因果系列

现状

因果关系整合到AI当中是目前ML领域的一个热门分支。11年的图灵奖得主的Judea Pearl则提到：“目前有太多深度学习项目都单纯关注缺少因果关系的粗糙关联性，这常常导致深度学习系统在真实条件下（明显不同于训练场景的条件下）进行测试时，往往拿不出良好的实际表现。”并在他的新书《The Book of Why: The New Science of Cause and Effect》当中提到，“如果没有对因果关系的推理能力，AI的发展将从根本上受到限制。”

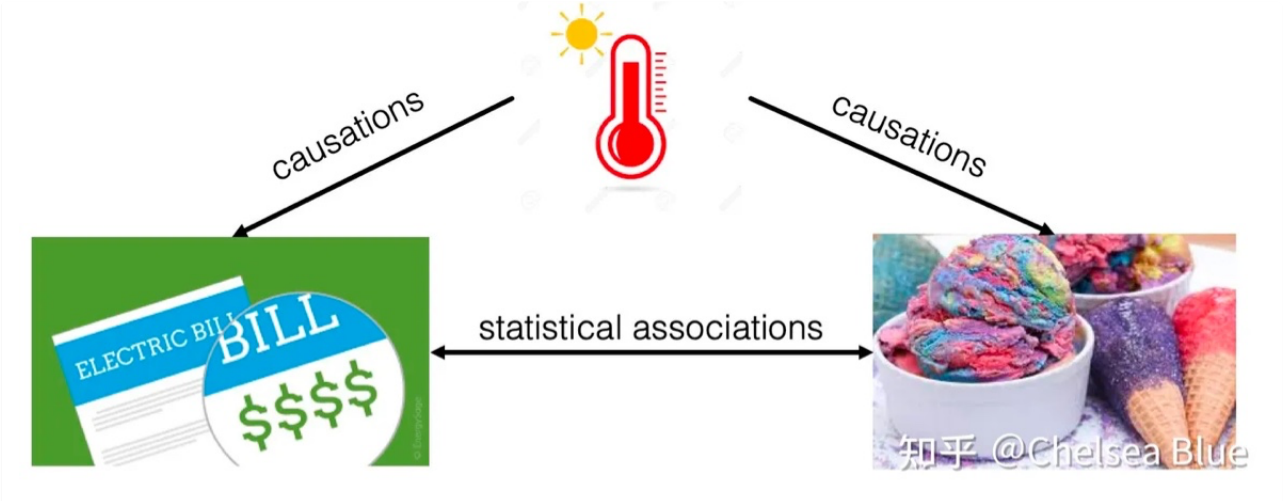
需要注明的是，传统因果很多方法都是通过做实验，即controlled experiment来得到结果，但是这种实验成本往往巨大，即便在体量相当大的公司，做一些商业上的ABtest也意味巨大的成本，在这个情况下，我们希望从观察到的数据，即observational data中得到因果推断的结果。

什么是causality (因果)

Formal Definition: Causality is a generic relationship between an effect and the cause that gives rise to it.

causality和statistical association的区别（因果性和相关性区别）

举个例子：如果我们发现，夏天的时候，一个冰淇淋店的电费上涨的同时冰淇淋卖的也很好，我们可以说他们互相之间有因果的关系吗？不见得，他们之间可能只是统计上的相关性，而真正给他们带来的因果性的因子叫做气温，是因为气温的上升，导致电费的增加，也是因为气温的上升，导致了冰淇淋销量的上升。这个例子很好的向我们阐述了因果性和统计相关性的区别。

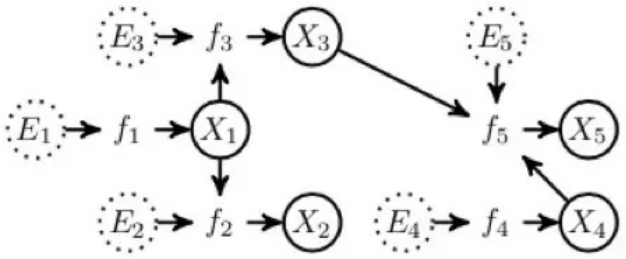


Two main questions

1. **Causal discovery** (因果关系挖掘)：比如研究：温度升高是否是电费增加的原因？或者在商品价格，商品转化率，商品上市时间，商品成本等几个变量之间探究一个因果图，即变量两两之间是否有因果关系？如果有，谁是因谁是果？
2. **Causal effect** (因果效应推理)：比如我们已经知道温度升高是电费增加的原因，我们想知道温度从20度升至30度，会对电费带来多少增加？

Two main frameworks

1. **Structural Causal Models**(SCM) – Judea Pearl: A causal model by SCMs consists of two components: the causal graph (causal diagram) and the structural equations. 即我们需要先得到一张因果图，然后对于因果图，我们去使用Structural Equations来描述它。f(X|E) ⇐ f(X|do(E))



$$\begin{cases} E_i & \sim \mathcal{E} \\ X_1 & = f_1(E_1) \\ X_2 & = f_2(X_1, E_2) \\ X_3 & = f_3(X_1, E_3) \\ X_4 & = f_4(E_4) \\ X_5 & = f_5(X_3, X_4, E_5) \end{cases}$$

箭头由因指向果，X和E都是变量。然后右边的一系列方程就是Structural Equations来描述这个图，每一个方程f都表示着由因到果的一个映射或者说一个表达式，这个方程可以是linear也可以是nonlinear的，取决于他们的因果关系是否线性。

- Pearl提出**小图灵测试**是实现真正智能的必要条件（机器如何迅速访问必要信息、正确回答问题，输出因果知识）。并提出**因果推理引擎**，以假设（图模型）、数据和Query输入，输出Estimand（基于do-calculus判断query是否可识别）、Estimate（概率估计）和Fit Indices（评估）。
- 其中do-calculus是判断因果问题是否可解的前提，原理就是贝叶斯网络中**D-seperation**（图分离与概率独立等价条件，参考PRML）
- **一般回答反事实问题需要SCM模型**，由图模型（表示因果知识）、反事实和干预逻辑（形式化问题）和结构方程（链接因果知识【图模型】和因果问题【反事实和干预逻辑】的语义）组成。**一般步骤**为 1. abduction（基于现有事实分布【先验】p(u|e)更新图概率p(u)） 2. action（基于结构方程更新x） 3. prediction（预测反事实）

2. **Potential Outcome Framework**– Donald Rubin: It is mainly applied to learning causal effect as it corresponds to a given treatment-outcome pair (D,Y). 简单来说，计算因果效应最直接的手段就是控制住所有的变量不变，只变化cause，比如把温度从20变到30度，然后直接看outcome变化，也就是直接用30度时的电费减去20度时的电费，既可以得到causal effect。

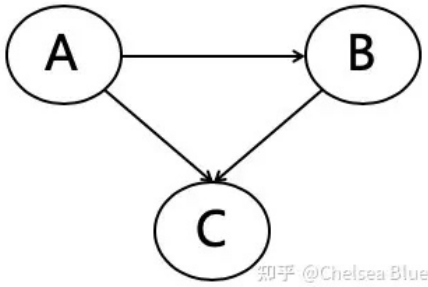
   基于**Potential Outcome Framework**，更加简单直观，统计和社科用的多。它**设想与观测相悖情况，是一种反事实因果**，被称为Experimental causality（但其一般回答干预层的问题）。因果**分析步骤**主要有 1. 定义问题构建粗粒化因果图 2. Do-Calculus（干预）基于概率计算效应。

   如果这个世界有两个平行时空，那么我们可以做这个实验，但是如果没有呢？我们知道温度不可能在同一个地方，同一个时间，既20度又30度，那么必然20度的时候，30度时的电费就叫potential outcome。而这个framework，就是想法设法从能观察到的数据中得到这个potential的结果，然后二者相减，就是我们想要的答案啦！

Causal Discovery

首先我们从贝叶斯网络聊起，贝叶斯网络是一个DAG(directed acyclic graph)，即有向无环网络。然后我们可以把它当作一个概率图，也就是可以概率表达它。举个例子：对于下图，我们可以表达为P(A,B) = P(B|A)*P(A)，why?，因为A指向B，而无箭头指向A，我们就可以得到A和B的联合分布

加一点难度，对于下图，可以表达为P(A,B,C) = P(C|A,B)*P(B|A)*P(A)。因为C有AB两个变量指向他，而B同样只有A指向它。



因果效应估计

套用一个发券和购买转化率的关系，已知发优惠券与购买转化率有因果关系，发优惠券是因，购买转化率是果，我们想知道，当发券的情况下，购买转化率会增加多少？这个问题就是一个典型的因果效应估计的问题。

因果效应的估计其实非常的广，细分了很多领域，比如对ITE的估计，对ATE的估计，其中根据因果类型和数据特性，又有更多分类，比如对于连续性treatment的估计，对multi-cause的估计，对time-varying treatment的估计等等。这里我们主要focus在ATE和ITE的两种估计。

定义什么是ATE?什么是ITE?

先说一下什么是treatment，就是我们感兴趣的那个因，比如我们研究温度升高一度对电费的影响，那么温度不变（t=0），温度升高一度（t=1）；比如我们研究给用户发优惠券对购买转化率的影响，那么不发券（t=0），发券（t=1）。

现在假设我们的treatment有两种 $t \in \{0, 1\}$，对于一个instance i，比如一个用户，他的转化率y是果，是否发券的t是因，y^t表示在treatment为t的情况下转化率y的值。我们就有 **ITE (Individual treatment effect)** 公式如下：

$$\mathrm{ITE}_i = \tau_i = y_i^1 - y_i^0$$

表示一个individual的treatment effect。那么如果我们想看一个大群体（一个普遍现象），就是ATE(average treatment effect)，ATE is the expectation of ITE over the whole population i=1,...,n：

$$\mathrm{ATE} = E_i[\tau_i] = E_i[y_i^1 - y_i^0] = \frac{1}{n}\sum_{i=1}^{n}(y_i^1 - y_i^0)$$

那么介于两者中间呢，就有一个 **CATE(conditional average treatment effect)**，也就是一个subpopulation的average treatment effect。

$$\mathrm{CATE} = \tau(X) = E_{i:x_i \in X}[\tau_i]$$

其实ITE就是CATE的变种，只不过这个subpopulation缩小到了一个人。

估计ATE

估计ATE的作用是做一些宏观的决策，或者对于整体population是否施加treatment做一些决策。举个例子，我们想要评估打疫苗对病变的效果，我们要评估一个overall的疫苗效应，这个时候我们去预估ATE就够了。

评估效果的时候，我们需要一个ground truth的ATE τ 以及我们infer出来的ATE $\hat{\tau}$，我们评估的指标就是MAE：

$$\epsilon_{MAE\_ATE} = |\tau - \hat{\tau}|$$

估计ITE (CATE)

那么什么时候我们需要估计ITE呢？当整个population是heterogeneous的时候，即人群有异质性的时候，ATE可能会误导结论。举个例子，我们衡量大众点评评分对餐馆的销量影响的时候，ATE可能会误导，因为大城市的餐馆可能会更多被大众点评影响，小城市或农村可能影响更小。这时候其实我们要评估的每一个subpopulation的ATE，也即CATE（或者细粒度到每个individual的ITE）。那么我们怎么去定义各个subpopulation呢？就是靠除了treatment t之外的其他特征X，每一组X的取值就代表了一个subpopulation。

评估效果的时候，我们可能需要AB的环境，对于某个样本做treatment=0和1的两次实验得到结果 y^0和 y^1，然后我们算一个MSE：

$$\epsilon_{PEHE} = \frac{1}{n}\sum_{i=1}^{n}(y_i^1 - y_i^0 - \hat{\tau}(x_i))^2$$

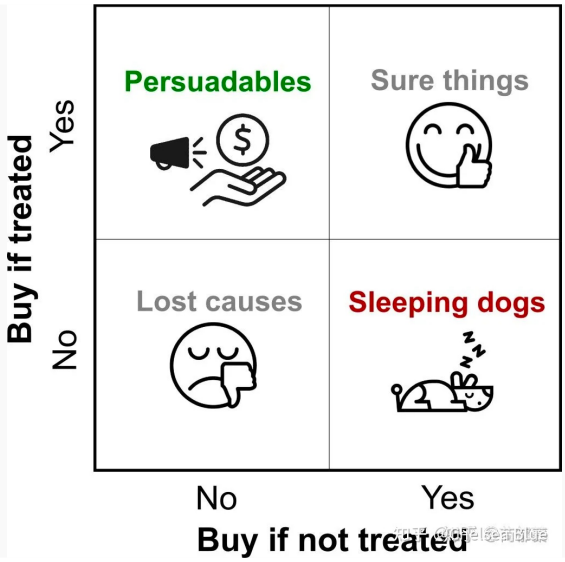
一个重要假设

SCM这个结构中有一个重要假设叫sufficiency assumption，即我们没有unobserved confounder，confounder就是同时对t和y都有因果影响的变量，这里要求所有的confounder都在我们的数据特征X中。所以前期的很多方法都需要满足这个假设，不过大家也知道这个条件其实在现实生活中的假设是很难被满足的，这时候我们就会有一些方法可以relax这个假设。所以我们也可以把方法依据是否能在有unobserved confounder的情况下使用分成两类。

估计uplift

ITE的估计或者说CATE的估计本质上就是目前现阶段在业界最火的uplift模型的估计。

uplift模型



先放一张任何 uplift 都会文章都会放的图来解释 uplift。

这里我们可以把 treat 当作给他优惠券，我们一共有四种人，左上角是给券买，不给就不买；右上角是给不给券都买（不价敏）；左下角是给不给券都不买；右下角是给券不买，不给券就买（反人类）。我们希望找到图里左上角那部分人（**即给券就买，不给就不买**），这里 treatment 就是给券，我们自然避免把钱花在其他三种人上。

我们希望找到这种对补贴十分敏感的人。于是我们就需要去评价一个人对于给券比不给券带来的额外购买意愿。**注意！！！一般的预测模型是预测这个人给券后的购买概率，这不是我们要的，我们要的是这个人给券比不给券购买概率的增量**，那么问题来了，你不可能在一个世界里既给他券又不给他券（除非这个人有分身或者有平行世界）。那怎么办？这时我们就用到了我们的 uplift 概念：the effect of an action on some customer outcome.

uplift 与因果

因果科学

因果科学是研究因果关系或回答因果问题的学科。现代因果定义为在保证其他因素不变，改变 X 引起 Y 的改变，则 X 为 Y 的一个原因。

简单从历史上看，因果研究一般分为 3 类

- 物理学定义因果（最清晰定义的因果）
- 物理系统演化动力学，**基于时间讨论因果**
- 哲学定义因果
- Type Causality：某原因导致什么结果，由因推果，**干预主义思想**（因果效应定义），用来帮助**预测**
- Actual Causality：关注事物发生的原因，由果推因，与**反事实思维相关**。【反事实：主要是研究“若非 (but for)”，“若非“过去 A 事件发生，结果事件 B 可能不发生，常用于因果检测。

统计中因果推断

从现实数据中提取出某些变量间的因果关系，**主要就是 Judea Pearl 的因果信息革命。其中也囊括了目前最火热的机器学习方法。他提出的因果关系之梯，根据因果问题的可答性，对比了目前的机器学习 (深度学习) 和因果推断区别。**

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Figure 1: The Causal Hierarchy. Questions at level i can only be answered if information from level i or higher is available.

这里他把任务分为递进的3类，**association**（what is）-> **intervention**（what if）-> **counterfactuals**（retrospection）。传统机器学习（关联层）是在问你what is?，即给定属性，问你是什么的概率； 而第二层（干预层）是在问what if？即，如果我对你做了什么，你会怎么样？终极（反事实层）的当然是回答哲学反事实问题，如果我当时那样做了，会怎么样？(唱："想回到过去，试着让故事继续"）这里层层递进，**高层可以回答低层的问题，反之则不行，因为不具备充足的信息。**

基于业务的目标我们可以发现，这里从哲学角度来看主要是一个type Causality类型的干预主义的预测问题，而非是一个反事实推理问题。 所以从分析框架上，我们很自然选择的是rubin的RCM在干预层面的分析工具，**所以uplift model是一个干预模型，且为RCM框架。**

我们来想想我们需要怎么得到这个uplift呢？这里就要求我们去estimate the difference between two outcomes that are mutually exclusive for an individual（即counterfactual）。当我们看到counterfactual就自然会想到因果！！！**我们需要通过因果推断，来给到一个反事实的预测，也就是一个what if**的问题。就是what if我们给他了券，那他的转化率购买意愿增量是多少？

uplift modelling目标是精确学习给定一个干预（发券）后，对结果（是否购买概率）增量。即它要建模出增量效应（发券对比不发券对结果的影响）。直白一点，用经济学语言就是，学习出边际效用，且好的模型能**最大化边际效用。**

uplift定义为：$u(x) = P(O = 1|T; x) - P(O = 0|\hat{T}; x)$，$T$为是否发券，$O = 1$为买单，$x$为属性。

干预问题和ML的相关问题最直接的区别就是，**ML监督模型都是由label的，而因果干预问题是缺失label的**（对于每一个人他只有发券或者不发券的结果【另一个是反事实的】），但是我们建模目标是干预（do操作），而非简单的相关（转化率）。我们希望模型能找到最多的可以被干预转化（发券购买）的用户（**我们希望看到用户行为的变化**）。这就类似强化学习RL中最大化reward，uplift model要最大化uplift（找到最有可能受到券激励消费的用户），RL中核心操作是决策action获得的额外reward。这里action对应do算子，reward对应uplift，所以uplift model和RL天然如出一辙。RL中学习过程通过E&E(explore & exploitation)来探索利用action，那uplift model最好也需要EE，即**随机试验，因为互联网产品都可以做ABtest，do可以作为ABtest分组，进而进行统计计算。**（传统医学因为随机试验可能存在伦理问题，比如分析吸烟是否导致肺癌，很难去随机强迫人吸烟，导致难度增加）。

heterogeneous treatment effects

我们T=1作为给券(treated)，T=0作为不给券(untreated)，Y作为最后的购买概率(outcome)。X是这个人的一些特征(feature)

ITE：individual treatment effect

$$\mathrm{ITE} = Y_i(T = 1) - Y_i(T = 0)$$

ATE：average treatment effect

$$\mathrm{ATE} = E[Y(T = 1) - Y(T = 0)]$$

CATE：conditional average treatment effect

$$\mathrm{CATE} = E[Y(T = 1) - Y(T = 0)|X = x]$$

我们注意到，我们想要的是某个个体给券不给券的区别，而不是总体给券不给券的区别，所以这里就不是研究ATE（average treatment effect），而是研究ITE（individual treatment effect）。而由于我们是使用observational data，ITE是CATE的一个特殊情况，我们其实研究的是CATE：an average treatment effect specific to a subgroup of subjects, where the subgroup is defined by subjects' feature。举个例子：如果我们的特征包括性别，年龄，职业，app活跃度。那我们其实想知道的是：**一个app活跃度高的28岁男性程序员给券和不给券对购买概率的差异。**

而这里CATE模糊来讲，就是heterogeneous treatment effects。因为他认为总的population是heterogeneous的，所以我们要通过X来区隔出一个个subpopulation。于是我们成功从uplift模型过渡到了一个研究heterogeneous treatment effects的问题 (为了后续行文方便，heterogeneous treatment effects我们有时候会写成treatment effects或者CATE或者ITE，都理解为一个东西就好)。

Problem Setting

我们的对象有一系列特征 X和treatment T（有的文章用 W），$T=1$时为treated，$T=0$ 时为untreated。 Y^0是 $T=0$ 时的outcome，Y^1是 $T=1$时的outcome我们的目标heterogeneous treatment effects（CATE）标记为 $\tau(x)=E[Y^1-Y^0|x]$ ，即given特征x，$T=1$时的outcome和 $T=0$时的outcome的差的期望值。

Uplift Modeling的常规方法

- **tree-based**：和机器学习中的树基本一样 (分裂规则、停止规则、剪枝），核心为分裂规则，根据每个叶子节点的uplift，计算分裂前后的信息差异。常用比如基于分布差异 (KL、Euclidean等）；或者结合bagging、boosting技术等
- **regression-based**
  - two-model：实验对照组分别一个模型，之后做差。可能实验对照各自学的好，但是缺失了我们要的uplift增益效应
  - one-model：把T（是否发券）作为特征加入 (其他模型T不作为特征），最终输出$u(x)=f(x,t=1)-f(x,t=0)$
- 聚类：从目标上我们想圈一部分uplift高的人，且无label指导，从这个角度可能可以借鉴思路

S-Learner

这个模型最简单，直接把treatment作为特征放进模型来预测（很像大学时候计量经济学里做的小作业），就是regression adjustment的方法。

首先我们把T作为特征一起放进机器学习模型的特征，Y是目标，然后训练一个有监督的模型 $\mu(x)=E[Y|X=x,T=t]$ 。然后我们改变 T的值，就可以得到两个不同的结果，再一相减就好了:

$\hat{\tau}(x)=\hat{\mu}(x,T=1)-\hat{\mu}(x,T=0)$

估计uplift-深度学习方法

（略）

uplift模型评估

如果要评价孰优孰劣，或者选取其中一个效果好的模型进行使用，就势必需要一个评估uplift模型的框架/方法。

评估uplift模型的难点主要是在于不像普通的分类或者回归模型，uplift模型没有真正的“ground truth”，比如我们看对一个人群发券或者不发券带来的转化率的变化这个uplift，他真正的ground truth应该是："平行时空A下这个人群发券的转化率"减去"平行时空B下这个人群发券的转化率"，然而很可惜的是平行时空不存在。所以我们该如何评估呢？

一个办法是直接上AB，通过AB业务关心的指标来得到结果。另一种方法就是离线评估，比如通过AUUC包括Qini系数来评估。

Problem Setting

我们假设问题定义是一个样本在发券后是否购买这个商品，即T与Y都是binary，$T\in\{0,1\}$，$Y\in\{0,1\}$ 。

X是我们的validation set，用来测试模型

u 是我们训练得到的uplift模型，$u(x)$就是对于样本 x预测出的uplift值

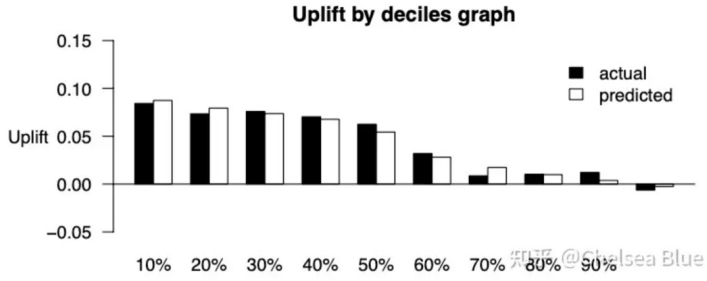
K是画图时用的bins的数量，即segment个数，当我们是deciles时，$K=10$

r_k^t,r_k^c 是k segment里treatment组中$Y=1$ 的样本个数和control组中 $Y=1$的样本个数

n_k^t,n_k^c是k segment里treatment组的样本个数和control组的样本个数

画图步骤

1. 对 X 中全部的 x 计算出其uplift $u(x)$
2. 把计算出来的 $u(x)$ 从高到低进行排序
3. 把这些 $u(x)$ 切成10份，并找到切分的边界 $b_0 \dots b_K, K = 10$
4. 计算每个segment k的predicted uplift u_{kp} ， $u_{kp} = \frac{1}{n_k^t + n_k^c} \sum_{x: b_{k-1} < u(x) < b_k} u(x)$,也就是均值
5. 计算每个segment k的actual uplift u_{ka} ， $u_{ka} = \frac{r_k^t}{n_k^t} - \frac{r_k^c}{n_k^c}$
6. 计算完每个segment的 u_{kp} 和 u_{ka} 后，我们就可以作图啦



结合图 - 一个简单但solid评估框架

1. Monotonicity of incremental gains：这个想说的是actual和predicted的uplift在单调性上是否一致，即是不是predicted uplift越大的segment，actual uplift同样也越大。也就是说，由于predicted uplift一定是单调下降的（因为我们是按这个大小排序的），actual uplift也应该是严格单调下降的。可以看到这个图上的模型基本满足，但是在20%，80%和90%这三个segment上不完全单调。
2. Tight validation：就是每个segment里，actual uplift和predicted uplift在数值上是否足够接近，即两根柱子是不是一样长，越一样越好。表明一个预测的准确性
3. Range of Predictions：就是最大的predicted uplift（最左那根柱子）和最小的predicted uplift（最右那根柱子），差距是否足够大。为什么要衡量这个呢？假设我们有一个模型，预测出来每个样本的uplift都是一样的，即使这个模型平均来看是比较准的，但是没有实际意义，因为我们实际上必须要求模型能够区分出uplift较大和较小的两群人，不然我们怎么发券？怎么投放？所以这里就要求模型能够在uplift上预测出足够的区分度。

更好量化的评估框架

主要问题定义同上面的评估框架，增加如下notations

π代表把 $u(x)$ 从大到小降序排的一个order，我们有 $u^{\pi}(x_i) > u^{\pi}(x_j), \forall i < j$

$\pi(k)$代表按照 π进行排序后的前 k 的 $u(x)$ 样本，对于$\pi(k)$ ，我们有 $u^{\pi}(x_l) \leq u^{\pi}(x_i), \forall l > k, i \leq k$

$R_{\pi(k)}$代表前 k个样本中 $Y = 1$的样本个数

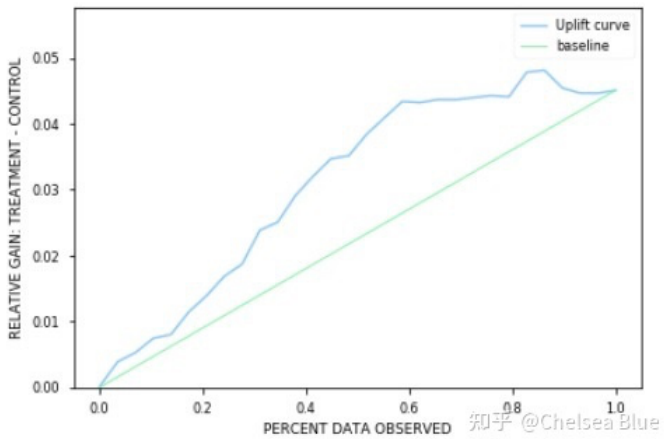
$R_{\pi}^T(k) = R_{\pi}(k)|T = 1, R_{\pi}^C(k) = R_{\pi}(k)|T = 0$ 代表前 k个样本中 $Y = 1$里在treatment和control组的个数

$\bar{R}^T(k) = k * E[Y|T = 1], \bar{R}^C(k) = k * E[Y|T = 0]$代表任意 k个样本中平均 $Y = 1$里在treatment和control组的个数

$N_{\pi}^T(k)$和$N_{\pi}^C(k)$是前 k个样本中在treatment和control组的个数

uplift curve & AUUC

横轴表示前 k/n%的observed样本，纵轴是这些observed样本中的treatment组中$Y = 1$ 的个数减去control组中 $Y = 1$的个数。按照上述提到的序$\pi(k)$，即把uplift降序排列，这条曲线 (uplift curve) 的纵轴就是$R_{\pi}^T(k) - R_{\pi}^C(k)$。而如果是任意抽样$k$个样本，这条曲线 (baseline) 的纵轴就是$\bar{R}^T(k) - \bar{R}^C(k)$。画图如下[3]：



这里uplift curve直观解释就是uplift最大的前 k 个样本里，treatment组中 $Y=1$的个数比control组中 $Y=1$的个数的差值。所以这个曲线最后一定会和baseline交汇，因为在全部样本下，uplift curve和baseline的计算结果必定相等。然后我们希望在k越小的地方，treatment组中$Y=1$的个数比control组中 $Y=1$ 的个数的差值越大，证明uplift大的样本确实是那些给treatment就更能转化的样本。

如果理解了上面的uplift curve和baseline怎么画的，AUUC就很简单了，就是两条线中间的面积，越大越好。明细公式如下：

$$\mathrm{AUUC}\pi(k)=\sum i=1^{k}(R_{\pi}^{T}(i)-R_{\pi}^{C}(i))-\frac{k}{2}(\bar{R}^{T}(k)-\bar{C}^{T}(k))$$