Thinking1:在 CTR 点击率预估中，使用 GBDT+LR 的原理是什么？

LR 是广义线性模型，与传统线性模型相比，LR 使用了 Logit 变换将函数值映射到 0~1 区间，映射后的函数值就是 CTR 的预估值。

优点：LR 这种线性模型很容易并行化，处理上亿条训练样本不是问题

缺点：线性模型学习能力有限，需要大量特征工程预先分析出有效的特征、特征组合，从而去间接增强 LR 的非线性学习能力。

GBDT（Gradient Boost Decision Tree）是一种常用的非线性模型，它基于集成学习中的 boosting 思想，每次迭代都在减少残差的梯度方向新建立一颗决策树，迭代多少次就会生成多少颗决策树。GBDT 的思想使其具有天然优势可以发现多种有区分性的特征以及特征组合，决策树的路径可以直接作为 LR 输入特征使用，省去了人工寻找特征、特征组合的步骤

Thinking2：Wide & Deep 的模型结构是怎样的，为什么能通过具备记忆和泛化能力（memorization and generalization）？

Wide and deep 模型是 TensorFlow 在 2016 年 6 月左右发布的一类用于分类和回归的模型，并应用到了 Google Play 的应用推荐中。wide and deep 模型的核心思想是结合线性模型的记忆能力（memorization）和 DNN 模型的泛化能力（generalization），在训练过程中同时优化 2 个模型的参数，从而达到整体模型的预测能力最优。

记忆（memorization）即从历史数据中发现 item 或者特征之间的相关性。

泛化（generalization）即相关性的传递，发现在历史数据中很少或者没有出现的新的特征组合。

为什么泛化能力强：DNN 这种 embedding 类得模型，可以通过学习到得低维稠密向量实现模型得泛化能力，包括可以实现对未见过得内容进行泛化推荐

Thinking3：在 CTR 预估中，使用 FM 与 DNN 结合的方式，有哪些结合的方式，代表模型有哪些？

在 CTR 预估中，为了解决稀疏特征的问题，学者们提出了 FM 模型来建模特征之间的交互关系。但是 FM 模型只能表达特征之间两两组合之间的关系，无法建模两个特征之间深层次的关系或者说多个特征之间的交互关系，因此学者们通过 Deep Network 来建模更高阶的特征之间的关系。

因此 FM 和深度网络 DNN 的结合也就成为了 CTR 预估问题中主流的方法。有关 FM 和 DNN 的结合有两种主流的方法，并行结构和串行结构。

并行结构：FM 部分和 DNN 部分分开计算，只在输出层进行一次融合得到结果。常见模型：DeepFM，DCN，Wide&Deep。

串行结构：将 FM 的一次项和二次项结果（或其中之一）作为 DNN 部分的输入，经 DNN 得到最终结果。常见模型：PNN，NFM，AFM

thingking4：GBDT 和随机森林都是基于树的算法，它们有什么区别？

随机森林：采用 bagging 思想，即利用 bootstrap 抽样，得到若干个数据集，每个数据集都训练一颗树。构建决策树时，每次分类节点时，并不是考虑全部特征，而是从特征候选集中选取若干个特征用于计算。弱特征共有 p 个，一般选取 m=sqrt(p)个特征。当可选特征数目很大时，选取一个较小的 m 值，有助于决策树的构建。当树的数量足够多时，RF 不会产生过拟合，提高树的数量能够使得错误率降低。

GBDT: 采用 Boosting 思想（注意是 Boosting，不是 Boostrap），不采用 Boostrap 抽样的方法（RF 采用了），每次迭代过程都会使用全部数据集（会有一些变化，即采用的是上一轮训练后得到的预测结果与真实结果之间的残差（残差是由损失函数计算得到的））。GBDT 的每棵树是按顺序生成的，每棵树生成时都需要利用之前一棵树留下的信息（RF 的树是并行生成的）。GBDT 中树的数目过多会引起过拟合（RF 不会）。

两者不同点：

1、组成随机森林的树可以是分类树，也可以是回归树；而 GBDT 只由回归树组成

2、组成随机森林的树可以并行生成；而 GBDT 只能是串行生成

3、对于最终的输出结果而言，随机森林采用多数投票等；而 GBDT 则是将所有结果累加起来，或者加权累加起来

4、随机森林对异常值不敏感，GBDT 对异常值非常敏感

5、随机森林对训练集一视同仁，GBDT 是基于权值的弱分类器的集成

6、随机森林是通过减少模型方差提高性能，GBDT 是通过减少模型偏差提高性能

Thinking5：item 流行度在推荐系统中有怎样的应用？

基于流行度的算法非常简单粗暴，类似于各大新闻、微博热榜等，根据 PV、UV、日均 PV 或分享率等数据来按某种热度排序来推荐给用户。这种算法的优点是简单，适用于刚注册的新用户。缺点也很明显，它无法针对用户提供个性化的推荐。基于这种算法也可做一些优化，比如加入用户分群的流行度排序，例如把热榜上的体育内容优先推荐给体育迷，把政要热文推给热爱谈论政治的用户。

物品冷启动:物品冷启动主要解决如何将新的物品推荐给可能对它感兴趣的用户这一问题

物品冷启动在 新闻网站等时效性很强的网站中非常重要。一般来说，物品的内容可以通过向量空间模型 1 表示，该模型会将物品表示成一个关键词向量。如果物品的内容是一些诸如导演、演员等实体的话，可以直接将这些实体作为关键词。但如果内容是文本的形式，则需要引入一些理解自然语言的技术抽取关键词。如果物品是电影，可以根据演员在剧中的重要程度赋予他们权重。向量空间模型的优点是简单，缺点是丢失了一些信息，比如关键词之间的关系信息。不过在绝大多数应用中，向量空间模型对于文本的分类、聚类、相似度计算已

经可以给出令人满意的结果。在给定物品内容的关键词向量后，物品的内容相似度可以通过向量之间的余弦相似度计算。得到物品的相似度之后，可以利用 ItemCF 算法的思想，给用户推荐和他历史上 喜欢的物品内容相似的物品。

基于物品的 CF TopN 推荐只是主体在于物品：

1. 分析各个用户对 item 的浏览记录。
2. 依据浏览记录分析得出所有 item 之间的相似度；
3. 对于当前用户评价高的 item，找出与之相似度最高的 N 个 item；
4. 将这 N 个 item 推荐给用户。