法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

- 课程详情请咨询
  - 微信公众号：小象学院
  - 新浪微博：小象AI学院

详解AdaBoost算法

目录

集成模型简介

个体分类器

又称为基分类器，由现有的算法从数据集中产生，例如：

- 决策树(C4.5，CART)
- 逻辑回归
- SVM

个体分类器的异同

同质：全部的个体分类器属于同一类别(可以有不同参数)

异质：包含两种或更多的个体分类器

集成模型简介

为什么要构建集成模型

和单一分类器比，集成模型具有如下优点：

- 泛化能力强
- 预测性能准确跟结构复杂的分类器相当
- 稳定性高
- 对数据的容忍度更高

集成模型简介

- 关于集成模型的准确性的简单推导

考虑二分类问题$y \in \{-1,1\}$和真实函数f,，假设基分类器h_i的错误率为ϵ，若采用简单投票法则结合T个基分类器，且超过一半的基分类器正确就认为集成分类器正确：

$$H(x) = sign(\sum h_i(x))$$

在各个基分类器错误率独立的假设下，集成分类器的错误率为：

$$P\big(H(x) \neq f(x)\big)$$

$$= \sum_{k=0}^{T/2} \binom{T}{k} (1-\epsilon)^k \epsilon^{T-k} \leq \exp\left(-\frac{T(1-2\epsilon)^2}{2}\right)$$

集成模型简介

关于集成模型的准确性的简单推导(续)

可以看出，当T增大时，集成的错误率将指数下降，最终趋向于0。

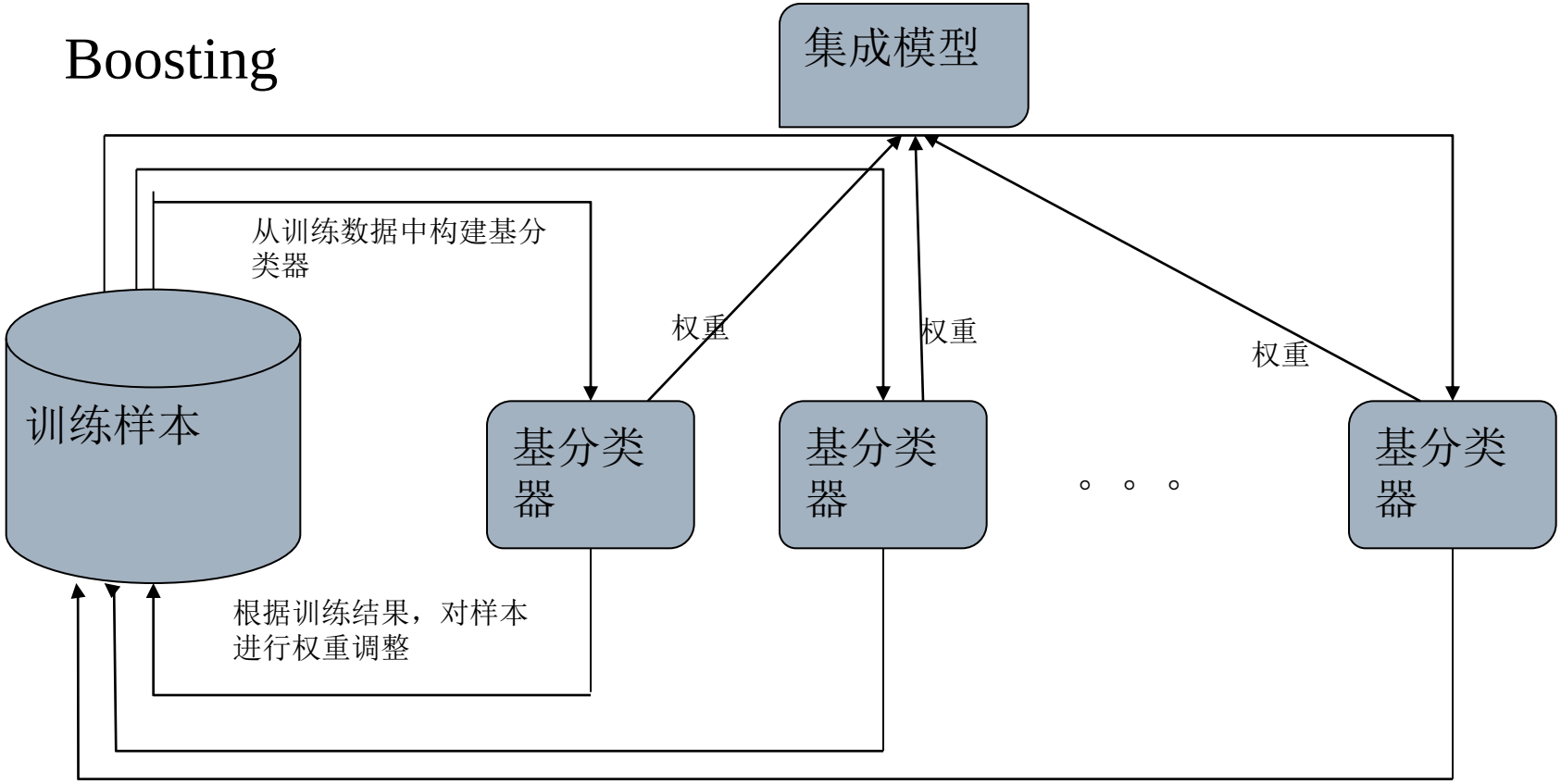
注意：

- 基分类器的“错误率独立”的假设往往不成立。
- 目标：产生好而不同的个体分类器

集成模型简介

- 两种常用的集成：Boosting和Bagging

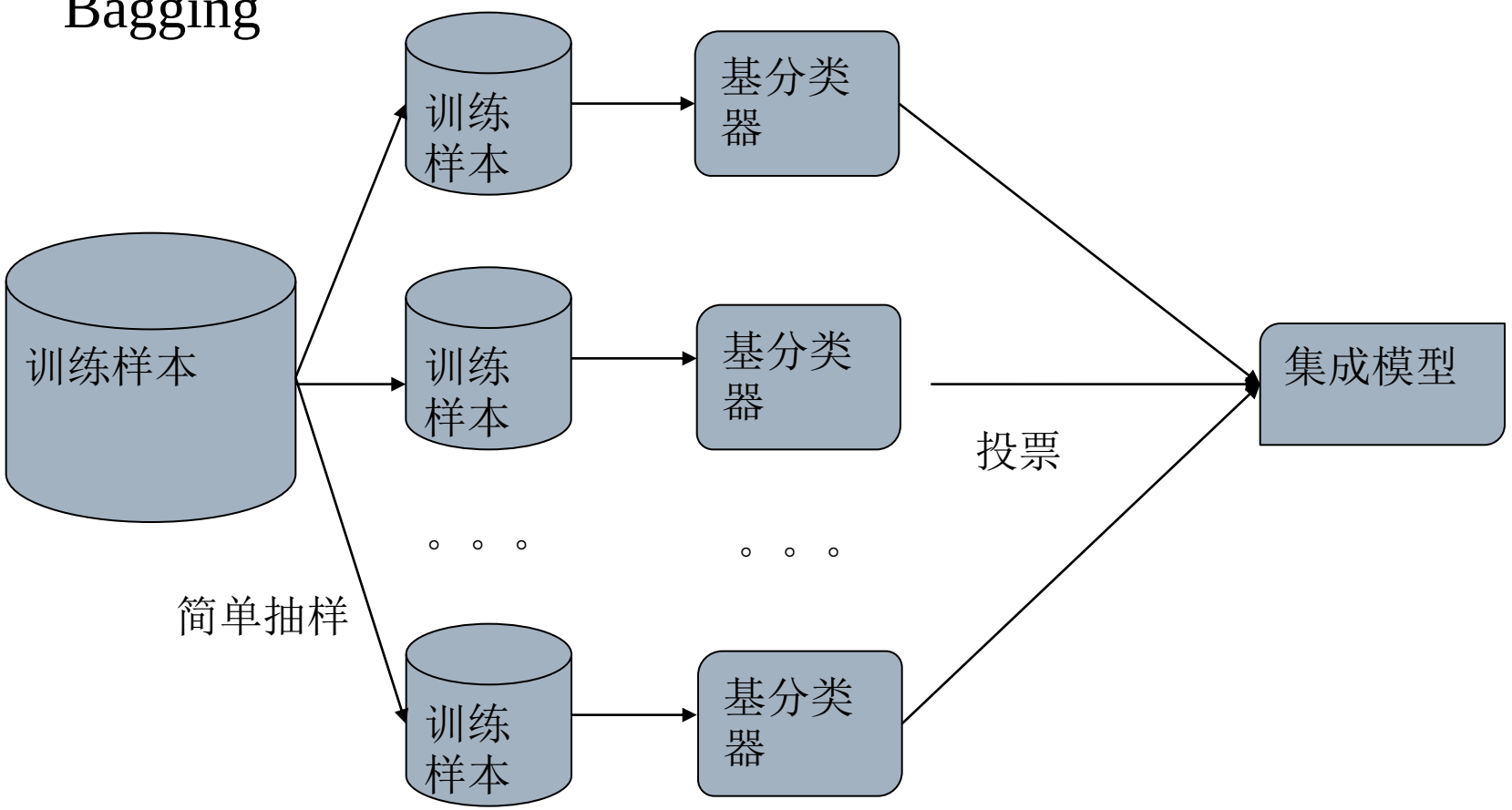
Boosting



集成模型简介

- 两种常用的集成：Boosting和Bagging

Bagging



目录

集成模型简介

决策树简介：以CART为例

AdaBoost算法

案例：基于CART的AdaBoost算法

决策树简介：以CART为例

- CART：分类与回归树

采用一种二分递归分割的技术，将当前 的样本集分为两个子样本集，使得生成的子节点都有两个分支。因此，CART算法生成的决策树是结构简洁的二叉树。

分类：

叶节点上出现频率最高的类别

回归：

叶节点上目标变量的均值

决策树简介：以CART为例

- Gini指数

在分类问题上，Gini指数是反应样本纯度的指标。Gini越小，纯度越高。

$$Gini(D) = 1 - \sum p_k^2$$

D:数据集

p_k:第k类样本的占比，$\sum p_k = 1$

决策树简介：以CART为例

- Gini指数(续)

属性a在数据集D中的基尼指数：

$$Gini(D,a)=\sum\frac{|D_v|}{|D|}Gini(D_v)$$

D_v是属性a的第v个值对应的全部样本，$\{D_v\}$的一个划分。

- Gini指数的作用

遍历所有可能的属性并计算出Gini指数，选择最小Gini指数对应的属性进行划分

决策树简介：以CART为例

- 连续属性

由于CART是二叉树，对于连续属性x需要进行切分。

步骤：

1. 确定划分的阈值m,将数据集D划分为{x<=m}和{x>m},计算出当前的Gini(D,x|m)
2. 所有的Gini(D,x|m)中，最小的Gini(D,x|m)就是x在D上的Gini，此时的m就是最优切分

决策树简介：以CART为例

- 离散属性

当离散属性取值个数较少时

枚举出所有可能的组合，对每种组合计算Gini。最小的Gini对应的切分方案视为最优切分方案。

例如：婚姻={已婚，未婚，其他}，枚举方案有：

- ✓ {已婚}，{未婚，其他}
- ✓ {未婚}，{已婚，未婚}
- ✓ {其他}，{已婚，未婚}

决策树简介：以CART为例

- 离散属性(续)

当离散属性取值较多时，枚举法将变的很困难。

解决方法：

对离散属性进行数值编码，例如用目标变量的浓度进行编码，将离散属性变为连续属性，再进行分割

注：

次序变量(如年级，学历等)的分割，要保证次序不变。

决策树简介：以CART为例

建立CART的步骤

输入：训练数据集D，停止计算的条件；
输出：CART决策树。

根据训练数据集，从根结点开始，递归地对每个结点进行以下操作，构建二叉决策树：

1.设结点的训练数据集为D，计算现有特征对该数据集的Gini系数。

2.在所有可能的特征A以及它们所有可能的切分点a中，选择Gini系数最小的特征及其对应的切分点作为最优特征与最优切分点。依最优特征与最优切分点，从现结点生成两个子结点，将训练数据集依特征分配到两个子结点中去。

3.对两个子结点递归地调用步骤1~2，直至满足停止条件。

4.生成CART决策树。

决策树简介：以CART为例

建立CART的步骤(续)

终止条件：

- 当前没有特征可以选择
- 所有样本的类别一致
- 所有样本在所有属性上的取值一致
- 其他预设条件，如深度、叶节点个数、Gini值得等

关键思想：

- ✓ 分而治之
- ✓ 递归

目录

集成模型简介

决策树简介：以CART为例

AdaBoost算法

案例：基于CART的AdaBoost算法

AdaBoost算法

AdaBoost基本原理

一种基于Boosting思想的迭代算法，通过改变数据分布来实现基分类器的迭代(分类性能越来越强)。

它根据每次训练集之中每个样本的分类是否正确，以及上次的总体分类的准确率，来确定每个样本的权值。将修改过权值的新数据集送给下层分类器进行训练，最后将每次训练得到的分类器最后加权融合起来，作为最后的决策分类器。

权值调整思想：

上一次分类正确的样本，降低权值；上一次分类错误的样本，提高权值。

AdaBoost算法

错误率及权重调整

错误率：

$$\epsilon = \frac{\text{未正确分类的样本数目}}{\text{所有样本数目}}$$

权重：

AdaBoost模型有2种权重：

- 第t步迭代中的分类器的权重α^t
- 第t步迭代中每个样本的权重D_i^t

AdaBoost算法

错误率及权重调整(续)

分类器的权重α^t的计算公式：

$$\alpha = \frac{1}{2}\ln(\frac{1-\epsilon}{\epsilon})$$

样本的权重D_i^t的计算公式：

对于分类正确的样本，

$$D_i^{t+1} = \frac{D_i^t e^{-\alpha}}{sum(D)}$$

对于分类错误的样本，

$$D_i^{t+1} = \frac{D_i^t e^{\alpha}}{sum(D)}$$

AdaBoost算法

- 错误率及权重调整(续)

其中$sum(D)$是归一化因子，使得$\sum D_i^{t+1} = 1$

思考：

当错误率$\epsilon > 0.5$时，会有什么样的影响，此时应该怎么操作

AdaBoost算法

- 基分类器的集成

在T次迭代中，训练出T个分类器$\{G_t\}$及T个权重$\{\alpha_t\}$，t=1,2,…,T,最终的分类器为：

$$f(x) = \sum_{t=1}^{T} \alpha_t G_t(x)$$

目录

案例

- 基于CART的AdaBoost模型在违约预测中的应用

CART中样本的权重调整

由于CART使用Gini作为属性选择的指标，且样本权重无法反映在Gini中，故无法直接将样本的权值带入CART的构建中去。

解决方法：

找出所有样本的权重$\{w_i\}$的最小值w_{min},对于样本i，重复$\frac{w_i}{w_{min}}$次放入样本中。

思考：

采用错误率作为属性选择的指标，就可以直接将样本的权值带入基分类器的构建中去

案例

- 基于CART的AdaBoost模型在违约预测中的应用(续)

4521条记录，

14个属性：年龄、婚姻、职业、历史违约记录等

1个目标变量：在贷后中是否违约

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
age	job	marital	education	default	spending	housing	cash_loan	contact_number_type	maturity	app_channel	max_late_charge	cash_withdraw_freq	poutcome	y
33	services	married	secondary	no	4789	yes	yes	cellular	220	1	339	4	success	no
35	manageme	single	tertiary	no	1350	yes	no	cellular	185	1	330	1	success	no
35	manageme	single	tertiary	no	747	no	no	cellular	141	2	176	3	success	no
43	services	married	primary	no	200	yes	yes	cellular	313	1	147	2	success	no
31	blue-collar	married	secondary	no	360	yes	yes	cellular	89	1	241	1	success	no
37	admin.	single	tertiary	no	2317	yes	no	cellular	114	1	152	2	success	no
67	retired	married	unknown	no	696	no	no	telephone	119	1	105	2	success	no
33	manageme	married	secondary	no	3935	yes	no	cellular	35	1	342	2	success	yes
38	manageme	single	tertiary	no	11971	yes	no	unknown	244	2	101	3	success	no
51	blue-collar	divorced	secondary	no	203	yes	no	cellular	134	1	170	5	success	no
40	unemploye	married	secondary	no	219	yes	no	cellular	204	2	196	1	success	no
52	services	married	secondary	no	657	no	no	telephone	33	2	460	2	success	yes
30	admin.	single	tertiary	no	261	no	no	cellular	233	1	137	20	success	no
38	manageme	single	tertiary	no	493	yes	no	cellular	188	1	367	7	success	no
32	manageme	single	tertiary	no	574	yes	no	cellular	259	2	145	3	success	no
33	blue-collar	single	secondary	no	200	no	no	cellular	76	2	207	1	success	no
51	manageme	single	tertiary	yes	200	yes	no	cellular	281	2	266	6	success	no
32	blue-collar	single	secondary	no	228	no	no	telephone	176	1	288	3	success	no
48	admin.	married	unknown	no	200	yes	no	cellular	85	1	168	2	success	no
29	admin.	single	secondary	no	428	yes	yes	cellular	54	1	345	2	success	no
55	technician	married	secondary	no	273	yes	no	cellular	84	3	183	3	success	no
28	manageme	single	tertiary	no	200	no	no	cellular	311	2	146	2	success	yes
37	admin.	married	secondary	no	200	yes	no	cellular	147	2	347	1	success	no
37	admin.	married	tertiary	no	200	yes	no	cellular	65	2	119	1	success	no
36	technician	married	secondary	no	200	yes	no	cellular	152	2	347	1	success	no

案例

- 基于CART的AdaBoost模型在违约预测中的应用(续)

尝试构建深度为9的单一CART模型，和迭代10次、深度为3的AdaBoost模型，试验结果如下：

CART：错误率60%

AdaBoost:错误率40%

疑问

- 问题答疑： http://www.xxwenda.com/
  - 可邀请老师或者其他人回答问题

联系我们

小象学院：互联网新技术在线教育领航者

- 微信公众号：小象学院
- 新浪微博：小象AI学院