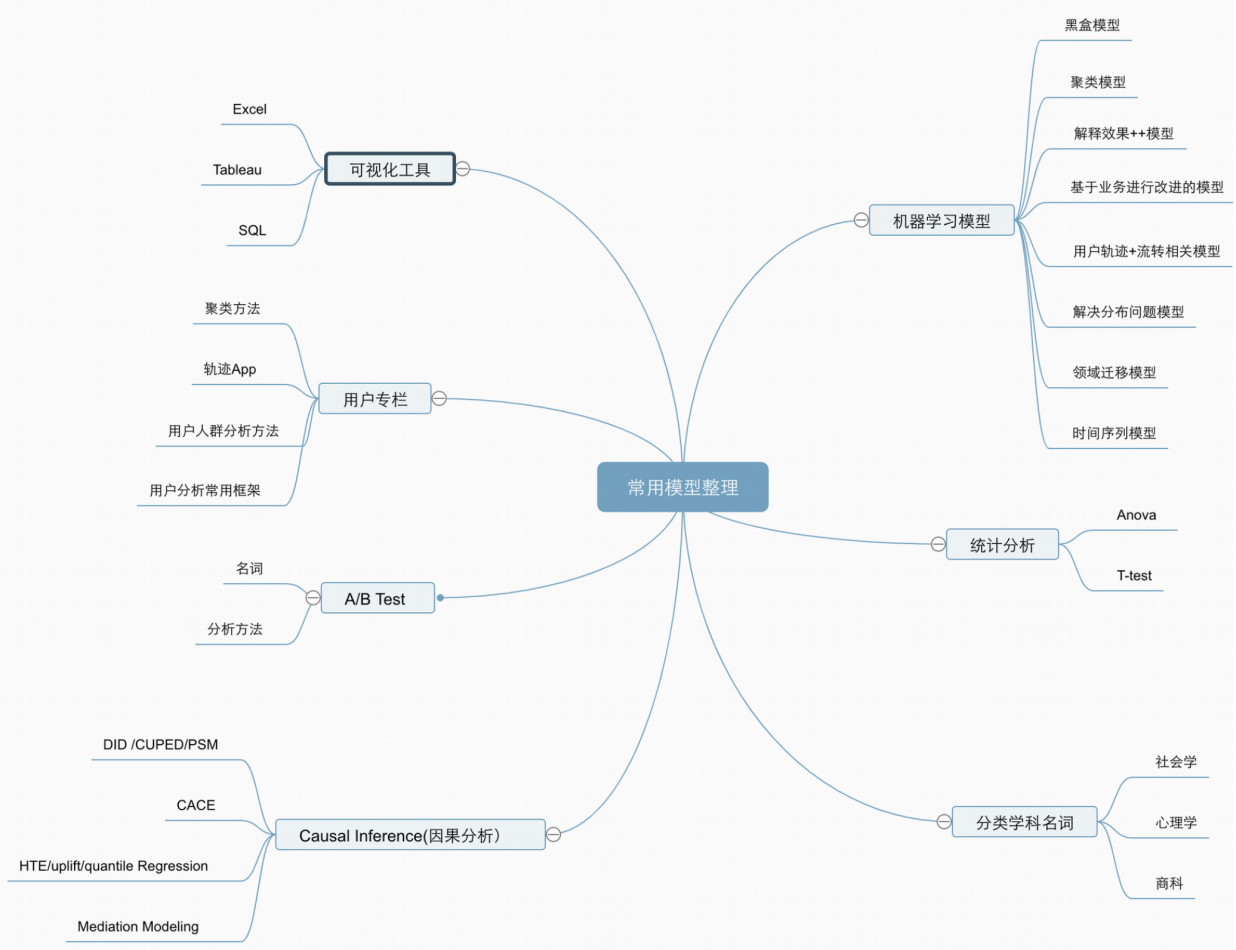
DS常用方法整理

请大家不要私下转载哦，文档会进行人数统计哦，后面我会陆陆续续补充 @Cindy

一. 关于此总结

此doc，主要结合在工作中遇到的各种问题和在文件中所看到的各种方法的总结，主要集中于描述什么场景使用什么方法，并且附上了sample code。 对其中的数学原理使用的方法并为进行详细的提及。因为我主要focus 在用户，所以在分类学科名词中，会主要提及到一些商科的框架模型用来做用户研究。

二. 总结框架



三. 机器学习模型总结

3.1 黑盒模型白盒化

【问题】：很多时候，我们建立了一个模型如（SVM, Xgboost, Neuro Network,甚至于一些NLP的task：sentiment analysis), 我们很难去解释，到底是什么Feature起到了重要的作用导致了这个结果？然而我们模型也需要业务方去理解，所以如何把上面提到的黑盒的模型白盒化就十分重要。

【经典例子】：

1. 我们想要看对于用户在A视频的观看时长，用户有哪些行为， 和哪些行为对观看时长影响最大?

【解决方法】：

1.**SHAP VALUE**

作用： 用来描述对于模型全局来说，每个变量对于全局的贡献程度。且可以看到正负的贡献程度，也可以看到两个相关的变量的相互影响的情况。同时也可以对应我们可以看到每一个个体的对于总总体的贡献程度，和对于每一个个体，每个特征的贡献程度。

Github代码源：https://github.com/slundberg/shap

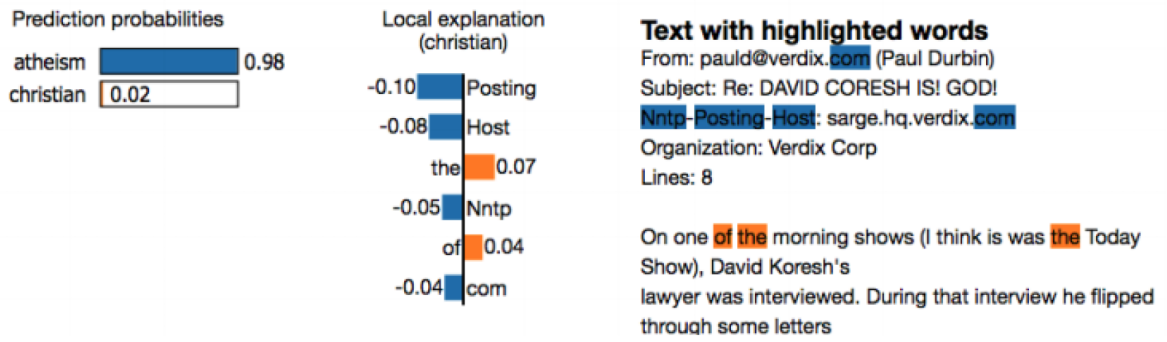
注意： 关于这个包有两个API，一个是原始的API Dmatrix, 另一个是sklearn 里面有shap value的interpret 的package, 两个的结果会有一定的区别，如果结果有很大的差别。需要多放一些数据，然后尝试多运行几次。根据业务进行选择。

2. **LIME**（Local Interpretable Model-agnostic Explanations）

作用：The overall goal of LIME is to identify an interpretable model over the interpretable representation that is locally faithful to predictions of any classifier。

Lime也可以用于对于总体来说，每一个部分的贡献程度[图4]。并且可以用作interpret, 在消失某一个变量的时候，其他的变量的贡献程度和总体是如何变化的。

Github代码源（LIME 解释应用于NLP）： https://github.com/marcotcr/lime



注意：这个part, 其实我们组用的比较多的是SHAP，但是其实都是可以用的

3.2 聚类模型

【问题】：很多时候我们需要看用户到底分为几类，还有不同类别的用户对视频的喜好是什么样子的。所以这个时候，就涉及到用户的聚类模型。

【经典例子】：

1. 根据用户的消费行为，我们想要研究用户分为哪几类？找出有心智的用户?

【解决方法】聚类及降维方法分享

1.Kmeans -基于距离的聚类模型

- 是我们最先尝试的一种方法
- 优点在于： 可视化效果好，可以比较好且容易的选择cluster (based on Elbow Curve 和 silhouette score)

- 缺点在于：有时候不容易找到最聚合的点，会不收敛
- 这里多说一点，在viz 的时候，最后可以选本身不相关的纬度看一下，也可以用ploty 的包画一个3D的图

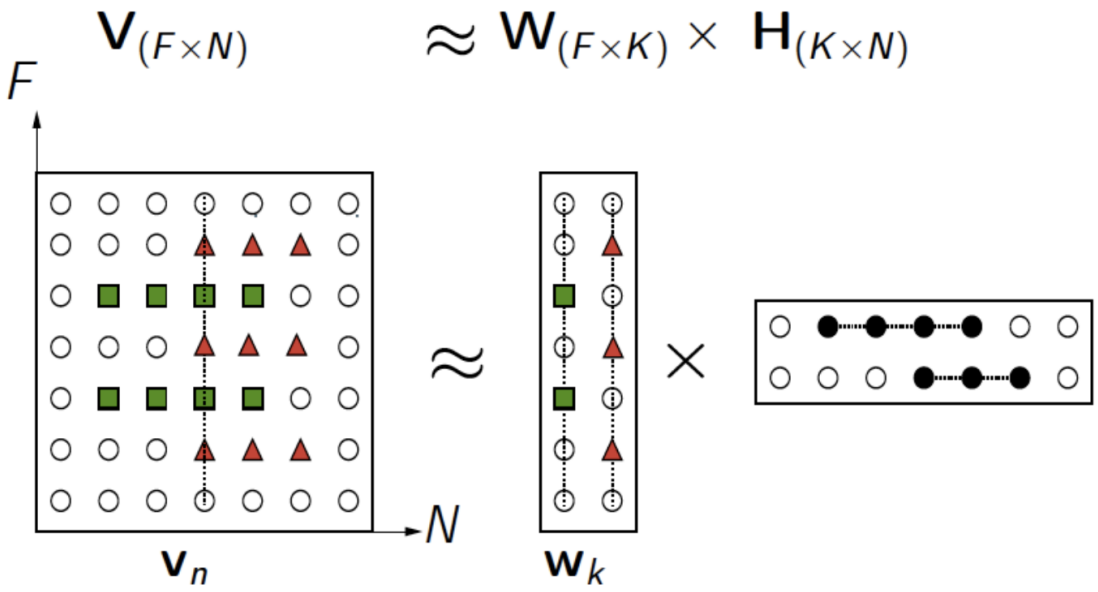
2.DBscan - 基于密度的

- 这边用这种方法比较多，比如说我们对地区进行聚类的时候，我们想要看每个地区用户活跃程度是如何相似的。用的此方法。
- 但是Viz 的部分不是很好解释

3.Gussian Mixture -基于概率密度分布

4.NMF+TF-IDF （NLP迁移方法）

及对非负矩阵分解有用的一些资料链接。NMF，全称为non-negative matrix factorization，中文为"非负矩阵分解"。



NMF的思想：V=WH（W权重矩阵、H特征矩阵、V原矩阵），通过计算从原矩阵提取权重和特征两个不同的矩阵出来。属于一个无监督学习的算法，其中限制条件就是W和H中的所有元素都要大于0。

3.3 解释效果++ 模型

问题： 一旦遇到，让定义threshold 的问题， 一定要先想到决策树，当然random forest 我也把它放到里面，因为random forest 也是可以pick 其树中的一棵树来看。比如 EconML 里面，也是建立了causal forest , 最后pick 一颗causal tree进行解释。R中的'grf '包做HTE有异曲同工。

【经典例子】
1. 定义我们根据什么指标进行不同用户的划分？然后划分的阈值是什么？

【解决方法】：
决策树：决策树，最主要的是挑参数，可以先从split 参数和max-depth 参数有奇效。其他的参数变化不大，最好in smaple 和out of sample 看一下会不会overfitting ,大概率是很容易overfiiting 的。并且要对决策树的理解正确

注意：一般的化，决策树不会跑的那么好，到了第四个分层就会有点乱了。所以要学会后剪枝，在这里我主要推荐REP(Reduced Error Pruning)[3]

优点：
•REP 是当前最简单的事后剪枝方法之一。
•它的计算复杂性是线性的。
•和原始决策树相比，修剪后的决策树对未来新事例的预测偏差较小。

缺点：
•但在数据量较少的情况下很少应用. REP方法趋于过拟合(overfitting) , 这是因为训练数据集中存在的特性在剪枝过程中都被忽略了, 当剪枝数据集比训练数据集小得多时 , 这个问题特别值得注意.

3.4 基于业务进行改进的模型

Survival analysis

LTV (Life Time Value ）

【问题】： 当我们遇到要计算一个视频作者的生命周期或者是用户的生命周期的时候，我们需要根据其历史数据进行预测。这个时候，LTV的计算。

【经典例子】：
关于客户和公司之间的customer lifetime value

3.5 用户轨迹+流转相关模型

【问题】： 很多时候，我们需要去看
1. 不同的cohort 之间的用户是怎么进行流转的， 用户之间按照一定的概率，有多少用户在cohort 1 , 未有外界干预的情况下流转到cohort 2
2. 用户在不同的tab 之间是如何流转的，从第一个tab 开始，有多少人进入了第二个tab , 然后分析流转的轨迹?

【经典例子】
1.5 类 cohort 之间是如何进行流转的？有多少percent 会进行转化？哪一类别的用户更有可能转化成栏目有心智的用户？/ 同城这边也做过这个，是高中低活用户的相互转化
2.用户在不同tab 之间的流转分析？ 比如哪些用户在tab 的流转轨迹是相同的（包括其路径和停留时间）
3. 地理位置来说，用户不同轨迹上的相似性。找出轨迹相似的Cohorts.

【方法】

1.Sanky Diagram

用来直观的展现用户的流转情况。（并非预测）

2. Markov Chain :

马尔可夫链的应用： 是基于Baysian Statistics Approach 对用户下一个状

态进行的预测。 https://www.youtube.com/watch?v=vTUwEu53uzs This video is pretty helpful to understand what is MCMC (Markov Chian Monte Carlo)
例子

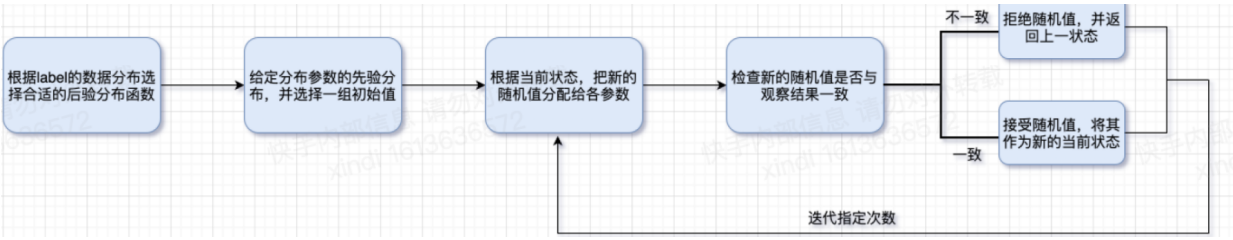
3. LSH (Minhash) Locality sensitive Hashing

利用局部敏感哈希的原理， 可以看到不同的用户之间的轨迹是不是有比较高的相似性

3.5 解决分布问题模型

【问题】当我们的模型预测的时候准确程度不是很高，是由于我们的数据比较有偏，这个时候我们可以用MCMC进行优化。

【实际例子】在实际业务场景中，我们经常遇到对视频播放时长、直播观看时长、DAU等连续变量进行建模排序的问题。当变量本身波动比较大，回归拟合的效果一般时，我们通常会考虑设定一个阈值来划分正负样本，将其转换为二分类问题，利用分类模型所得score（即属于正样本的概率）对样本进行排序。因此如何科学合理地设定正负样本阈值是一个值得探讨的问题，本文介绍了一种基于MCMC的阈值设定方法，即通过MCMC采样所得label后验分布的参数设定阈值。



3.6 领域迁移模型

【问题】+【方法】

1. NLP 的迁移应用

1） 评论生态分析 2） word2vec找相似的IP 3）时效关键词 4）时空分析 5）社交图鉴

2. 心理学问题的迁移应用 -韦伯 - 费纳希定理

1）如何衡量主播或者视频作者对涨粉的感知?

3. 定义多样性问题的迁移应用

1）内容的多样性
2）生态多样性

4. 地理信息的迁移应用

利用Moran’s I 和 Local Moran’s I对活跃用户的聚集效应进行衡量。

5. 传播统计模型

在研究COVID-19病毒传播的时候，我们会用到SIR模型，用来计算病毒的传播的影响范围。在我们的项目中我们会来分析和预测视频的传播情况。比如说我们也可以用来话题的传播速度等等。

3.7 时间序列模型

【问题】这个方法我们用的不是很多。大多数比如我们想要看股票下一个季度的价格，房价下一个季度的价格。还有比如病毒在什么时间将会减少到0.

【具体例子】实验分析中，我们在t1 时刻的，进行了treatment 1。 但是在T2时刻treatement 1并没有进行，但是我们需要以treatment 1作为对照组进行比较。因此我们用了时间序列的方法去预测，这个treatment 1在T2 时刻，如果没有干预的情况下会是什么结果?

【方法】1. ARIMA / Auto-ARIMA
我个人比较推荐auto arima 尤其是对于新手， arima 有三个参数pqm, 需要自己确定，因此很多时候，是不容易进行确定的。而auto_arima 就是会比较好的解决这个问题。

2. Holt-Winters 模型
Holt-winters 的模型是可以加入seasonality 的，比如是以一个星期为纬度。还是以特定的一个月。[5][6] 霍尔特-温特（Holt-Winters）该方法对含有线性趋势和周期波动的非平稳序列适用，利用指数平滑法（EMA）让模型参数不断适应非平稳序列的变化，并对未来趋势进行短期预报。Holt-Winters 方法在 Hol t 模型基础上引入了 Winters 周期项（也叫做季节项），可以用来处理月度数据（周期 12）、季度数据（周期 4）、星期数据（周期 7）等时间序列中的固定周期的波动行为。引入多个 Winters 项还可以处理多种周期并存的情况。

4. Exponential Smoothing 的方法 （Excel也可以实现）

4. Facebook的package: Prophet
这个是基于MLP的模型，本身是regression , 但是需要大量的数据模型才会比较准确，所以这个模型再平常实验的数据是不合适的。

●Causal Impact用来解决类似问题：
当然啦，时间序列的模型有时候，很难很solid去说明这个东西的准确性。所以我们业务上用的比较少。当我们为了排除季节性，或者特殊节日的影响。去看t2时刻的，某一个cohort 对于某一个exp 的反应。我们用到更solid 的办法是causal Impact。 【后面causal inference部分我再补充】

四. 统计分析

4.1、假设检验 (源自DS工具箱)
参数检验 - 在已知总体分布的情况下(通常要求服从正态分布) 对一些主要的参数(均值、方差、 相关系数、百分位数)进行的检验(Z or T 体验)
○ 单样本 t 检验:推断该样本来自的总体均数 μ 与已知的某一总体均数 μ0 (常为理论值或标 准值)有无差别
○ 配对样本 t 检验:当总体均数未知时，且两个样本可以配对，同对中的两者在可能会影响 处理效

果的各种条件方面扱为相似
○ 两独立样本 t 检验:无法找到在各方面极为相似的两样本作配对比较时使用

● 非参数检验 - 不考虑总体分布是否已知，而是针对总体的某些假设进行检验
○ 正态性检验:很多统计方法都要求数值服从或近似服从正态分布
■ Q-Q图
■ P-P图
■ W检验
■ 非参数检验的 K-量检验
○ 非参检验主要方法 ■ 卡方检验
■ 二项检验
■ 秩和检验
■ K-量检验等

4.1.1 T-Test

【问题】我们经常遇到两种问题，一种是我们需要看两个群体在某一个连续的变量上面是不是有差异， 比如说 我们想看离职员工和未离职员工在工资的水平上是不是有显著的不同。

【例子】在ab test 中，我们想要看对于exp 组和base 组，在某个指标上是不是有显著性的差异

【方法】： t -test
T-test 的应用是在于连续变量上是不是有显著性的差异

4.2 Anova Analysis

【问题】：在看用户喜好的方面，我们想要看到哪些因素对于用户特定品种video 下的观看时长比较显著。

【方法】：
因此这个时候，方差分析及Anova，就是适用的方法。可以根据p-value 的大小，P-value 越小则我们的相关性就越强/P-vale< 0.05的时候，我们可以说是显著相关。* 相关性不等于causality

五. 分类学科名词

5.1 社会学

1.马太效应（Matthew Effect）

是指强者愈强、弱者愈弱的现象，广泛应用于社会心理学、教育、金融以及科学领域。
【例子】
微博背后的马太效应和长尾效应
们不得不承认，微博在经历了一夜爆红、低谷和二次崛起后，依然是最具影响力的媒体平台。这种影响力来自于马太效应和长尾效应的叠加。微博的马太效应在于，社交媒体一旦形成用户规模，其会获得内容“独家性”，因为内容发布者为了追求最大的传播声量，必然要选择这个平台；微博的长尾效应则表现在，三四线城市的不断深耕和挖掘，让那些不被看好的碎片市场，无论用户还是内容产出，都能形成相当规模的市场。微博亮眼的财报背后，不仅仅是业绩数字好看，而是微博品牌不断成长，微博内容不断丰富，以及三四线用户的增长共同形成的结果。微博增长，这也吸引了越来越多广告主，带动广告效果和
广告收入的增加。微博增长背后，是马太效应和长尾效应叠加的结果。

2. 长尾效应：

最初由《连线》的总编辑克里斯.安德森于2004年发表于自家的杂志中，用来描述诸如亚马逊公司、Netflix之类的网站之商业和经济模式。是指那些原来不受到重视的销量小但种类多的产品或服务由于总量巨大，累积起来的总收益超过主流产品的现象。

【例子】微博在三四线城市的发展
微博的长尾效应则表现在，三四线城市的不断深耕和挖掘，让那些不被看好的碎片市场，无论用户还是内容产出，都能形成相当规模的市场。

3.纳金斯箱定理

4. 卡诺模型（KANO模型)

是对用户需求分类和优先排序的有用工具，以分析用户需求对用户满意的影响为基础，体现了产品性能和用户满意之间的非线性关系。在卡诺模型中，将产品和服务的质量特性分为四种类型：⑴必备属性；⑵期望属性；⑶魅力属性；⑷无差异属性。

●魅力属性：用户意想不到的，如果不提供此需求，用户满意度不会降低，但当提供此需求，用户满意度会有很大提升；
●期望属性：当提供此需求，用户满意度会提升，当不提供此需求，用户满意度会降低；
●必备属性：当优化此需求，用户满意度不会提升，当不提供此需求，用户满意度会大幅降；
●无差异因素：无论提供或不提供此需求，用户满意度都不会有改变，用户根本不在意；
●反向属性：用户根本都没有此需求，提供后用户满意度反而会下降

5.2 心理学

1.韦伯 费希纳

2.马斯洛需求理论

3.Angle effect

4. 10 Principle Marketing Phychology

八. 因果分析

1、因果分析：

跨过相关性分析 -> 因果分析
使用范围：过去项目的评估，现在项目的可测量优化，未来项目的方向预估获取长期价值
特点：不仅关心A与B的correlation，更关心A与B之间的因果关系
愿景：任何被提出的关于商业逻辑的猜想都需要经过严密的因果分析，确保这些猜想对于涉及的用
户和商家都是正确的
差异性：
- 相关性分析：两件同时发生的事情放在一起分析，叫做相关性分析
- 因果分析：分析一件事情发生的原因，叫做因果分析

2、统计检验

T检验：对比两组连续型变量的显著性
卡方检验：对比两组二元变量的显著性
Ab实验：2个base组【aa实验：对比波动是否显著】，一个exp组

3、DID

计算过程：实验后的差值 - 实验前的差值 = 双重差分， a3的显著性为DID的显著性
DID重要假设：平行趋势假定，即实验组和控制组在政策前后是具有一致的趋势

5. HTE

因为我主要参与了HTE的部分，所以我主要说一下，HTE是用来找到实验最正向的一群用户或者是最负向的用户。因为当我们有很多feature 的时候，我们不容易进行总结，去看到底哪部分群体对实验有比较大的reaction。

Summary：

上面只是一部分的
知识，其中关于一些我没有整理的部分,希望这个可以作为一个范本给后面需要的同学，当遇到不同的问题的时候，可以找到适当的方法。也欢迎大家在上面补充和修改。谢谢大家

Reference：

[1]Shapley Additive Explanation (SHAP) Lundberg and Lee (2017); Lund- berg et al. (2020)
[2]Local Interpretable Model-agnostic Explanations (LIME) Ribeiro et al. (2016)
[3] https://zhuanlan.zhihu.com/p/30296061
[4]https://courses.lumenlearning.com/suny-natural-resourcesbiometrics/chapter/chapter-10-quantitative-measures-of-diversity-site-similarity-andhabitat-suitability/
[5]https://baike.baidu.com/item/Holt-Winters%20%E6%96%B9%E6%B3%95/24137738
[6]https://otexts.com/fpp2/holt-winters.html
[7] https://www.zhihu.com/question/22989667