报名

1. 注册华为账号
2. 点击比赛网址，创建队伍并按照要求提供团队信息
3. 报名：广告-信息流跨域CTR预估
4. 下载比赛数据，共包含四个csv文件

赛题理解

1. **什么是CTR：**

   Click-Through-Rate 点击通过率，指网络广告的点击到达率

   $$CTR = \frac{\text{广告的实际点击次数}}{\text{广告的展现量}}$$

2. **为什么要做CTR预测：**

   在广告实际投放情况中，需要综合考虑CTR以及平均点击价格ACP，决定广告的投放量

3. **什么是比赛数据中的目标域和源域数据：**

   目标域：收集广告任务产生的历史数据，但是用户行为数据稀疏，行为类型相对单一（用户个人信息、用户所点击的广告信息等）

目标域用户行为数据

字段名称	字段含义	是否可为空	字段类型	取值样例
label	是否点击，0：否，1：是	否	int	0，1
user_id	用户id	否	String	1，2...
age	年龄	是	String	1，2，3...
gender	性别	是	String	1，2...
residence	常住地-省份	是	String	1，2...
city	常住地-市-编号	是	String	1，2...
city_rank	常住地-市-等级	是	String	1，2...

源域：同一媒体的跨域数据，可以通过同一广告用户在其他域的行为数据，深度挖掘用户兴趣，丰富用户行为特征

源域用户行为数据

字段名称	字段含义	是否可为空	字段类型	取值样例
u_userId	用户标识	否	String	0001
u_phonePrice	用户手机价格	是	String	13
u_browserLifeCycle	浏览器用户活跃度	是	String	10
u_browserMode	浏览器业务类型	是	String	11
u_feedLifeCycle	信息流用户活跃度	是	String	12
u_refreshTimes	信息流日均有效刷新次数	是	String	16
u_newsCatInterests	信息流图文 点击 分类偏好	是	[String,]	[1^2...]

4. **赛题任务：**

   根据历史的用户行为特征（两个域的数据），预测是否会点击广告（二分类问题）

   不可以使用穿越信息（T时刻样本使用T时刻之前的信息，不能使用T时刻未来的信息）

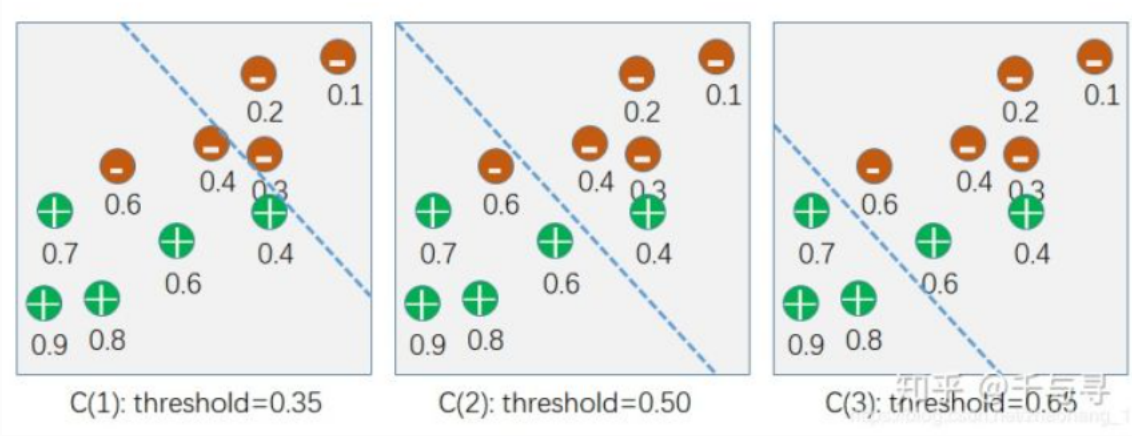
5. **评价指标：**

   $xAUC = \alpha * GAUC + \beta * AUC$，越高越优（初赛$\alpha = 0.7, \beta = 0.3$）

   下面结合**这篇文章**，介绍一下ROC、AUC、GAUC

(a) **ROC前身：通用的对分类模型的评价**

在对每一个样本计算出一个预测概率后，若选择不同的阈值，我们会得到不同的分类结果

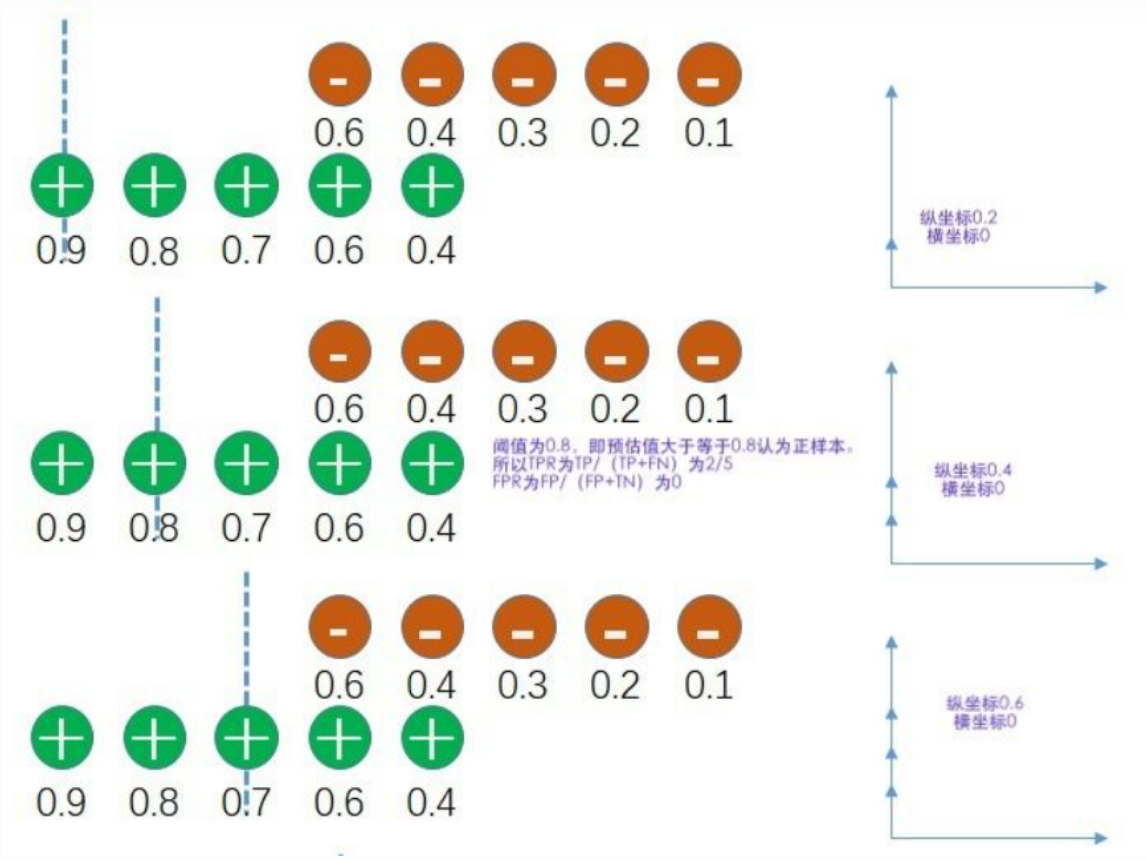


(b) **ROC曲线：**

在上面的情况中，我们可以遍历所有的阈值，查看每一个阈值下的分类情况，并绘制ROC曲线

纵坐标：$TPR = \frac{TP}{TP+FN}$

横坐标：$FPR = \frac{FP}{FP+TN}$



(c) **AUC：**

在绘制出ROC曲线后，我们就可以计算出AUC的值

$AUC = \frac{ROC曲线下面积}{在x*y区域面积}$

将上面公式的积分过程拆开，我们可以用另一个角度计算AUC

$AUC = \frac{\sum 每个预测为正的样本能比多少负样本大}{正样本数*负样本数}$

第一节中，原始有五个正样本：

p=0.9的真实正样本，它在所有5个负样本前面，因此记为5

p=0.8的真实正样本，它在所有5个负样本前面，因此记为5

p=0.7的真实正样本，它在所有5个负样本前面，因此记为5

p=0.6的真实正样本，它在4个负样本前面，因此记为4

p=0.4的真实正样本，它在3个负样本前面，因此记为3

交叉区域记为5*5=25

因此最终的AUC记为

$$AUC = \frac{5+5+5+4+3}{5*5} = 0.88$$

(d) **GAUC：**

在广告推荐领域虽然仍是二分类模型，但是是更细粒度的二分类（对每个人进行二分类），因此传统的粗粒度的AUC并不适用

GAUC其实是计算每一个用户的AUC，然后加权平均，这样可以减少不同用户之间不好比较的影响（因为不同用户之间对于广告偏好差距较大）

实际处理时权重一般可以设为每个用户view或click的次数，而且会过滤掉单个用户全是正样本或负样本的情况