怎样将Embedding融入传统机器学习框架?

原创 石塔西 夕小瑶的卖萌屋 2020-12-19 22:00

LR本身是一个经典的CTR模型，广泛应用于推荐/广告系统。输入的特征大多数是离散型/组合型。那么对于Embedding技术，如何在不使用深度学习模型的情况下（假设就是不能用DNN），融入到LR框架中呢？让我们来看看清华大学的博士石塔西大佬是如何解答的。

问题实战意义

其实这个问题可以再扩展一下，即，**如何在传统机器学习算法（LR/GBDT）中使用Embedding信息。**

这个问题并非空穴来风，而是有一定的实战价值。 目前DNN热度不减，基本上成为推荐、搜索系统的标配算法。传统机器学习算法，如LR、GBDT，纷纷被打入冷宫，得不到关注。至于为什么DNN能够成功上位，独占各位打工人的欢心，**请参考我的文章《无中生有：论推荐算法中的Embedding思想》[1]**。

但是，**DNN有一个致命缺点，就是上线困难。** 训练的时候，各位调参侠，把各种酷炫的结构，什么attention, transformer, capsule，能加上的都给它加上，看着离线指标一路上涨，心里和脸上都乐开了花，却全然无视旁边的后端工程师恨得咬紧了牙根。模型越复杂，离线和线上指标未必就更好，但是线上的时间开销肯定会增加，轻则影响算法与后端的同事关系（打工人何苦为难打工人），重则你那离线指标完美的模型压根没有上线的机会。虽说，目前已经有TF Serving这样的线上serving框架，但是它也不是开箱即用的，也需要一系列的性能调优，才能满足线上的实时性要求。

所以，如果你身处一个小团队，后端工程人员的技术能力不强，在线DNN就会成为一个难题，这个时候，**传统的LR、GBDT就凸显出优势。** 如果全部使用ID类特征（实数特征也桶化成ID类特征），那么LR在线上就简化成“查表取权重，再累加“，连乘法都省了，实时性自然有保证。

但是，如果你想鱼与熊掌兼得，**既不得不使用简单的传统机器学习算法，又想利用Embedding带来扩展能力上的提升，** 你该怎么办？唉，费了半天口舌，只是解了题而已，目的是为了说明这一问题的实战意义，引起大家对这一问题的重视。

不推荐直接使用Embedding本身

首先，如果你的主框架是传统机器学习算法，那么Embedding肯定就不能是End-To-End学习得到的，而需要离线用另外的算法先学习好。比如，你使用DeepWalk先学习用户的购买序列，离线学习好商品的Embedding。

第二个问题才是传统机器学习如何利用这些Embedding。当然最简单的方法就是直接使用，为了使用一个64维的向量，就相当于LR增加了64维特征。但是，我**不推荐**使用这种方式：

- 之所以线上使用LR，看中就是使用其处理高维、稀疏的ID类特征的能力，线上操作简化成“查表、累加权重”的快速便捷。如果你使用了向量这样的稠密特征，那么LR的优点就不复存在了。更何况有的Embedding，比如图片的Embedding可能上千维，破坏了稀疏性，线上的存储与计算都很困难。
- LR所使用的Embedding是离线计算得到的，黑盒，可解释性不强。而我们使用LR，图的就是其可解释性强，方便debug。
- 另外，Embedding还不稳定，因为计算Embedding的离线程序可能也需要升级。一旦升级，之前累积的训练样本就全部作废，因为新老Embedding肯定不处于同一个坐标系下，不能混用。

推荐使用基于Embedding的衍生指标

所以，**我不推荐在LR中直接使用Embedding。** 在我看来，正确的姿势，**应该是基于离线生成的Embedding，衍生出一系列衡量<user,item>相关度的指标，然后在LR中使用这些衍生指标。** 这种作法也并非我的空想，也是有出处、经过实践检验的。Airbnb的《Real-time Personalization using Embeddings for Search Ranking at Airbnb》[2] 中就采用这种方法，将离线计算好的Embedding，喂入他们的GBDT排序模型。

详细算法，请阅读Airbnb论文的第4.4节，我这里将Airbnb的做法简述如下：

1.前提，Airbnb已经将listing(房屋) embedding离线计算好

2.从多种角度来收集用户的历史，

1. 比如Hc代表用户过去2周点击过的listing集合，
2. Hs代表曝光给用户但被忽略的listing集合，
3. Hw是用户收藏的listing的集合，
4. Hb是用户预订过的listing的集合，……

3.将以上某个集合中所有listing的embedding取平均，当成user在这个行为（点击、忽略、收藏、预订、……）下的embedding

4.再拿user在某个行为下的user embedding，与当前要排序的listing embedding，计算cosine similarity，作为user对当前listing执行某动作（点击、忽略、收藏、预订、……）的倾向性。将这种“执行某动作的倾向性得分”作为实数特征，喂入GBDT，训练排序模型。

5.将这种“执行某动作的倾向性得分”作为实数特征，喂入GBDT，训练排序模型。

6.除了以上用户的长期兴趣（H*都是以周为单位收集的），Airbnb还计算当前待排序的listing embedding与用户最后一次点击的listing embedding的相似性，来刻画用户的短期兴趣。

Airbnb使用的全部基于listing embedding的衍生指标见论文中的表6

Table 6: Embedding Features for Search Ranking

Feature Name	Description
EmbClickSim	similarity to clicked listings in H_c
EmbSkipSim	similarity to skipped listings H_s
EmbLongClickSim	similarity to long clicked listings H_{lc}
EmbWishlistSim	similarity to wishlisted listings H_w
EmbInqSim	similarity to contacted listings H_i
EmbBookSim	similarity to booked listing H_b
EmbLastLongClickSim	similarity to last long clicked listing
UserTypeListingTypeSim	user type and listing type similarity

总结

- 在传统机器学习中使用Embedding，这个问题，有一定的**实战意义。特别是你想规避DNN模型复杂的上线流程，而又想获得Embedding带来的扩展性的提升的时候。**

- 在传统机器学习模型中使用Embedding，**我不推荐直接使用Embedding，而建议使用基于Embedding计算得到的衍生指标。**

参考文献

[1]https://zhuanlan.zhihu.com/p/320196402

[2]https://zhuanlan.zhihu.com/p/162163054

阅读原文　文章已于2021/01/11修改

喜欢此内容的人还喜欢

我不看好data2vec这类多模态融合的研究

夕小瑶的卖萌屋