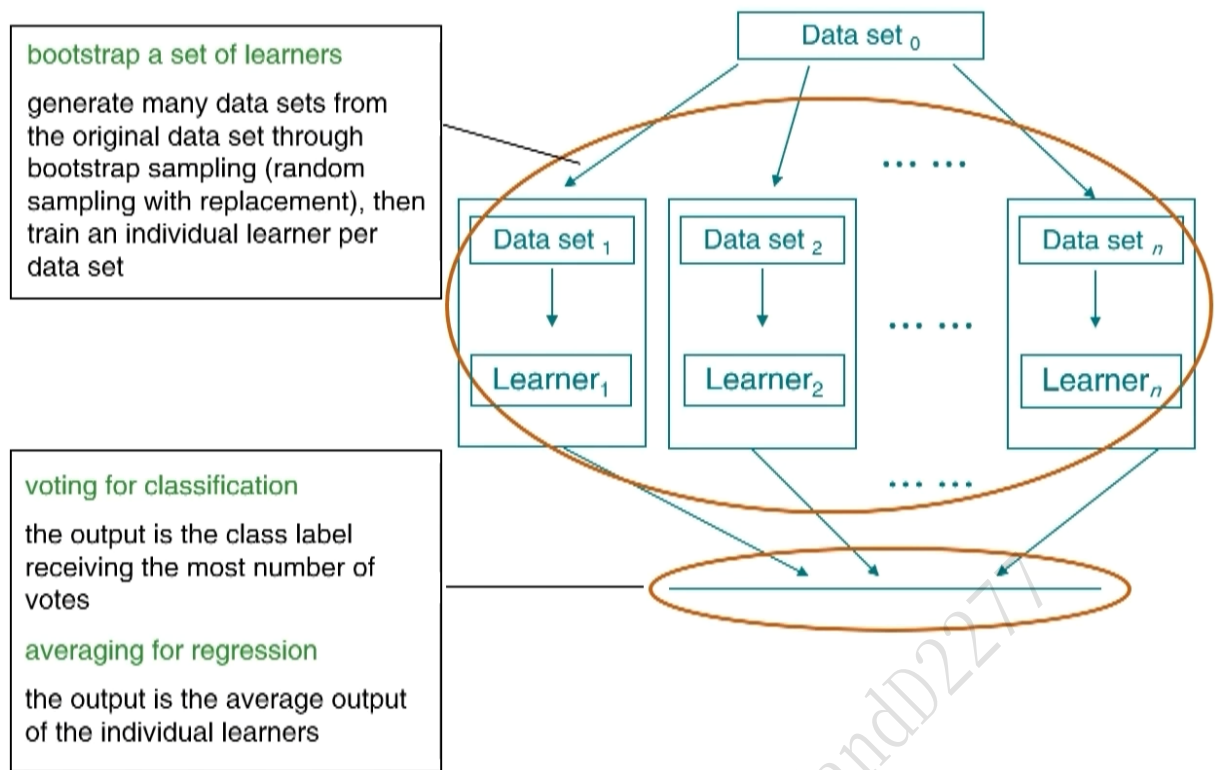


-
- 对原本的数据集进行训练得到学习器 1, 然后使用学习器 1、对数据集 1 中的数据进行判断, 分别给予其权值, 使得其构成数据集 2, 数据集 2 中的不同数据用于不同的权值, 对其进行训练、产生第二种学习器.....;

并行化方法 - Bagging

概念: **Bagging**

- 概念:
 - 通过一个数据集, 产生很多和学习器;



- 通过数据集 0 产生 n 个数据集的时候，用一个 **bootstrap** 可重复采样，随机可重复采样的选择时，可能会有 **63.7%** 的数据选择，剩下的一直都选不中，因此 n 个数据集都不太一样；
- 在不同的数据集基础上做学习器，然后把它们结合起来；
- 投票做分类，平均做回归；

多样性度量

概念：多样性

- 概念：
 - 预测一样的是 a 和 d ，预测不一样的是 c 和 b ；
 - 用 $c+b/a+d$ 或者 $(c+b)-(a+d)$ 就可以进行度量多样性；

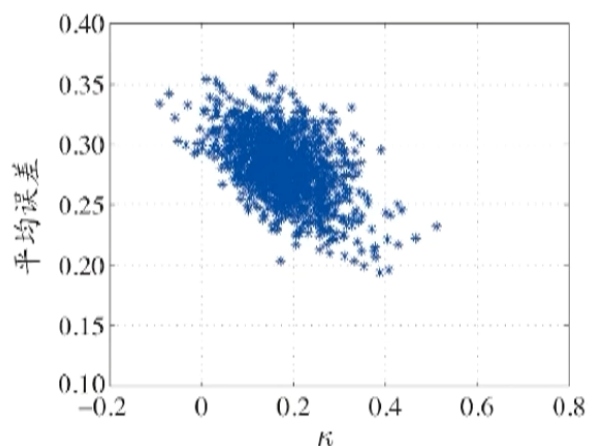
多样性度量

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

随机森林

概念：随机森林

- 概念：
 - 森林 - 指的是模型当中包含了很多的决策树，从而将这个包含了很多决策树的模型看成一个森林；
 - 随机 - 指的是森林中的每个决策树都随机的采样数据集中的一小部分进行训练，以保证每个决策树的输出相似但是不一样，从而在最好将它们整合在一起的时候可以作为一个输出；
- 过程：
 - (1) 预设模型的超参数，即用于几个树？每个树最多几层深度？
 - (2) 随机采样，训练每个决策树；
 - (3) 当每个决策树都完成了训练，将最后每个树的结果整合在一起；
 - 对于回归的问题，通常采用均值的方法将所有结果整合；
 - 对于分类的问题，通过采用求众数的方法输出最多的类别；
- 特点：
 - 优点：
 - (1) 模型随机性强，不易 `overfit`
 - 抗噪性强，表示对异常点 `outlier` 不敏感
 - (2) 处理高维数据相对更快；
 - (3) 树状结构，模型可解释度高，可以告诉你每个特征的重要性；
 - 缺点：
 - 模型往往过于 `General`，不具备正确处理过于困难的样本的能力；
 - 即：模型的起点高，但是天花板低；

Adaboost

Adaboost 基础概念

概念：Adaboost

- 概念：
 - **Adaboost**： `Adaptive Boosting` - 自适应增强算法
 - 先训练一个弱学习模型，再对这个模型进行评估，对于这个 `Weak Learner` 中做对的问题，将减少注意力。而对于它做错的问题，将增大它的注意力，进而在后续的模型中，更专注于克服前面的模型不能解决的困难点；
 - 最后，将所有的模型整合在一起，构成一个大的框架，大框架中有处理简单问题的模型，也有处理困难问题的模型，使得大框架的整体性能有所提高；
- 分析：
 - 相比于随机森林中的各个模型相互独立，`AdaBosst` 中的后一个模型永远是建立在前一个模型的输出的基础上，构成顺序，级联的结构；
- 特点：
 - (1) 顺序，级联的结构 + 利用权重组合结果

- 模型的天花板高;
- (2) 更容易 **Overfit**, 并且对于异常点过于敏感, 拉低了模型的性能低点;
- (3) 训练的速度更慢;
- (4) 非常的简单;
- 算法:

输入: 训练集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$;
 基学习算法 $\mathcal{L}$;
 训练轮数 T .

过程:

```

1:  $\mathcal{D}_1(\mathbf{x}) = 1/m$ .
2: for  $t = 1, 2, \dots, T$  do
3:    $h_t = \mathcal{L}(D, \mathcal{D}_t)$ ;
4:    $\epsilon_t = P_{\mathbf{x} \sim \mathcal{D}_t}(h_t(\mathbf{x}) \neq f(\mathbf{x}))$ ;
5:   if  $\epsilon_t > 0.5$  then break
6:    $\alpha_t = \frac{1}{2} \ln \left( \frac{1-\epsilon_t}{\epsilon_t} \right)$ ;
7:    $\mathcal{D}_{t+1}(\mathbf{x}) = \frac{\mathcal{D}_t(\mathbf{x})}{Z_t} \times \begin{cases} \exp(-\alpha_t), & \text{if } h_t(\mathbf{x}) = f(\mathbf{x}) \\ \exp(\alpha_t), & \text{if } h_t(\mathbf{x}) \neq f(\mathbf{x}) \end{cases}$ 
    $= \frac{\mathcal{D}_t(\mathbf{x}) \exp(-\alpha_t f(\mathbf{x}) h_t(\mathbf{x}))}{Z_t}$ 

```

8: **end for**

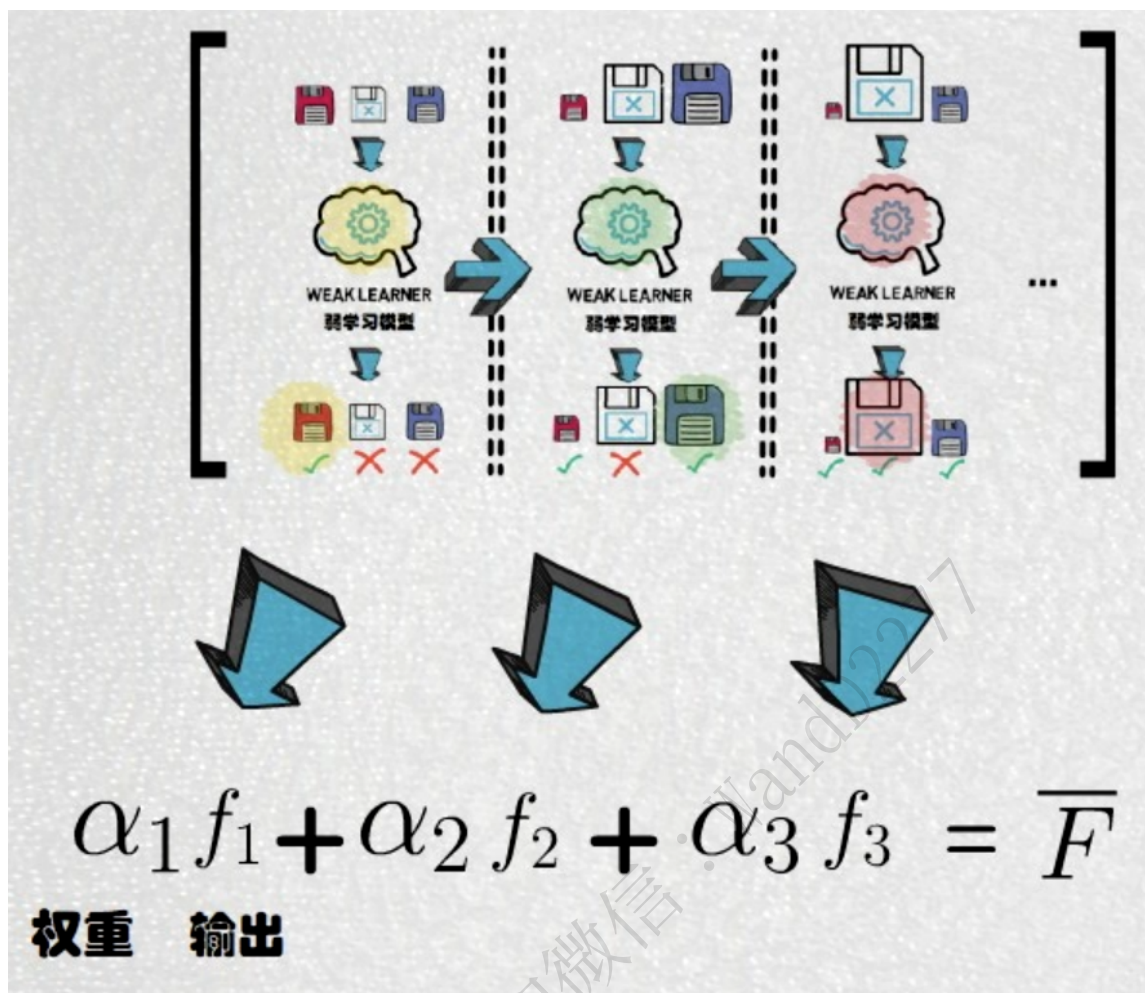
输出: $H(\mathbf{x}) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(\mathbf{x}) \right)$

- 实现:
 - (1) 对于训练每个 **weak learner** 计算样本困难度:
 - 1. 初始化所有样本的困难度为 $w_i = 1/N$
 - 2. 利用当前 **weak learner** 的训练结果, 更新所有样本的困难度:
 - 对于做对的样本, 减少它困难度

$$w_i^{new} = \begin{cases} \frac{1}{2(1-\epsilon)} w_i^{old} \\ \frac{1}{2\epsilon} w_i^{old} \end{cases}$$

ϵ

- 3. 基于当前样本困难度, 训练下一个 **learner**
- (2) 学习每个 **Weak learner** 的权重, 将它们整合

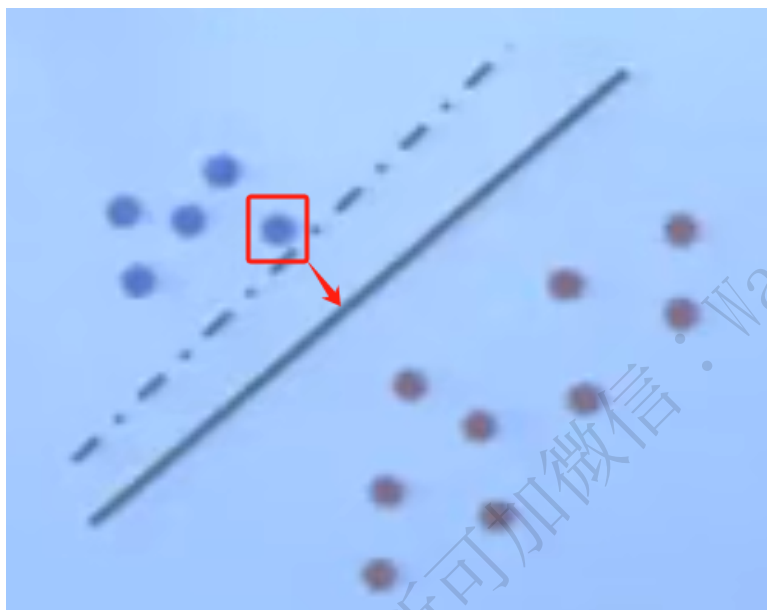


Adaboost 为什么不过拟合

分析: Adaboost 为什么不过拟合

- 链接:
 - [2 围绕adaboost为何不overfitting_哔哩哔哩_bilibili](#)
- 两类观点:
 - 间隔理论;
 - 很多都是计算机科学家;
 - 统计观点;
 - 很多都是统计学家;
 - 它们的观点很多都互相不接受, 统计学家认为没有一个统计量能够称之为“间隔”;
- 概念: 统计学家的观点
 - 你要做一个二分类的问题, 比如 0/1-loss
 - 如果预测给一个正的值, 就是 0, 如果是错的值, 就给你一个 1;
 - 但直接优化 0-1 函数很难, 不平滑, 因此会使用一个优化函数;
 - 因此统计学家认为, 这些计算机科学家就是: 第一, 找一个目标函数, 第二, 找一个方法去优化它;
 - 因此统计学家认为, Adaboost 的关键一步, 就是对权值的调整, 这个过程和指数函数很像, 基于这个进行推理;
 - 推理下来, 假设使用一个对数函数, 并估计它的概率, Adaboost 就是在优化这个指数函数, 只不过是用一个牛顿法的方式来一步步的迭代;

- 因此，只要换一个方法来进行优化，就可以得到其他的 `Boost` 算法；
- 但是，从统计学上推理得到的结果，和实际上并不是完全一致的；
- 并且，统计学家完全没有解释：为什么 `Adaboost` 不会 `Overfit`；
- 概念：间隔理论
 - 间隔：
 - 只要是一个分类问题，就可以理解为是找一个超平面，将两类点分隔开，而间隔就是划分超平面的线和其他点的距离；
 - 如果间隔越大，则划分的把握就越大；
 - 随着模型训练轮次的增加，即使模型的训练误差已经降低到 0 了，继续训练下去、它的间隔是可以提高的；



- 间隔提高了，就会使得模型的预测性能能够提高；
- 而且最小间隔的点最重要，然后可以对最小间隔最大化设计算法，但设计出来的结果依然比 `Adaboost` 差；
- 原因是其没有对模型的复杂度控制好，他们只是对决策树的树叶控制了，但是 `arc-gv` 生成出来的树比 `Adaboost` 的稍微深一点；
- 这次不使用决策树，换了一种树。基于这种树做出的结果，得出应该是平均的间隔重要；
- 后背用界来描述；
- 历史：

1989, [Kearns & Valiant], open problem

1990, [Schapire], proof by construction, the first Boosting algorithm

1993, [Freund], another impractical boosting algorithm by voting

1995/97, [Freund & Schapire], AdaBoost

1998, [Schapire, Freund, Bartlett & Lee], Margin theory

1999, [Breiman], serious doubt by minimum margin bound

2006, [Reyzin & Schapire], finding the model complexity issue in exps, emphasizing the importance of margin distribution

2008, [Wang, Sugiyama, Yang, Zhou & Feng], Emargin bound, believed to be a margin distribution bound

2013, [Gao & Zhou], a real margin distribution bound, shedding new insight ; margin theory defended

•

获取笔记持续更新可加微信: WandD2271