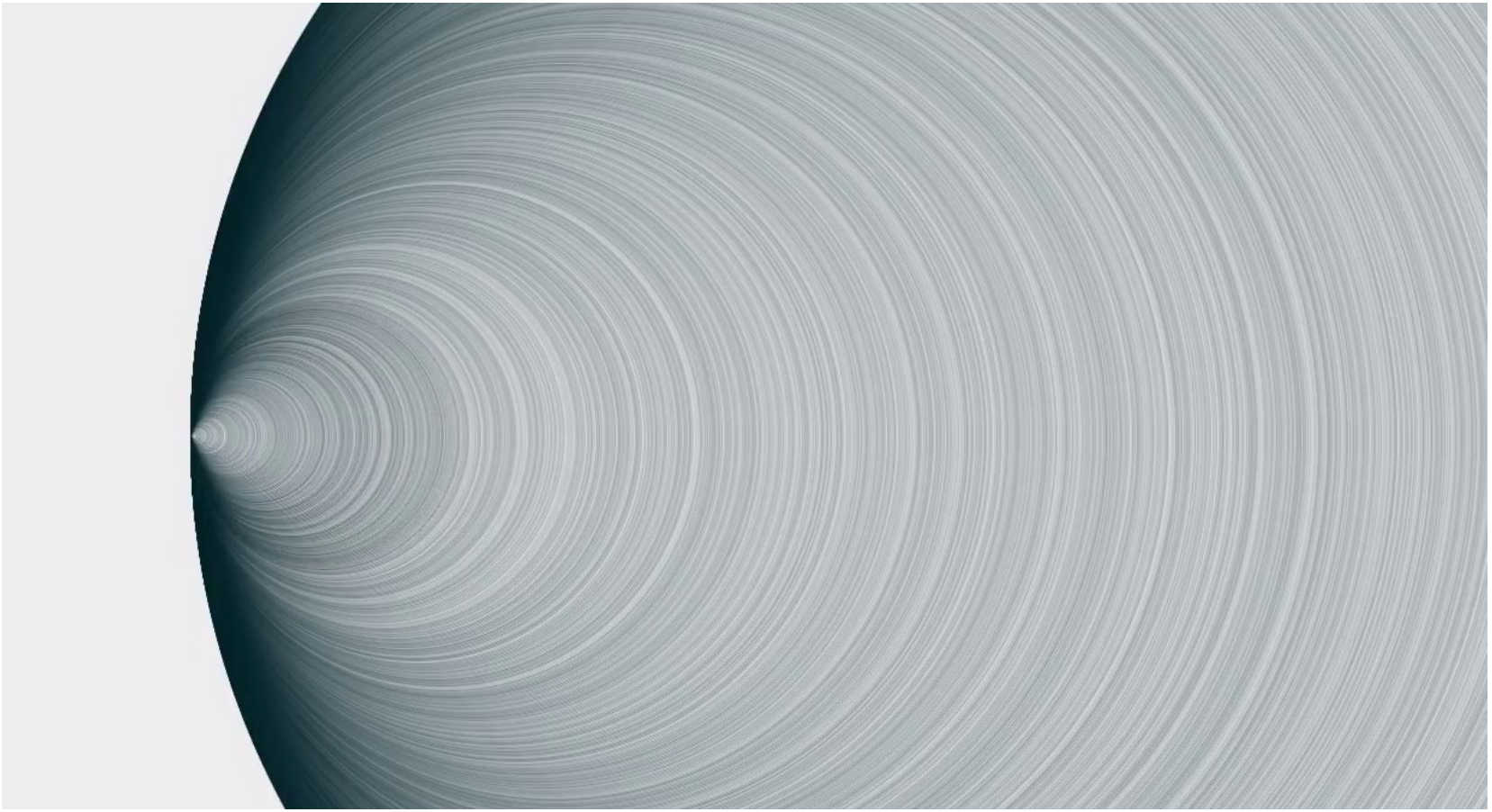
机器学习

Xin Tao



日程

I. 直觉类算法部分：

1. 信息论
2. 决策树
3. 集成算法

II. 应用讨论

III. 数学类讲解部分：

1. 概率论数学标记法
2. 贝叶斯统计数学标记法
3. 贝叶斯概率视角下的线性回归

信息论

信息是概率的另一面

Information Theory

熵(Entropy)衡量不确定性：

$$H = -\sum_{i=1}^{N} p_i \log_2(p_i)$$

抛正常硬币的熵为：

$H= -[0.5*\log_2(0.5)+ 0.5*\log_2(0.5)]= 1$

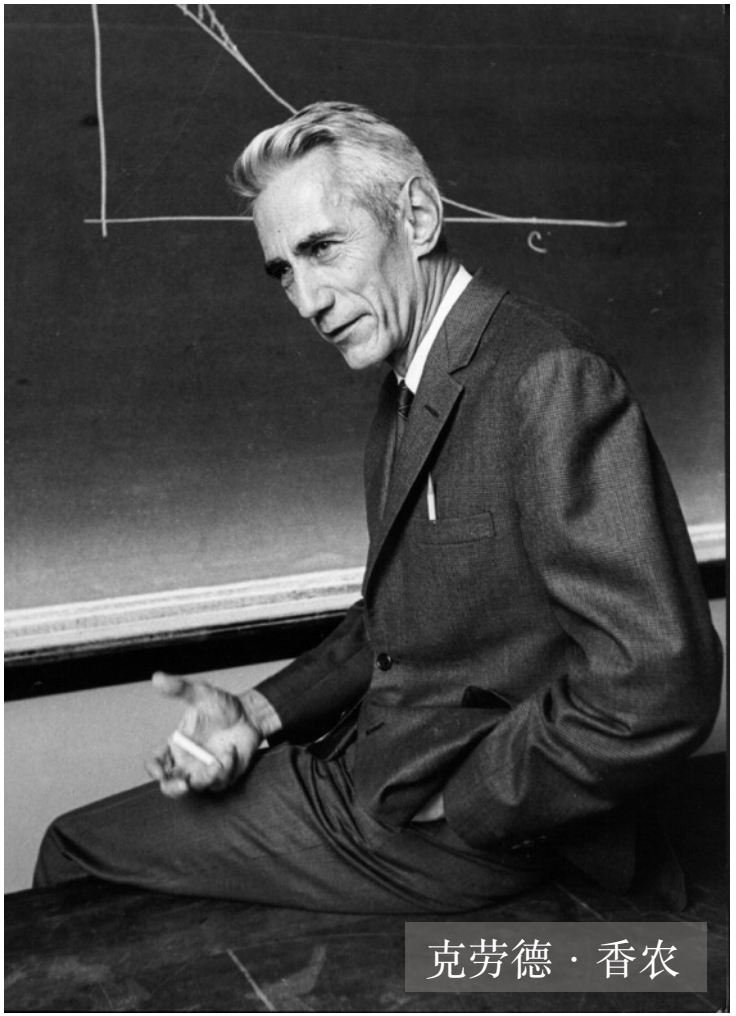
抛不均匀硬币（字的概率为0.25）的熵为

$H= -[0.75*\log_2(0.75)+ 0.25*\log_2(0.25)]= 0.81$

抛两面都是字的硬币的熵为：

$H= -1*\log_2(1) = 0$

- 正常硬币不确定性最高，而两面都是字的硬币没有不确定性。
- 概率分布范围越广越平均熵越高。
- 猜硬币的熵是1，而如果你已经知道答案，熵就等于零（概率坍塌）。信息会降低熵。



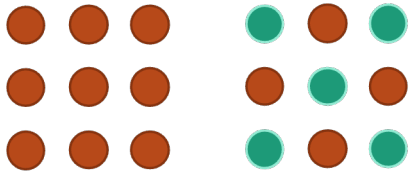
信息论与分类算法

成功的分类会增加信息，从而降低系统和每一部分的熵

Information Theory and Classification

熵是很好的的分类算法损失方程：

- 分类后类别内的样本应该比较单一（例如：都是红球）。
- 类别内、总体熵值下降、信息增加。



低熵

P(红)=100%

高熵

P(红)=44%

P(绿)=56%

决策树(1984)是最常用的分类基础算法之一：

- 市场研究、医疗诊断、赌球，Xbox Kinect…
- 能生成人可以使用的、直观的决策模型。
- 缺点是容易过度拟合
- 其集成版本（如：随机森林）可以有效解决过度拟合问题(1994)

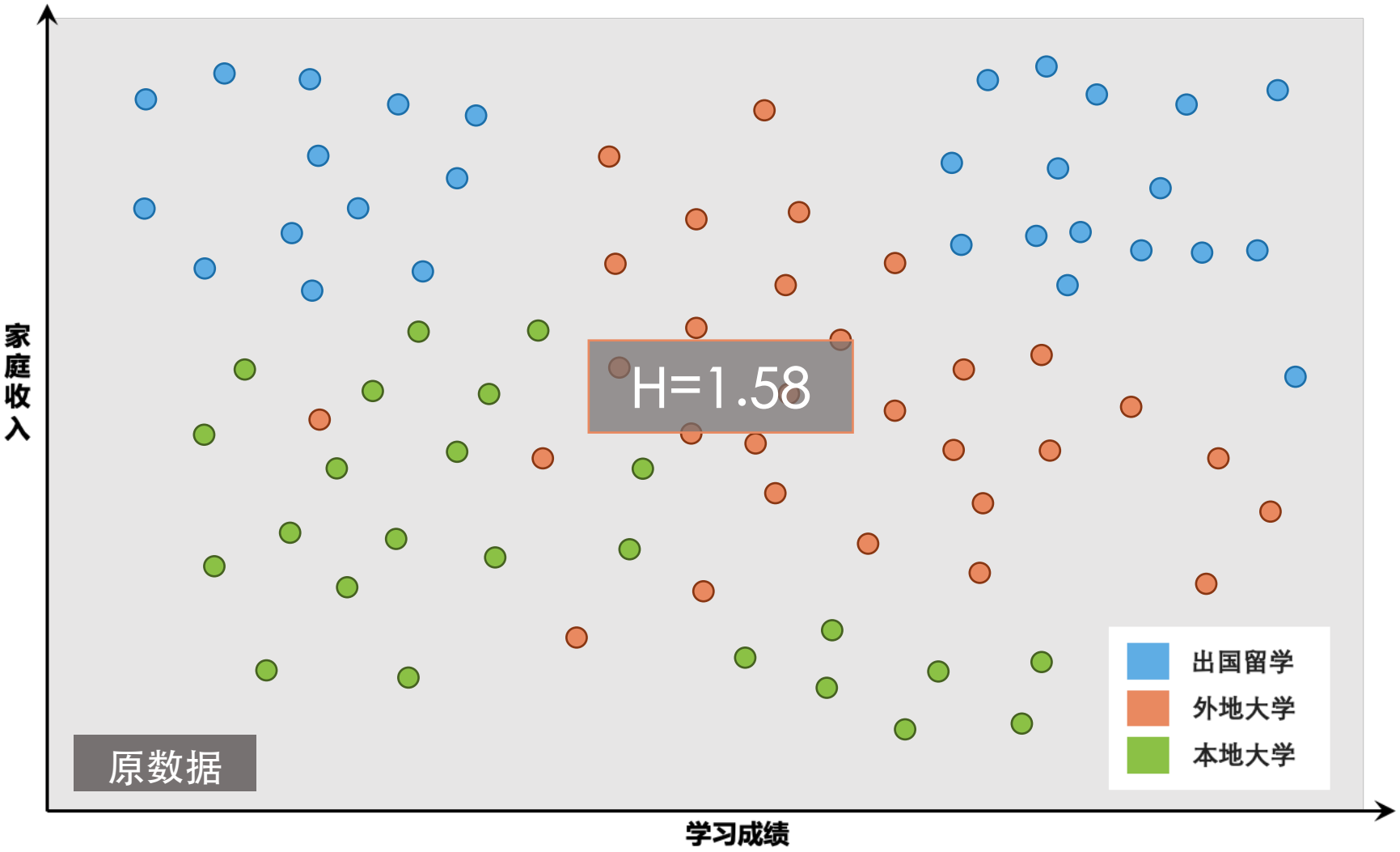
教育咨询例子：

- 一家教育咨询机构希望了解人们选择留学、在外地、或在本地留学的原因（市场调研）。
- 取得了1000多家庭的数据、包括上学地点、收入、小孩学习成绩。
- 现在拿出83个数据做分析。

决策树

怎么做决定?

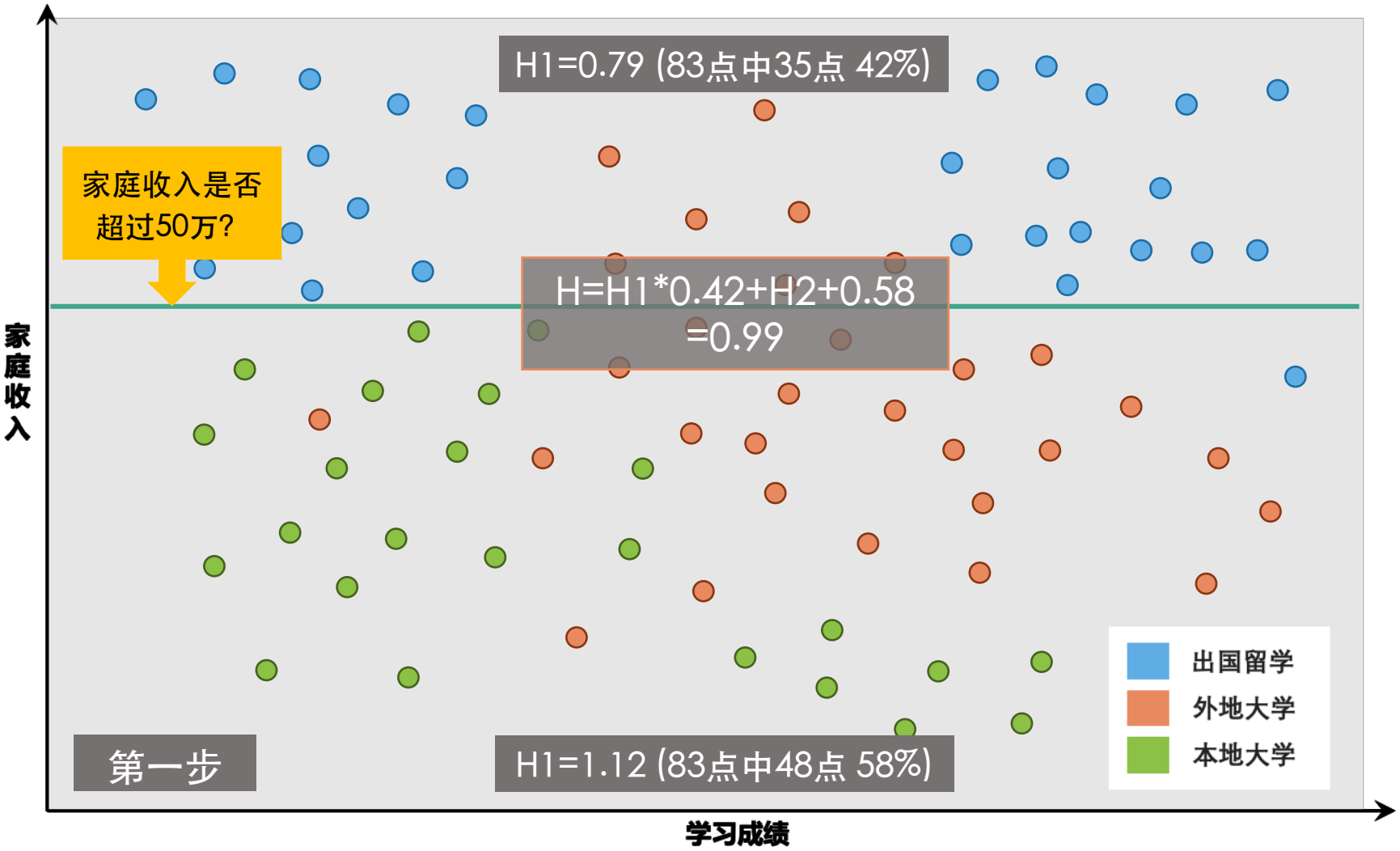
Decision Tree



决策树

分割的依据：分割后系统的熵值H下降最多

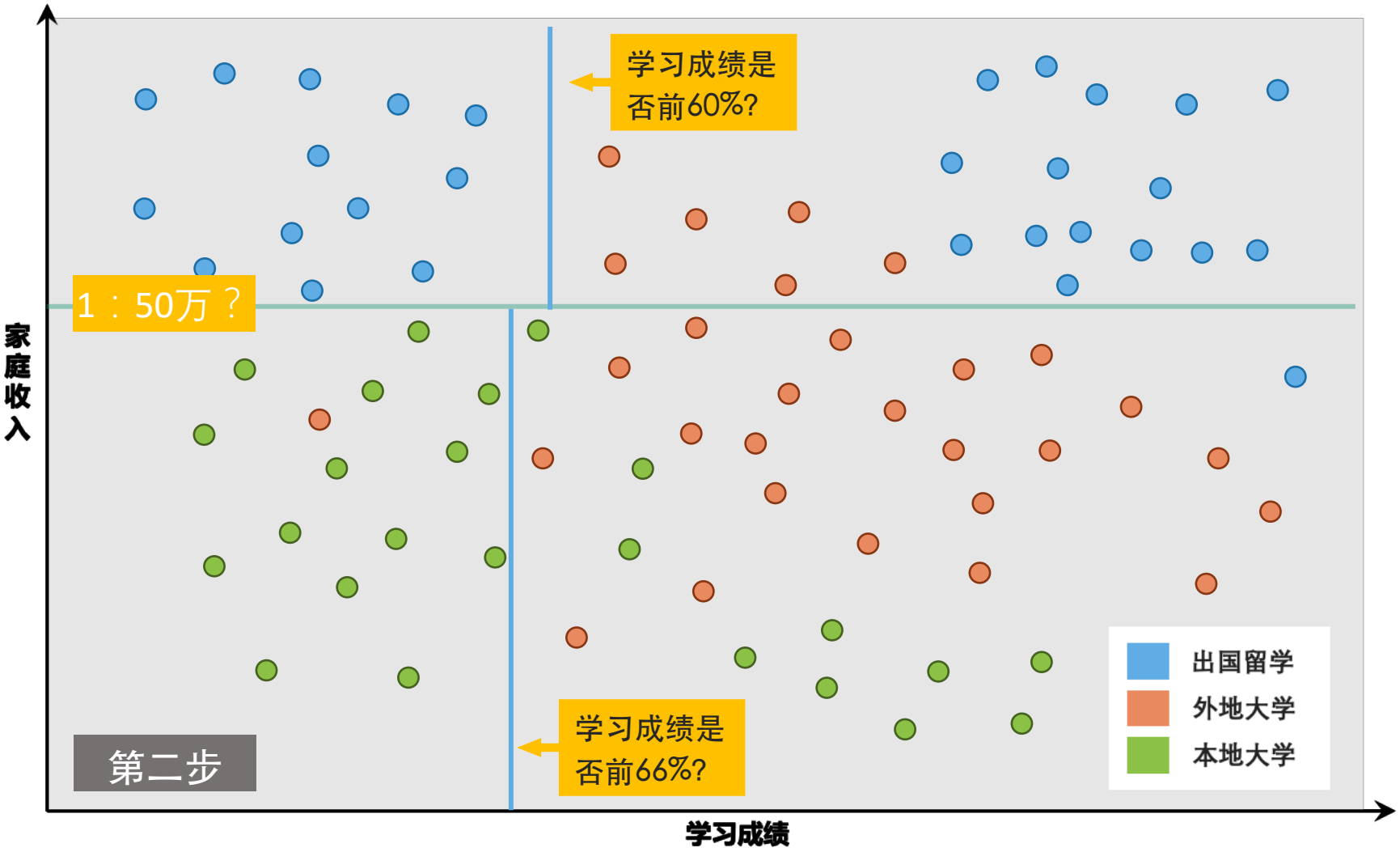
Decision Tree



决策树

继续分割，决策树第二层

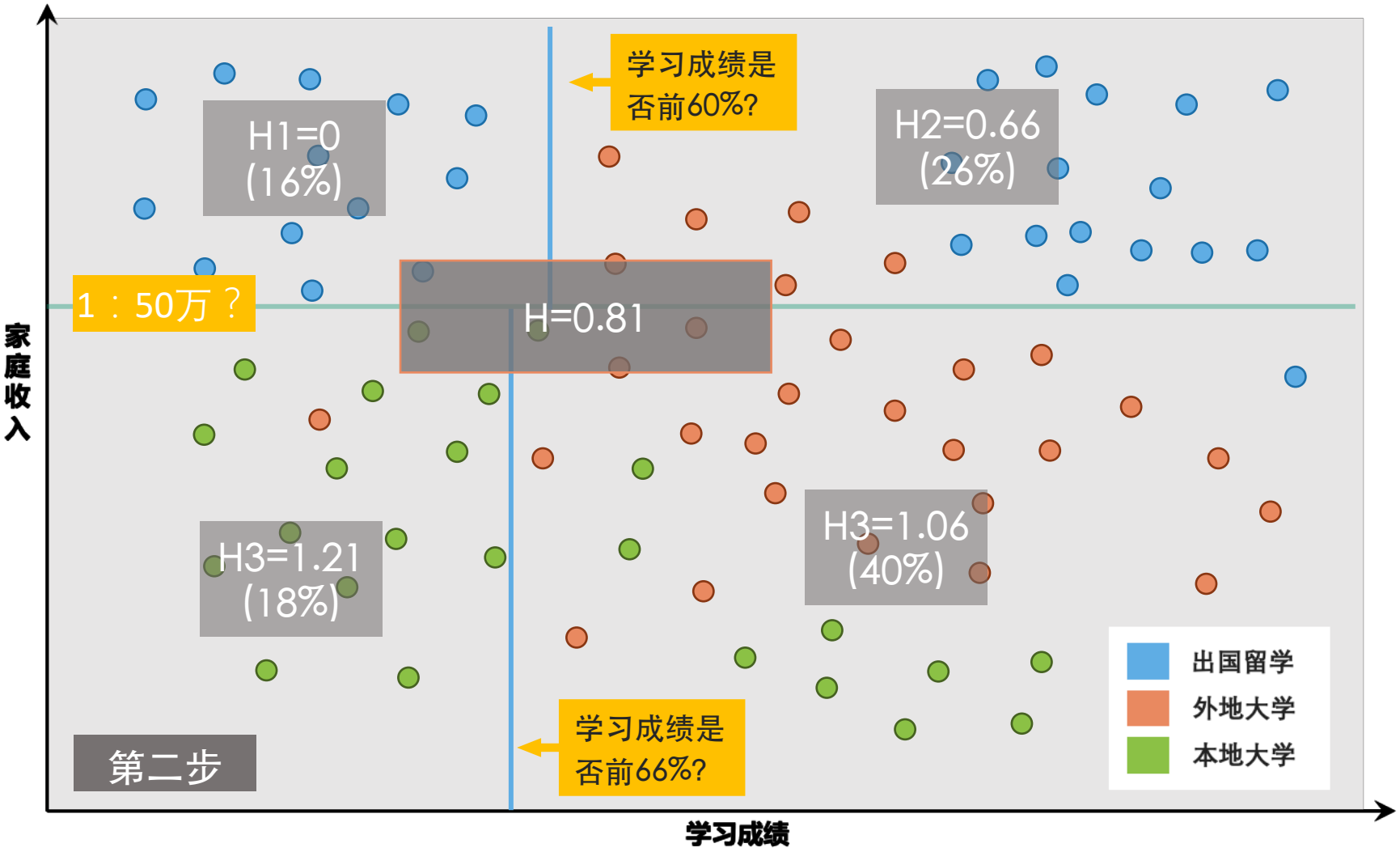
Decision Tree



决策树

系统熵值继续下降

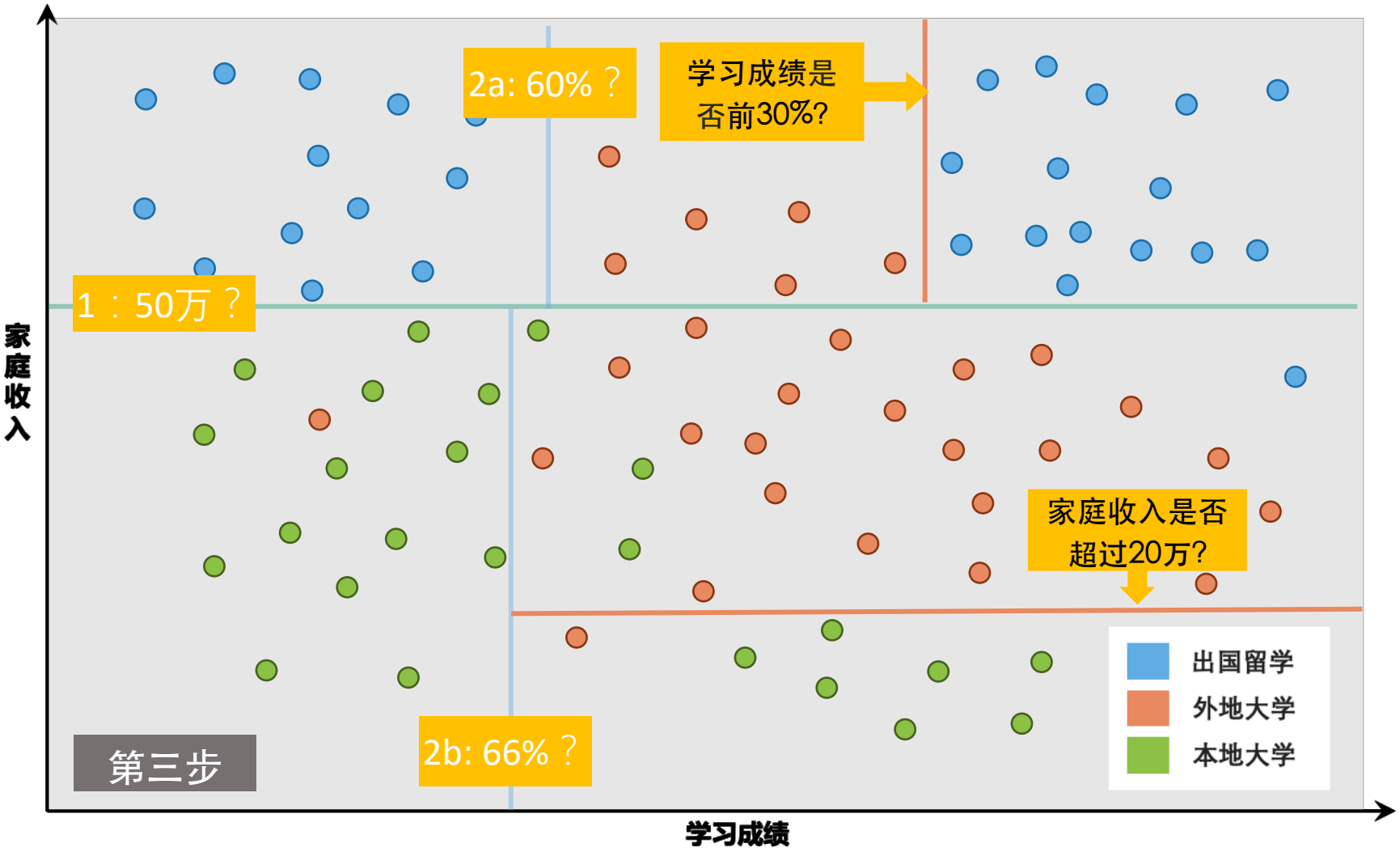
Decision Tree



决策树

继续分割，决策树第三层

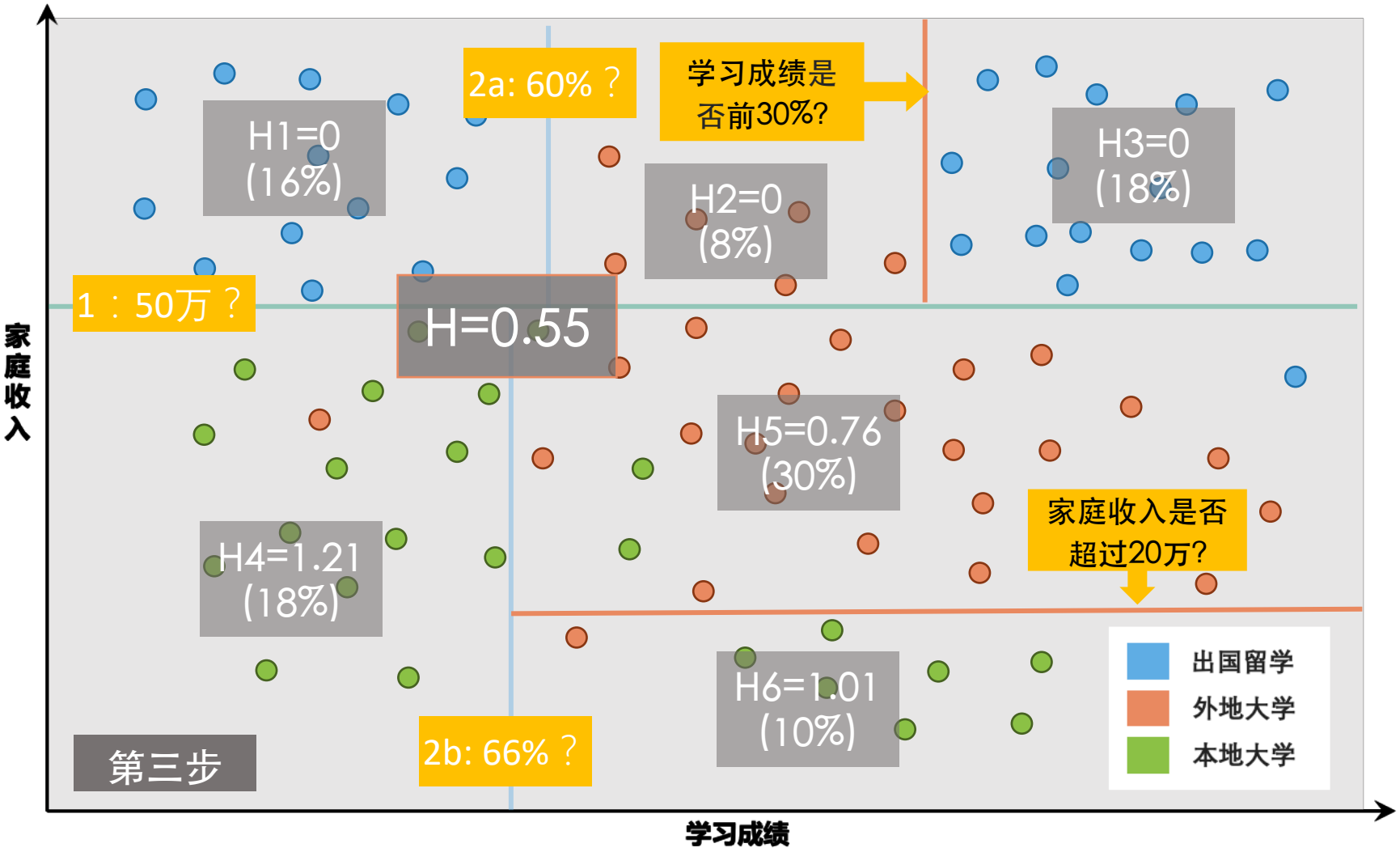
Decision Tree



决策树

停止，系统熵从1.58下降到0.55

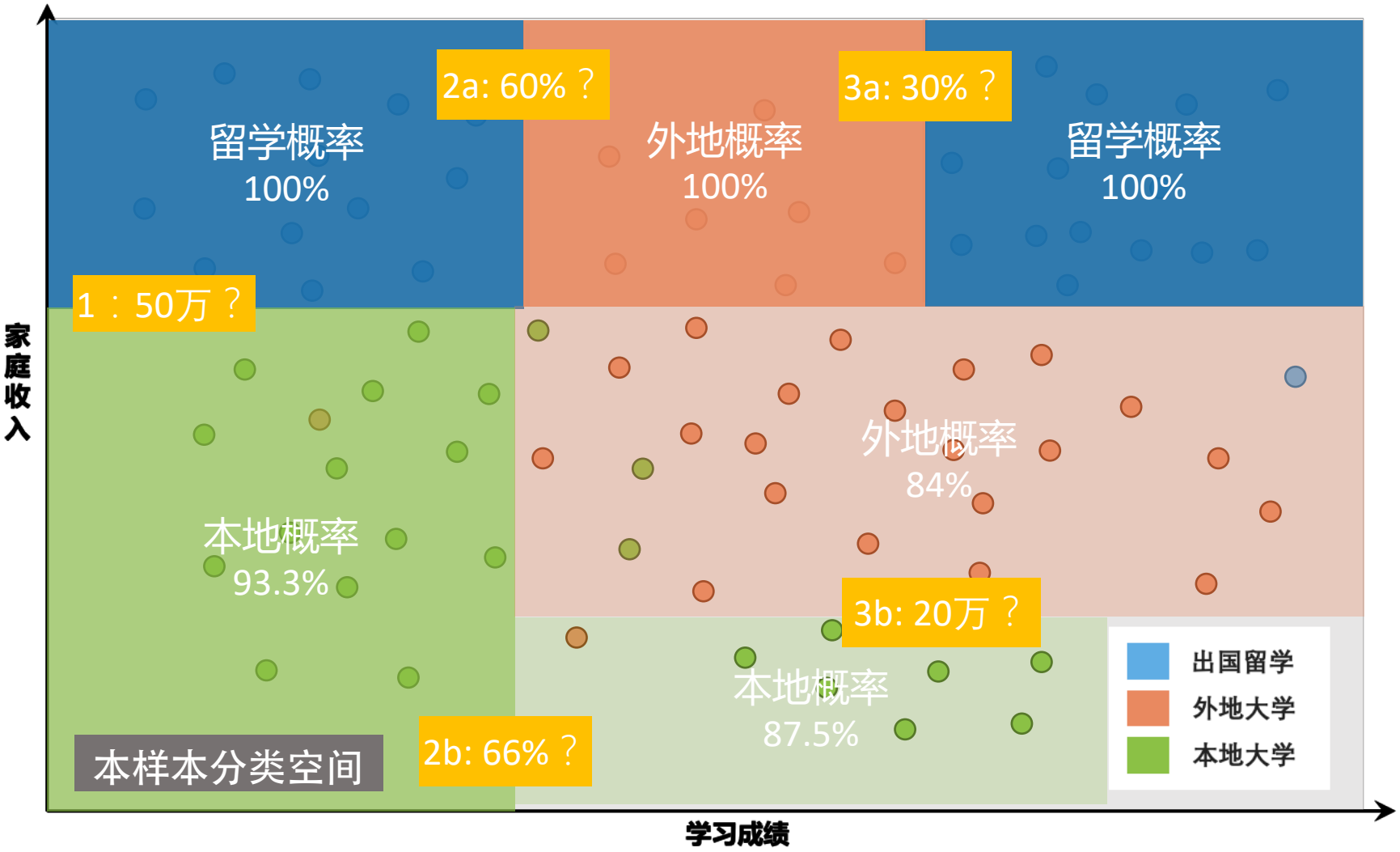
Decision Tree



决策树

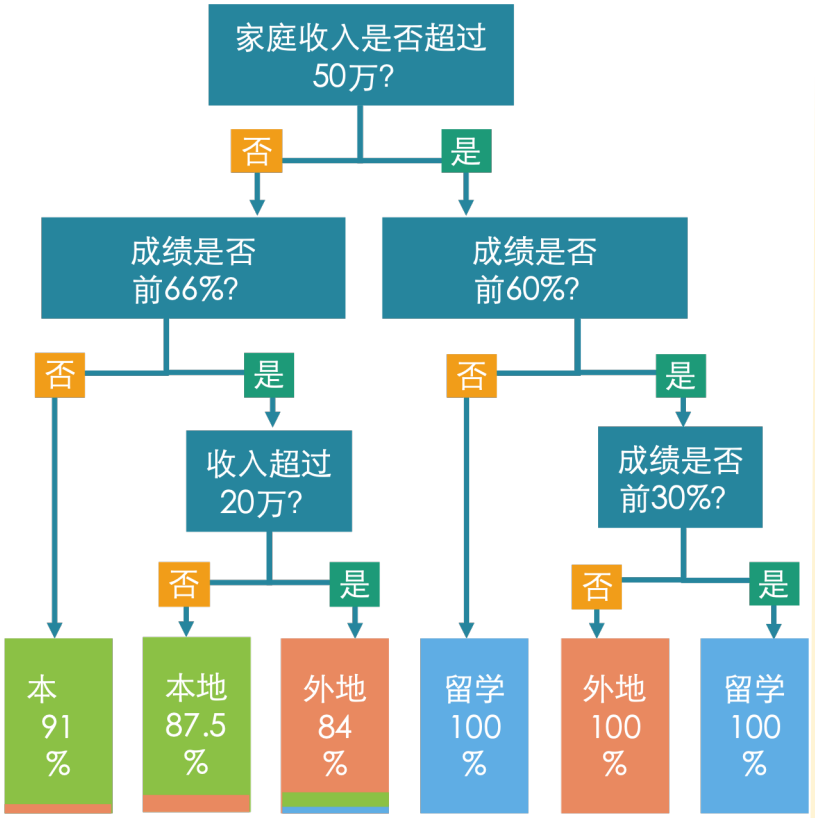
决策树产生的分类空间

Decision Tree

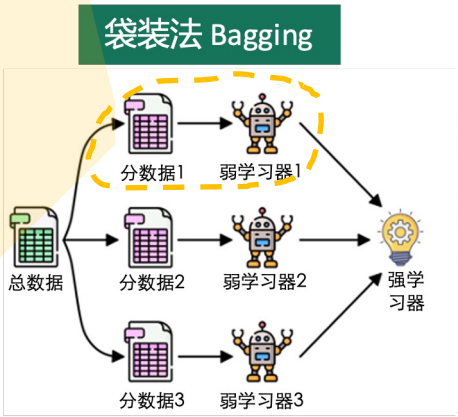


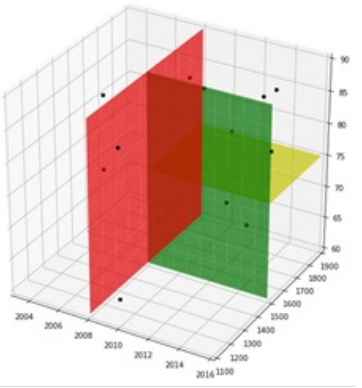
决策树与集成算法

算法生成的决策树



- 决策树的深度可以一直增加，但会产生过度拟合问题，因此必须有“停止分支”的标准，比如“熵小于X就停止”，这叫做“剪枝”(pruning)。
- 本例子是二维空间，决策树也可以在高维进行。
- 决策树可以“叠加”。例子中我们还有917个数据，我们可以拆分成多个样本，分别做成决策树，最后叠加决策空间。这种方法便是袋装法。

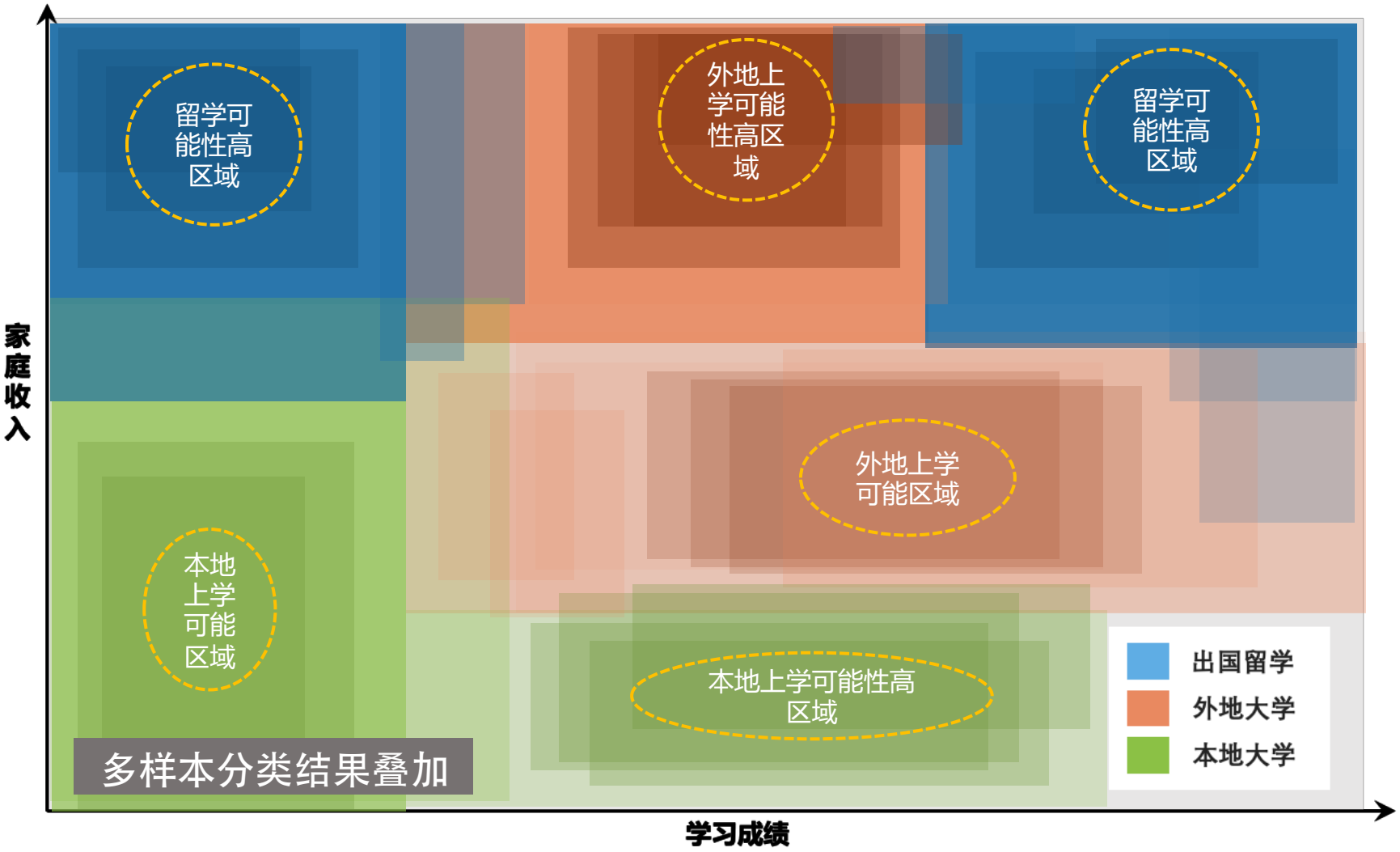




决策树集成（袋装法）

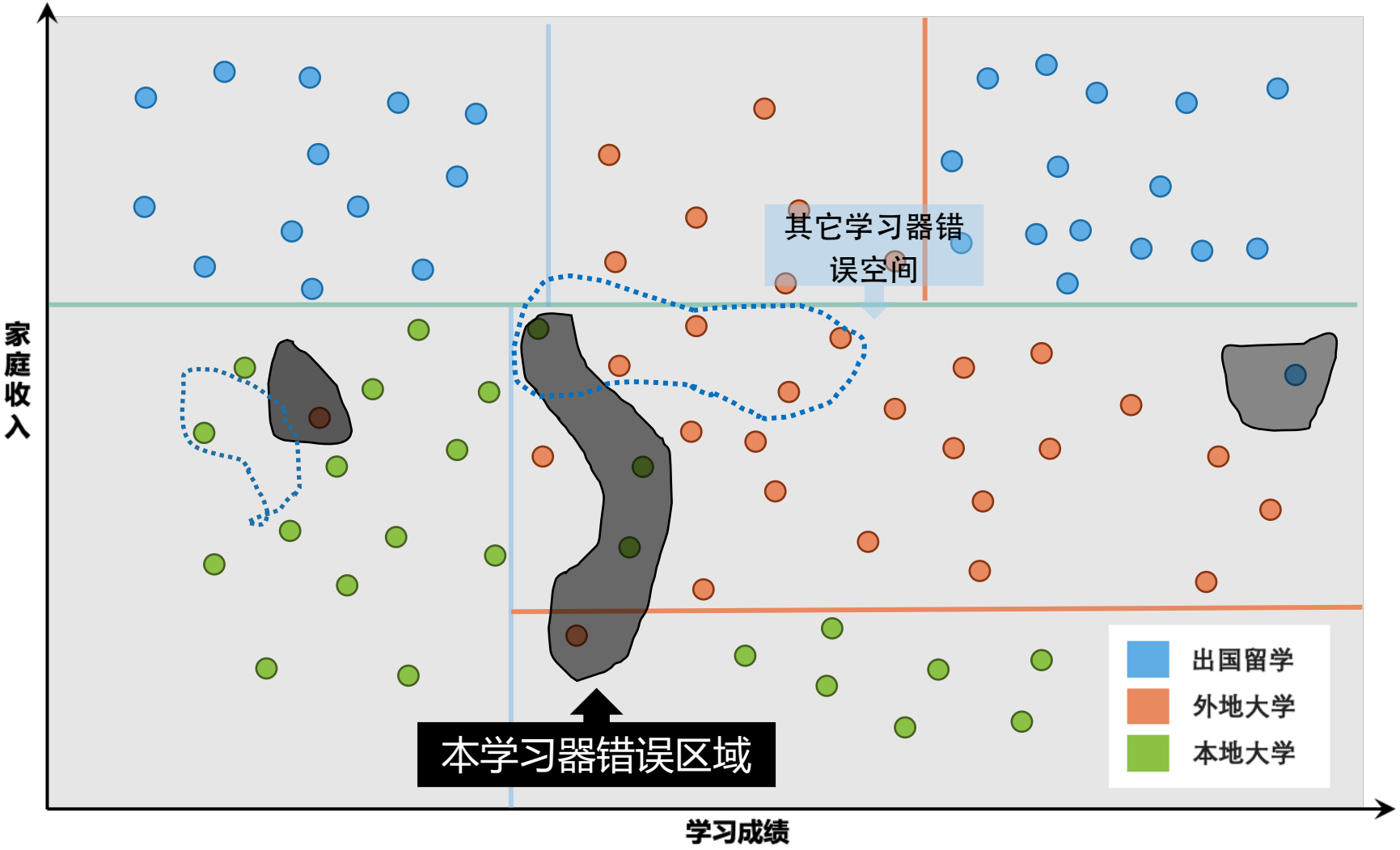
袋装法形成的决策空间

Decision Tree and Bagging



错误空间

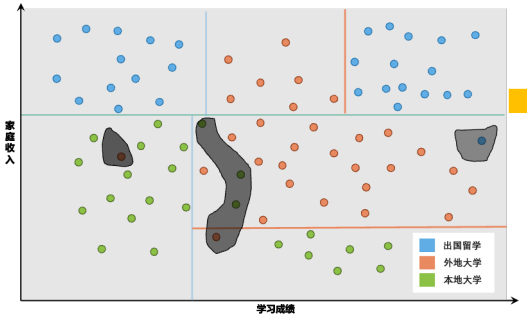
袋装法有效的基本基本原因注意是不同学习器的错误空间不完全叠加



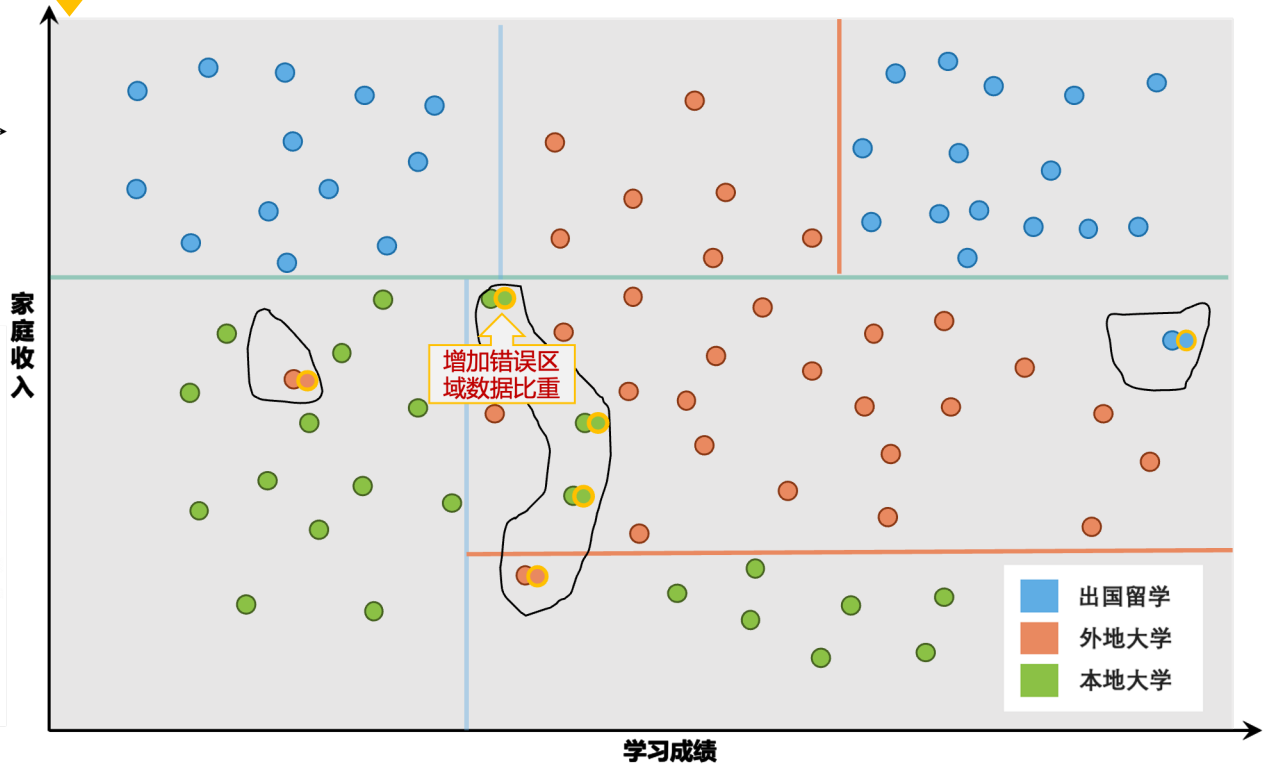
提振法

继续分割，决策树第三层

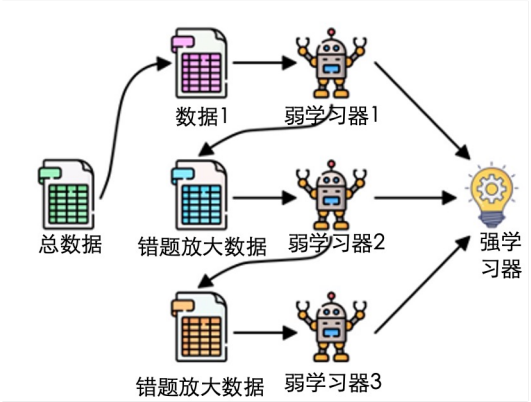
Boosting



提振法: 提高错误样本比重后再训练



提振法 Bagging



决策树和集成算法应用

微软XBOX 360 Kinect

Boosting

XBOX 360 Kinect

使用集成版本的决策树（随机森林）识别人和人的动作，并在此基础上革新了电子游戏行业。

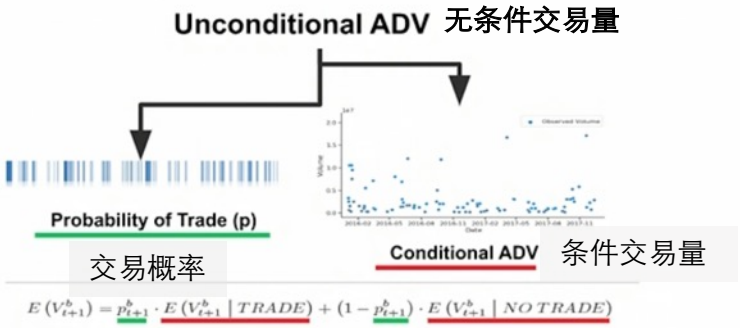
决策树和集成算法应用

贝莱德（BlackRock）交易量预测模型使用随机森林技术

Boosting

Volume Forecast Model – Overview

交易量预测模型–总论



The Volume Forecast Model is a single security predictive model of tradeable volume for corporate bonds globally.

Problem is decomposed into two sub-problems,

- the probability of a trade occurring
- the predicted volume if the bond does trade.

交易量预测模型预测全球范围内单个债券的交易量预测问题由两个部分组成：

- 交易出现的概率
- 出现交易后交易量

Random forest regression, known for dealing well with non-linearity and missing data.

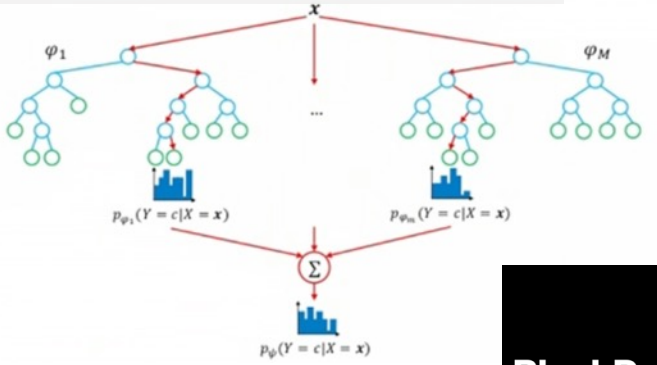
This as well allows us to include a large feature space to capture well the heterogeneity of bond markets.

Overall, the model performs very well

总体来看，本模型表现非常优秀

随机森林回归是著名的善于处理非线性和缺失数据的模型。

该模型帮助我们使用大的特征空间去抓住债券市场的独特性。



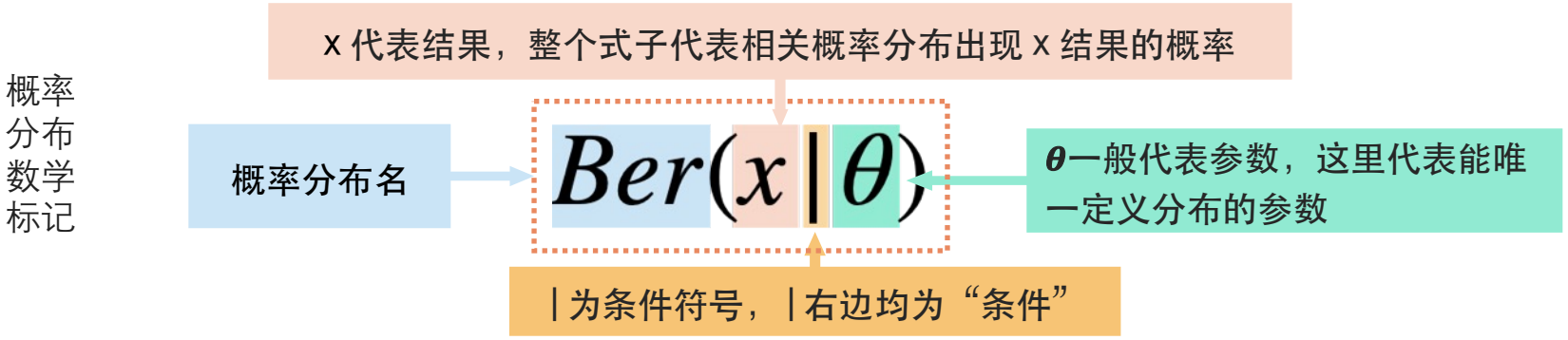
BlackRock


应用讨论

概率的常用数学标记法及意义

贝叶斯统计的一般数学标记法和意义

Notations and Bayes' Theorem



- Ber() 伯努利分布是一种两有种结果的概率分布，若伯努利实验成功，则伯努利结果取值为1。若伯努利实验失败，则伯努利结果取值为0。其参数$\boldsymbol{\theta}$ = （p,q），成功概率为p，失败概率为為q=1-p。
- 丢硬币为一种特殊的伯努利分布，其参数$\boldsymbol{\theta}$为（0.5,0.5），若视抛到正面为成功，则：

$$Ber(x=1|\theta=[0.5,0.5]) = p^{\delta_{x,1}}q^{\delta_{x,0}} = (0.5)^1(0.5)^0$$

$\delta_{a,b}$为克罗内克方程，当a=b时为1，否则为 0

贝叶斯统计的常用数学标记法及意义

后验、先验、似然、数据、最大后验

Notations and Bayes' Theorem

似然概率:
这组数据是由参数为θ的模型产生的概率

后验概率: ★
这组数据对应的是这个参数的概率

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

θ：参数（向量）
P(θ)参数概率，意思是出现这个参数的概率

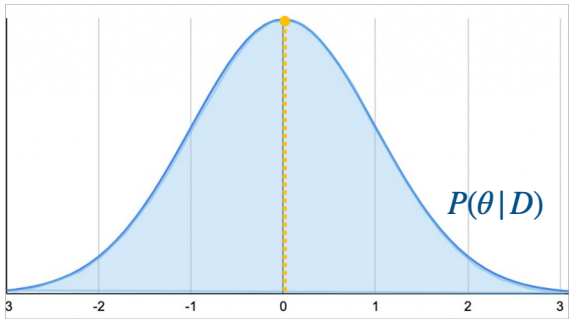
D：数据（矩阵）
P(D)证据概率，代表出现这组数据的概率

$$P(\theta = 0|D) = 0.4$$

点后验概率
后验分布点概率中最大的的点叫做最大后验（MAP）点

$$P(\theta|D)$$

后验概率分布



MAP θ=0

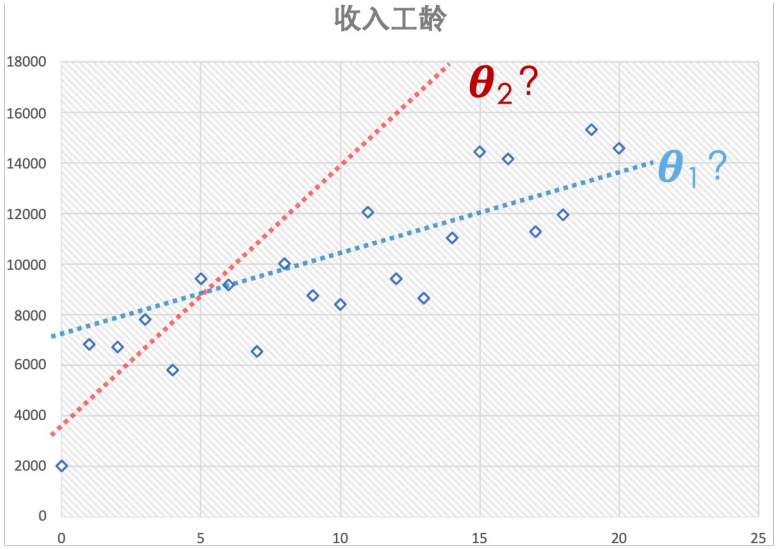
贝叶斯概率视角下的线性回归

四种概率

A Bayesian View of Linear Regression

例子：研究一家公司工龄与工资之间的线性关系

- θ 参数向量，可以看成一种薪酬体系，应包括：
  - a 斜率，每年工资上涨数 *(例：每年涨500)*
  - b 截距，起始工资 *(例：起薪2000)*
- D 数据，代表公司所有员工的工龄(x)工资(y)。
- $y = ax + b + \epsilon$



- P(D)：同行业公司里员工工龄-收入分布是D的占比。
- P(θ)：同行业公司里采用θ薪酬体系的占比。
- P($D|\theta$)：θ薪酬体系下，员工工资的情况是D的可能性有多大。

  回归线解释：回归线为θ时，出现D数据的概率。

- P($\theta|D$)：**员工工资的情况是D，薪酬体系是θ可能性有多大。**

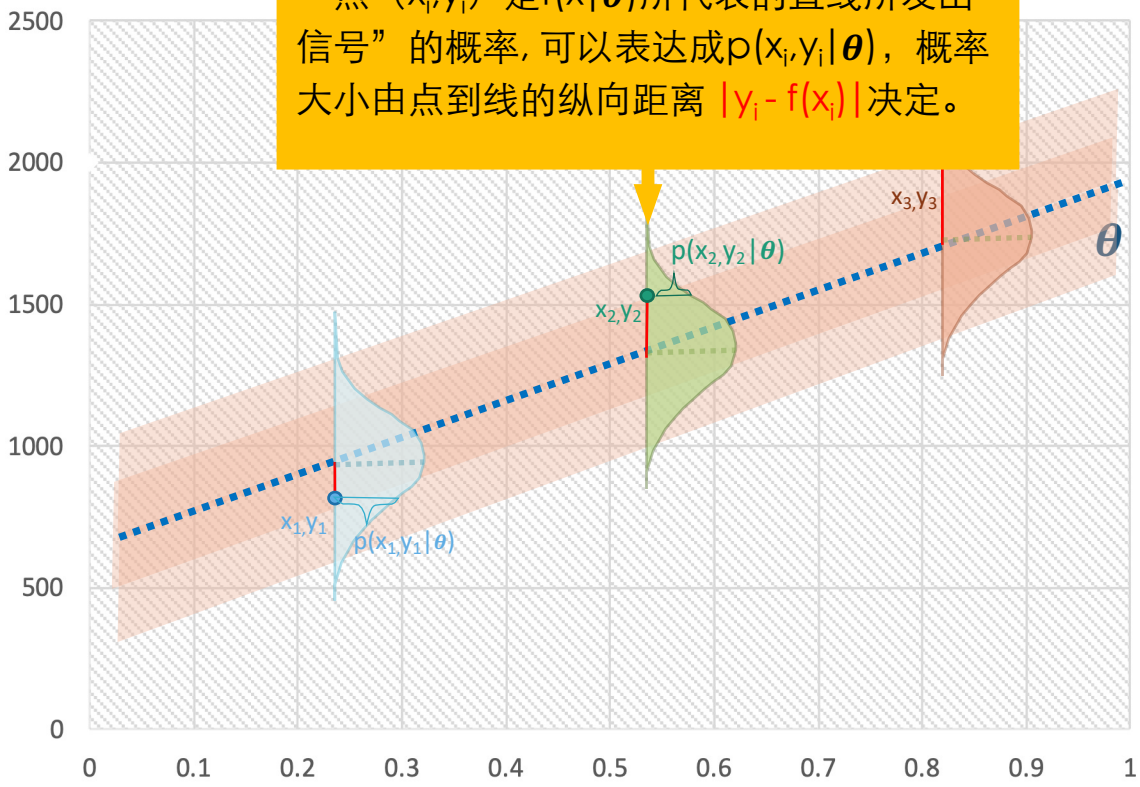
  回归线解释：D数据是由θ线产生的概率。

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

贝叶斯概率视角下的线性回归

线性回归的似然部分

A Bayesian View of Linear Regression



传统线性回归只找最大化似然$\boldsymbol{\theta}$

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

$P(D|\boldsymbol{\theta})$ 是回归线系数为$\boldsymbol{\theta}$时出现D 数据的概率

根据概率公式可知，整个数据集的概率等于单个数据出现概率的积，因此有：

$$P(D|\theta) = \prod_{i=0}^{n} p(y_i|x_i,\theta)$$

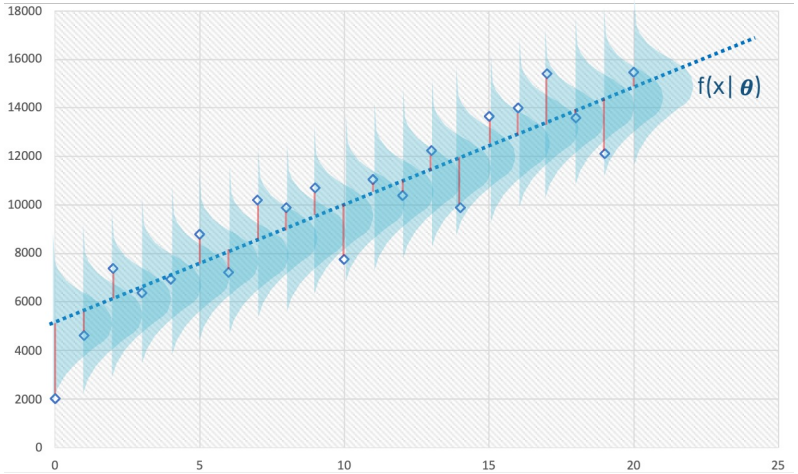
线性回归

传统（频率派）线性回的目标是最大化似然概率

Linear Regression

$$P(D|\theta) = \prod_{i=0}^{n} p(y_i|x_i,\theta)$$

$$log\left(P\left(D|\theta\right)\right) = \sum_{i=0}^{n} log\left(p(y_i|x_i,\theta)\right)$$



机器学习最常用的损失方程为**负对数似然概率**（Negative Log Likelihood, NLL）

$$NLL = -log\left(P\left(D|\theta\right)\right) = -\sum_{i=0}^{n} log\left(p(y_i|x_i,\theta)\right)$$

如果数据噪音为高斯分布*，则

$$NLL = -log\left(P\left(D|\theta\right)\right) = -\sum_{i=0}^{n} log\left(\frac{1}{2\pi\sigma^{1/2}} exp\frac{\left(y_i - f\left(x_i\right)\right)^2}{-2\sigma^2}\right) = \sum_{i=0}^{n}\frac{\left(y_i - f\left(x_i\right)\right)^2}{2\sigma^2} + n\frac{log\left(2\pi\sigma\right)}{2}$$

---

* 高斯分布公式为 $\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$

线性回归

传统（频率派）线性回的目标是最大化似然概率

A Bayesian View of Linear Regression

- 通过调整$\boldsymbol{\theta}$最小化 NLL，等价于最大化P(D|$\boldsymbol{\theta}$)，或者说调整$\boldsymbol{\theta}$使得f(x|$\boldsymbol{\theta}$)成为最可能发出所有数据信号x，y的那条线。

$$NLL = \sum_{i=0}^{n} \frac{\left(y_i - f\left(x_i\right)\right)^2}{2\sigma^2} + n\frac{log\left(2\pi\sigma\right)}{2}$$

常数，求极值时可以不考虑

- 最小化 NLL等同于最小化 $\sum_{i=0}^{n}\left(y_i - f\left(x_i\right)\right)^2$

因为有平方（二乘），所以叫最小二乘法

$$\sum_{i=0}^{n}\left(y_i - f\left(x_i\right)\right)^2 = \sum_{i=0}^{n}\left(y_i^2 - 2y_i(ax_i + b) + (ax_i + b)^2\right) = \sum_{i=0}^{n} x_i^2a^2 + 2x_i(b - yi)a + (b^2 - 2y_i + y^2)$$

如看成以A为自变量的一元二次方程，有极值公式

- 使上式最小化的 $a^* = \frac{\sum_{i=0}^{n} x_i(y_i - b)}{\sum_{i=0}^{n} x_i^2}$

- 中心化x，y为$x' = x - \bar{x}, y' = y - \bar{y}$ 则有 $b' = 0$（中心化后截距为零）

$$a' = \frac{\sum_{i=0}^{n} x_i'y_i'}{\sum_{i=0}^{n} x_i'^2}$$

$$a^* = a' = \frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=0}^{n}(x_i - \bar{x})^2}$$

（中心化后斜率不变）

一般线性回归斜率公式

将a*带回式子即可求得b*

- 因为$\boldsymbol{\theta}$=(a*,b*)使得似然概率P(D|$\boldsymbol{\theta}$)最大化了，我们称其为最大似然（max likelihood, MLE）参数，记做$\boldsymbol{\theta}_{MLE}$

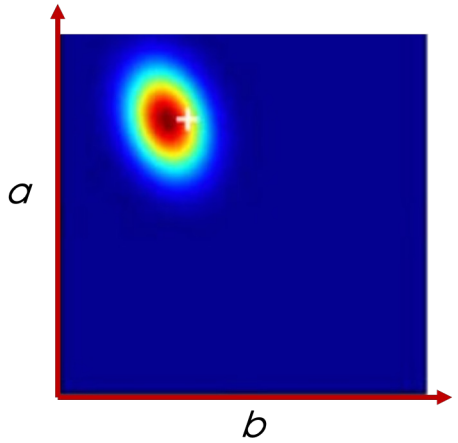
贝叶斯线性回归

- 由于P(D)常常无法直接观察，贝叶斯公式在应用时常变为：

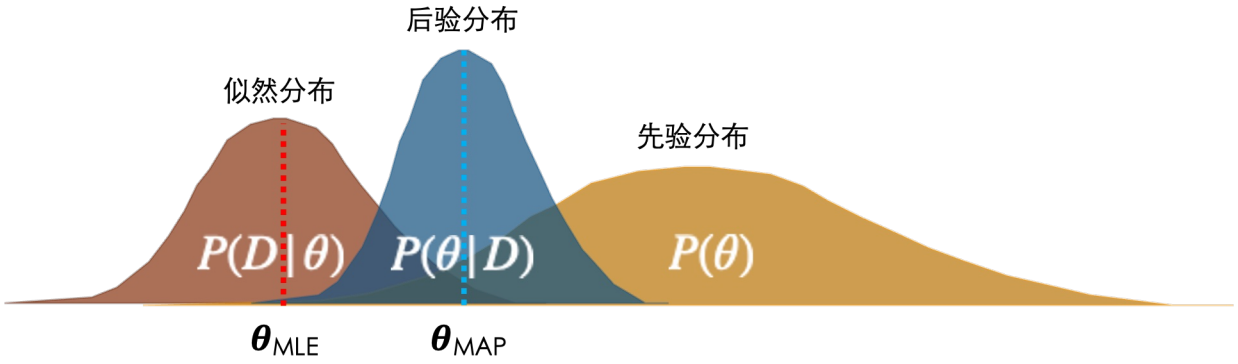
$$P(\theta|D)=\frac{P(D|\theta)P(\theta)}{P(D)}=\frac{P(D|\theta)P(\theta)}{\int P(D|\theta)P(\theta)\,d\theta}\ \text{或}\ \frac{P(D|\theta)P(\theta)}{\sum_{\theta}P(D|\theta)P(\theta)}$$

参数空间连续　　参数空间离散

- 对贝叶斯线性回归而言，我们已经知道似然部分的计算方法，而对于$\boldsymbol{\theta}$的分布，在有相关分布数据时我们参照该该数据，在没有的情况下，一般认为P($\boldsymbol{\theta}$)服从某高斯分布。
- 随后可以计算贝叶斯参数后验分布的情况，包括其最高点$\boldsymbol{\theta}_{\text{MAP}}$。
- 如P($\boldsymbol{\theta}$)不为高斯、积分不好计算、则必须借助MCMC等抽样法
- 似然、先验、后验分布，以及$\boldsymbol{\theta}_{\text{MLE}}$，$\boldsymbol{\theta}_{\text{MAP}}$ 的关系如下：



贝叶斯线性回归的后验参数是一个概率空间范围



贝叶斯线性回归

贝叶斯线性回归的计算复杂度远高于一般线性回归，什么时候有必要用贝叶斯线性回归?

- 数据样本比较缺失，而先验比较充分的情况下，贝叶斯线性回归能够比一般线性回归更为准确。
- 例如疫情统计：世界疫情的各种相关统计可能比较充分，而本地疫情刚刚开始，数据较少，但似乎和世界一般情况不同，这个时候基于本地数据的一般线性回归可能得出过激的判断，而贝叶斯线性回归可以借助“先验的统计力量”，给出更合理的判断。