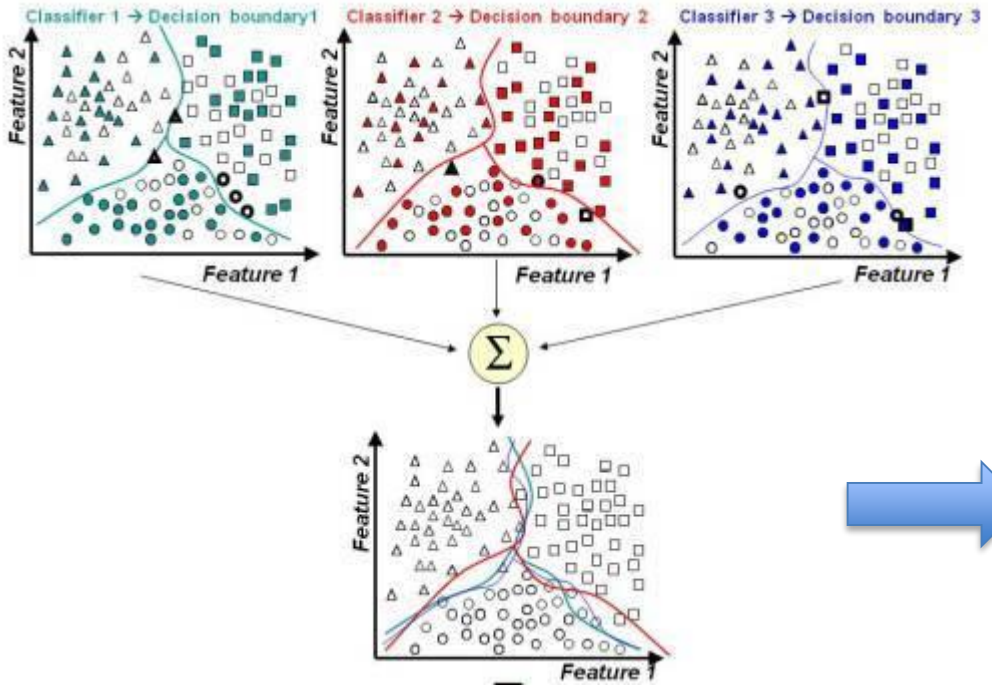
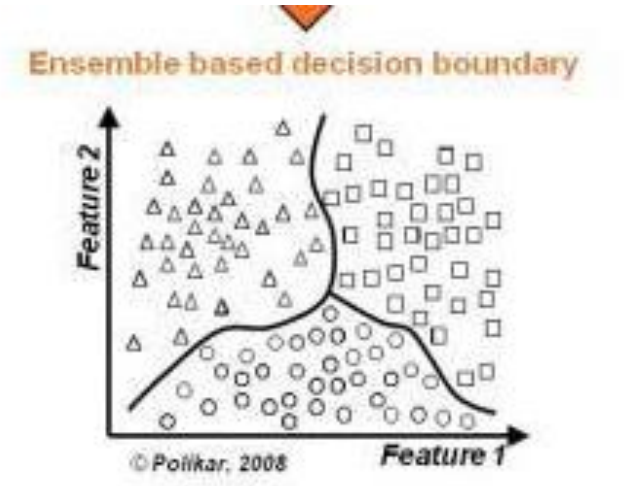
集成学习
Ensemble Learning

陈昱夫

集成学习

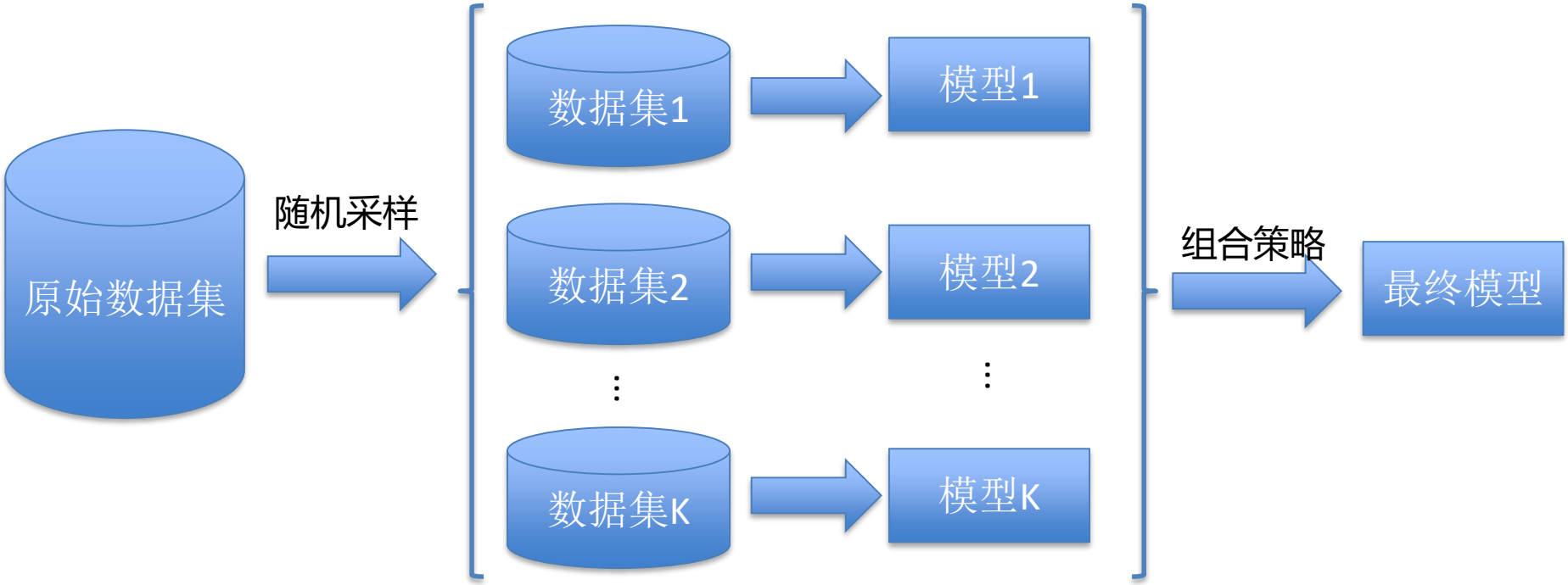
- 通过某种策略将多个模型集成起来，通过群体决策来提高决策准确率。





Bagging类方法

- 通过随机构造训练样本、随机选择特征等方法来提高训练数据集的独立性，从而提高每个弱模型的独立性



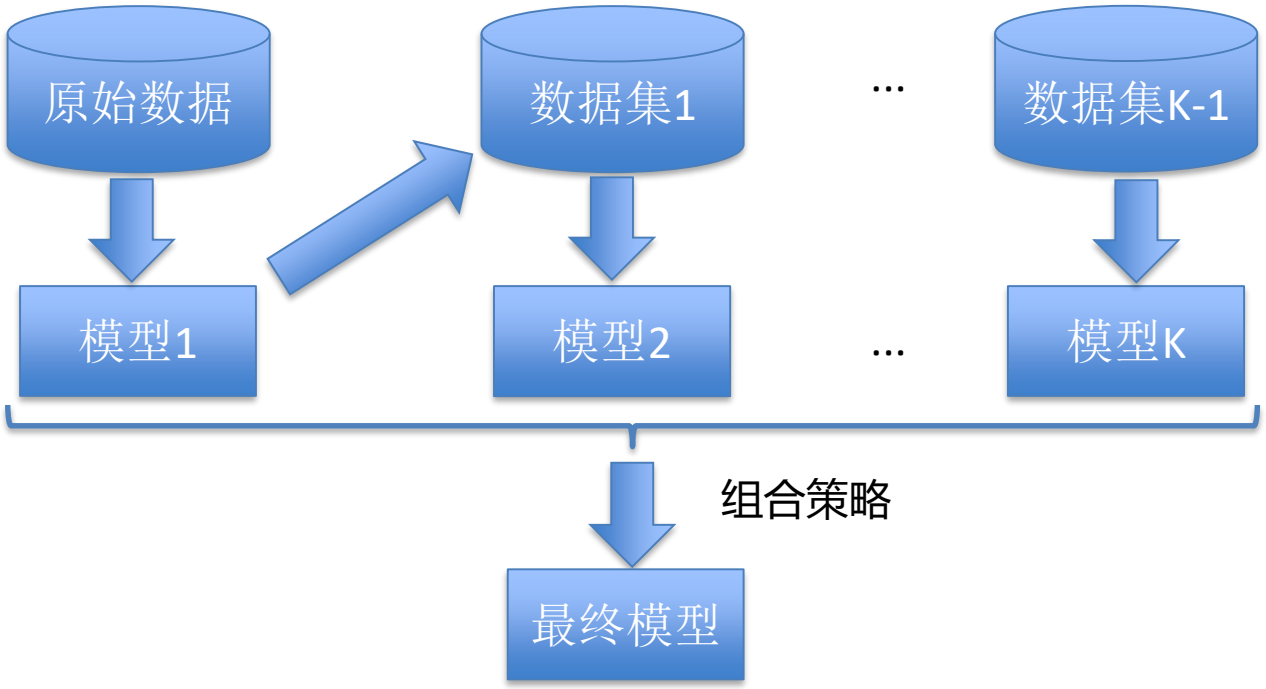
随机森林

- 随机采样：
  - 1. 对于每棵树而言，随机且有放回地从训练集中的抽取若干个训练样本（bootstrap sample），作为该树的训练集。
  - 2. 随机地从该数据集所有特征中选取一个特征子集，每次选取节点时，从这个特征子集中选择最优特征。

- 组合策略
  - 分类：所有决策树进行多数投票
  - 回归：所有决策树预测的回归值的均值

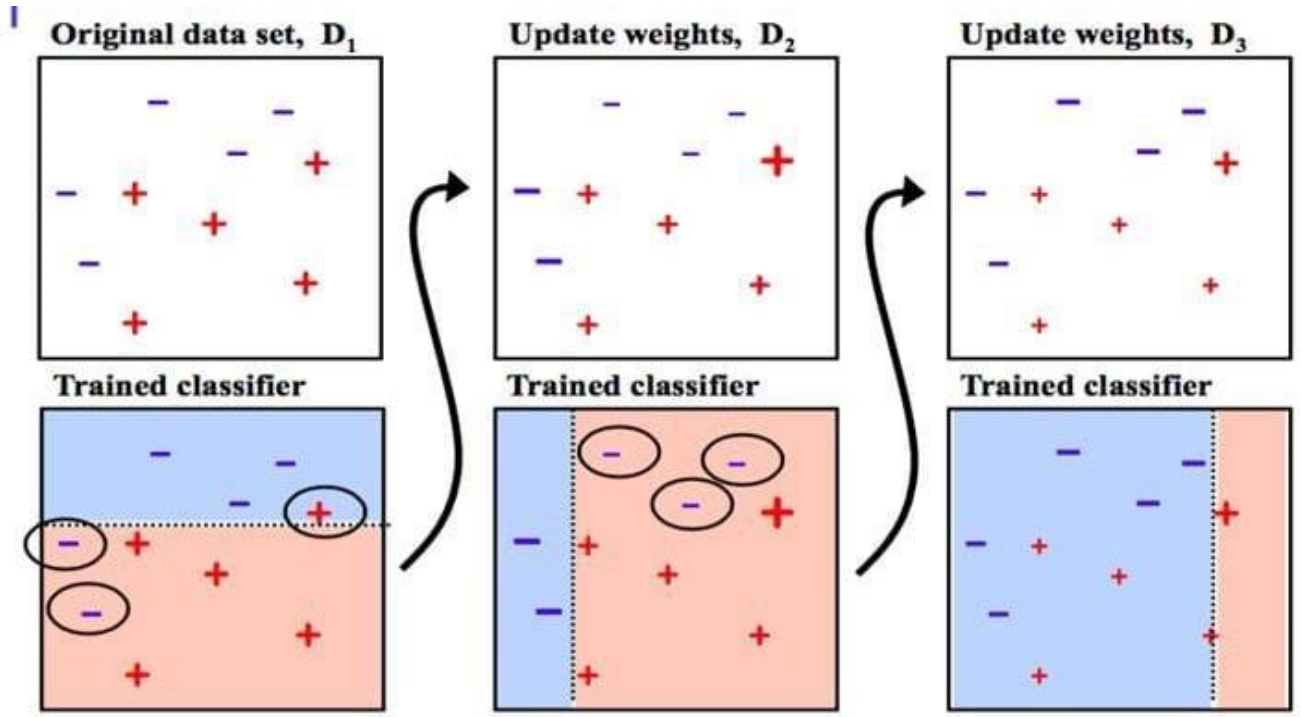
Boosting类方法

- 按照一定的先后顺序来训练不同的弱模型，每个弱模型都针对前一个弱模型的错误进行专门训练。根据前一个模型的结果来辅助下一个模型的训练，从而增加不同模型间的差异性。



AdaBoost

- 一种迭代式的线性算法，通过改变数据分布来提高弱模型的独立性。在每一轮训练中，增加分错样本的权重，减少分对样本的权重，从而强迫下一个弱模型学习新的特征。



AdaBoost

- Bootstrap

数据集	数据表示x	权重u
原始数据集	{(x1,y1), (x2,y2), (x3,y3), (x4,y4)}	(1, 1, 1, 1)
采样数据集1	{(x2,y2), (x2,y2), (x3,y3), (x4,y4)}	(0, 2, 1, 1)
采样数据集2	{(x1,y1), (x1,y1), (x1,y1), (x4,y4)}	(3, 0, 0, 1)

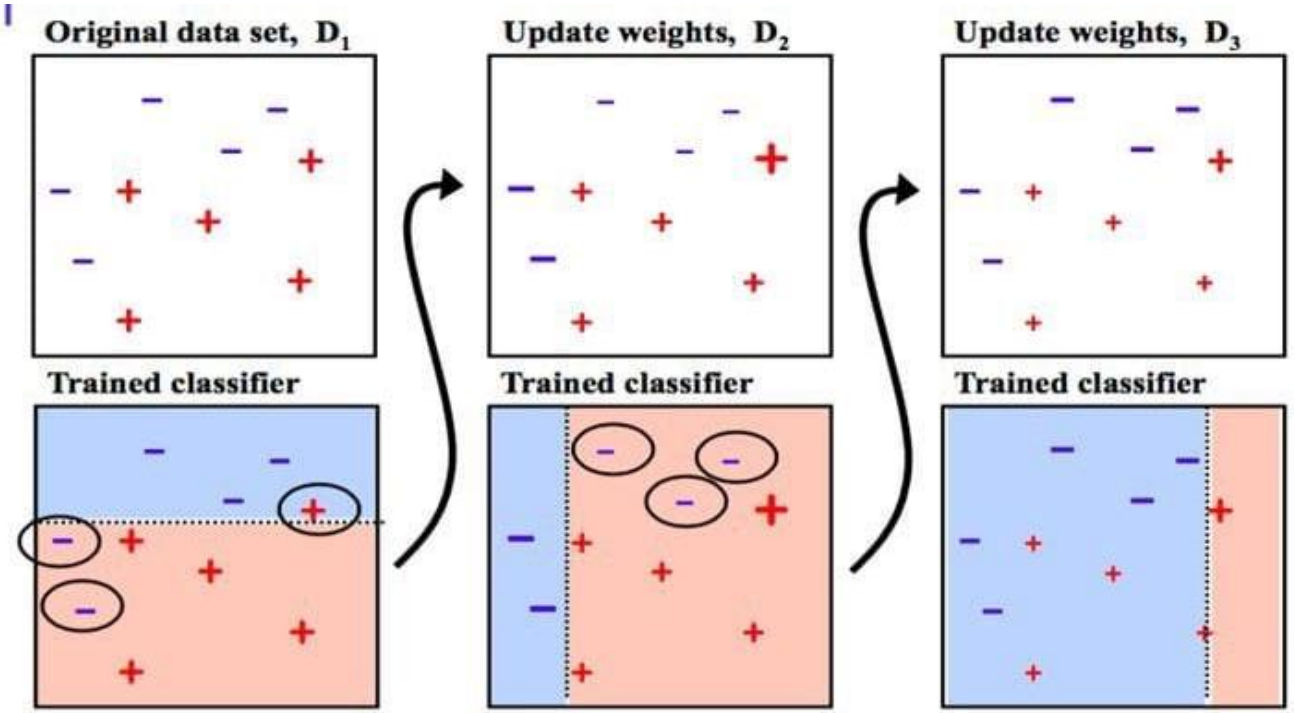
- $E^u = \frac{1}{N}\sum_{n=1}^{N} u_n err(y_n, h(x_n))$

- u_n可以视为每个数据点的权重

AdaBoost

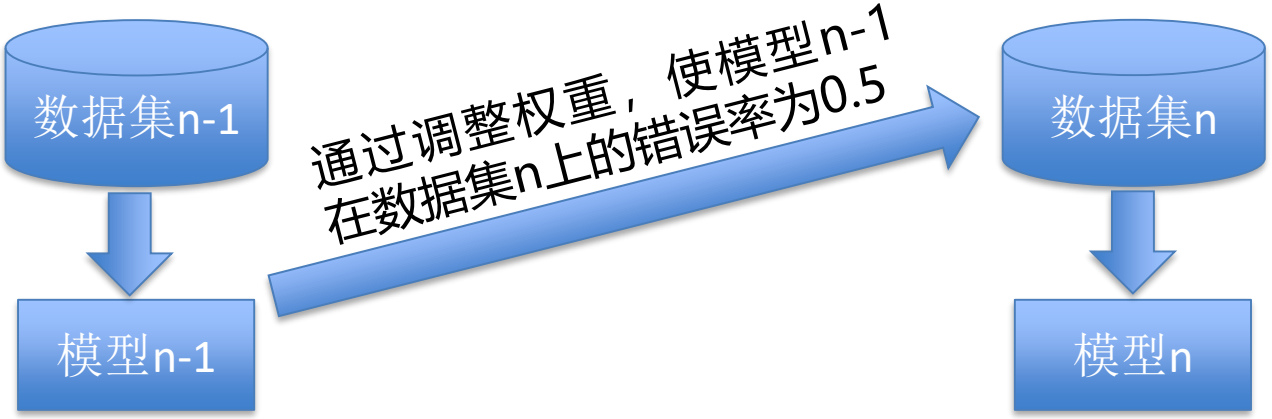
- 能不能控制权重u，让各个子模型学到不同的规则？

我们依次训练子模型。训练完一个模型后，在下一次训练开始时，将当前子模型**分类错误**的点的**权重增大**，**分类正确**的点的**权重减小**。那么下一个子模型就可以**更加关注当前分类错误的这些点**。



AdaBoost

- 能不能控制权重u，让各个子模型学到不同的规则?



在数据集n-1上
错误率e为0.2

在模型n上学习新的规则

- 为什么要保证新的错误率为0.5?

AdaBoost

- 调整权重，另新的错误率为0.5：

  u_incorrect ∝ 1-e

  u_correct ∝ e

- 实际操作中，我们令$s = \sqrt{\frac{1-e}{e}}$

  u_incorrect ← u_incorrect × s

  u_correct ← u_correct / s

AdaBoost

- 现在我们学到了N个子模型，如何进行组合？

使用加法模型将弱分类器进行线性组合，通过加权进行多数表决。错误率小的分类器具有更大的权值，错误率较大的分类器具有更小的权值。

- 怎么判断每个子模型的票数？

每次算子模型的时候，每个子模型都持有一个错误率e，直观的来想，e越偏离0.5，这个子模型的票数应该越高。

AdaBoost

- 怎么判断每个子模型的票数?

  实际操作中，我们使用$\alpha = \ln(s)$来作为每一个子模型的权重。 s为前面介绍的权重缩放参数，$s = \sqrt{\frac{1-e}{e}}$。

错误率e	缩放参数s	模型权重α
$e = 0.5$	$s = 1$	$\alpha = 0$
$e = 0.1$	$s = 3$	$\alpha = \ln(3)$
$e = 0.9$	$s = 1/3$	$\alpha = -\ln(3)$

AdaBoost

Input :

- A training set $S = ((\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m))$.

Initialization :

- Maximum number of iterations T;
- initialize the weight distribution $\forall i \in \{1, \ldots, m\}, D^{(1)}(i) = \frac{1}{m}$.

for $t = 1, \ldots, T$ **do**

- Learn a classifier $f_t : \mathbb{R}^d \to \{-1, +1\}$ using distribution $D^{(t)}$
- Set $\epsilon_t = \sum_{i: f_t(\mathbf{x}_i) \neq y_i} D^{(t)}(i)$
- Choose $a_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$
- Update the weight distribution over examples

$$\forall i \in \{1, \ldots, m\}, D^{(t+1)}(i) = \frac{D^{(t)}(i) e^{-a_t y_i f_t(\mathbf{x}_i)}}{Z^{(t)}}$$

where $Z^{(t)} = \sum_{i=1}^{m} D^{(t)}(i) e^{-a_t y_i f_t(\mathbf{x}_i)}$ is a normalization factor such that $D^{(t+1)}$ remains a distribution.

Output : The voted classifier $\forall \mathbf{x}, F(\mathbf{x}) = \text{sign}\left(\sum_{t=1}^{T} a_t f_t(\mathbf{x})\right)$

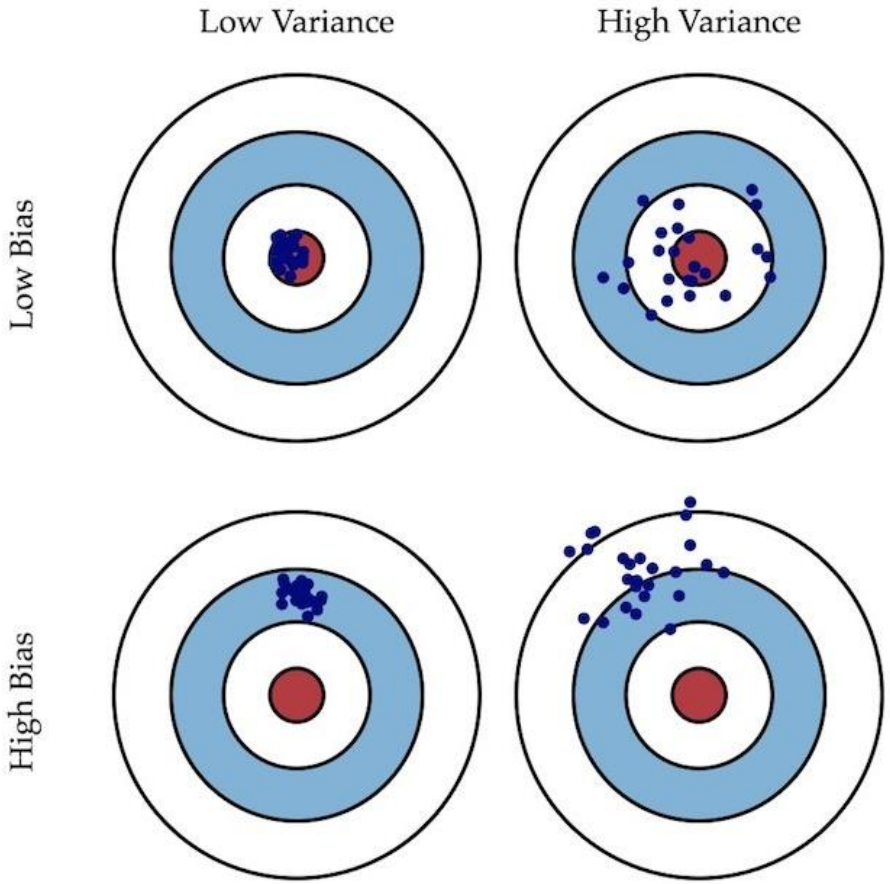
AdaBoost

- Adaboost直观总结：

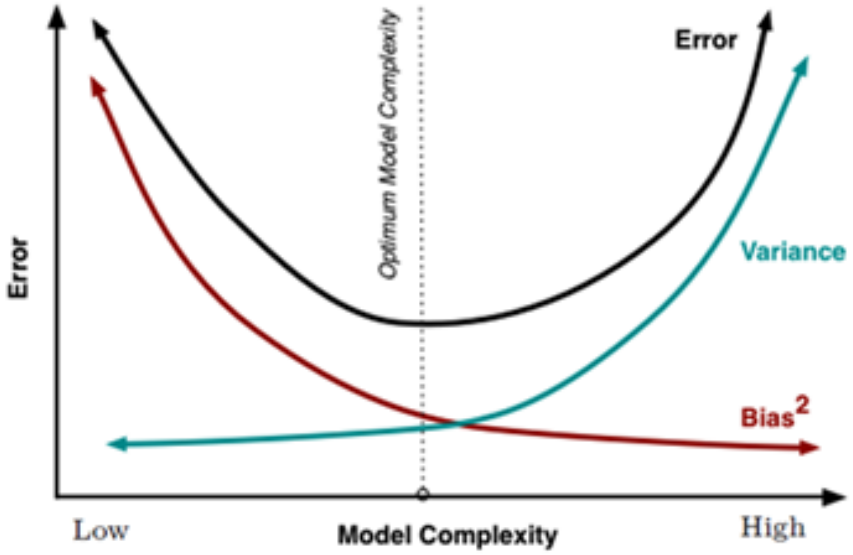
Adaboost = 弱模型g (学生)
+ 权重调整因子s (老师)
+ 加权求和α (班级)

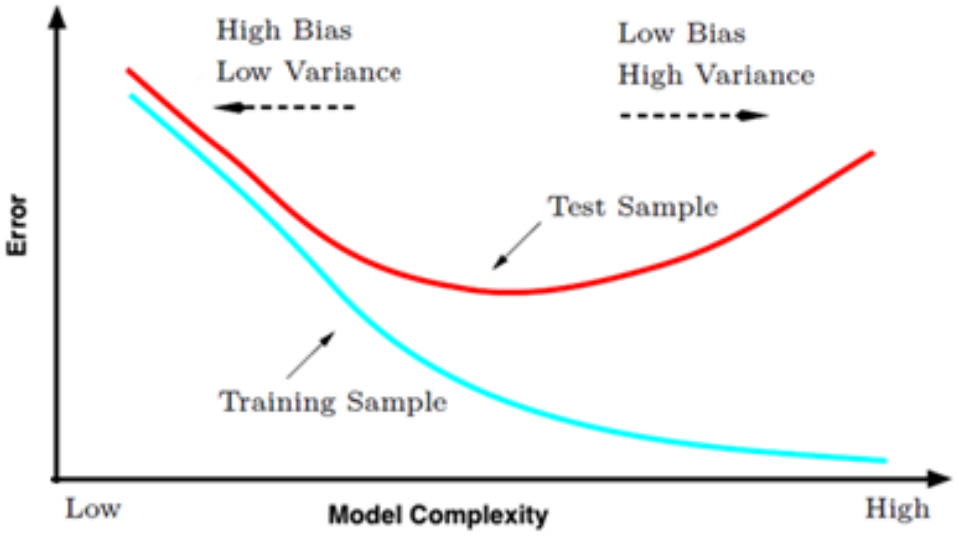
方差与偏差

- **偏差（Bias）**：描述模型输出结果的期望与样本真实结果的差距。
- **方差（Variance）**：描述模型对于给定值的输出稳定性。



方差与偏差





$$E((y-\hat{f}(x))^2) = \sigma^2 + Var[\hat{f}(x)] + (Bias[\hat{f}(x)])^2$$

样本噪音　　方差　　偏差

总结

- 基于方差偏差的策略
  - [高方差] 采集更多的样本数据
  - [高方差] 减少特征数量，去除非主要的特征
  - [高偏差] 引入更多的相关特征
  - [高偏差] 采用多项式特征
  - [高偏差] 减小正则化参数 λ
  - [高方差] 增加正则化参数 λ

总结

- Bagging（融合）
  - 减少方差，通常也可以避免过拟合，但基准模型需要低偏差
  - 基准模型互相独立，运行速度快
- Boosting（提升）
  - 减少偏差，基准模型需要低方差，避免过拟合
  - 各个基准模型只能顺序生成，运行速度较慢

Stacking（堆叠）

- 一种分层模型集成框架。使用一个元模型或者多个元模型来整合多个模型的集成学习技术。基础模型利用整个训练集做训练，元模型将基础模型的特征作为特征进行训练。

Algorithm Stacking

Input: training data $D = \{x_i, y_i\}_{i=1}^{m}$
Ouput: ensemble classifier H
Step 1: learn base-level classifiers
for $t = 1$ to T **do**
    learn h_t based on D
end for
Step 2: construct new data set of predictions
for $i = 1$ to m **do**
    $D_h = \{x'_i, y_i\}$, where $x'_i = \{h_1(x_i), ..., h_T(x_i)\}$
end for
Step 3: learn a meta-classifier
learn H based on D_h
return H

期中project

- 报告评分：
  - 每组同学共同完成一份报告
  - 参考论文要有参考文献，参考的代码要标记来源
  - RNN模型和CNN模型的实现各占50分
  - 报告提交DDL：11月7日晚11:00

评分项	说明	分值
实验原理	总结两种模型的原理	20
网络结构	画出自己模型的网络结构示意图	10
结果分析	展示并分析不同结构下的实验结果	40
创新	可以借鉴现有方法，但需总结原理	20
排版	整体美观性	10
组员分工	总结组员各自做了什么工作	0

期中project

- PPT展示：
  - 小组成员共同完成验收，时间为5~10分钟，超时扣分
  - 通过PPT来展示期中project小组成员完成的工作
  - PPT展示会占据一定的期中project分数
  - 验收时间：11月8日实验课

评分项	说明
实现思路	总结这段时间内实现进度
网络结构	介绍自己模型的网络结构
结果分析	展示并分析实验结果
创新	介绍有哪些创新与尝试