详解 Google 多任务学习模型 MMoE (在推荐场景中应用广泛)

杨镒铭 智能推荐系统 2019-06-11


Research Track Paper KDD 2018, August 19-23, 2018, London, United Kingdom

Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts

Jiaqi Ma[1*], Zhe Zhao[2], Xinyang Yi[2], Jilin Chen[2], Lichan Hong[2], Ed H. Chi[2]
[1]School of Information, University of Michigan, Ann Arbor [2]Google Inc.
[1]jiaqima@umich.edu [2]{zhezhao, xinyang, jilinc, lichan, edchi}@google.com

文章作者：杨镒铭 阿里妈妈 算法专家
内容来源：记录广告、推荐等方面的模型积累@知乎专栏
出品社区：DataFun

文章发表在 KDD 2018 Research Track 上，Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts 。地址：

https://www.kdd.org/kdd2018/accepted-papers/view/modeling-task-relationships-in-multi-task-learning-with-multi-gate-mixture-

在工业界基于神经网络的多任务学习在推荐等场景业务应用广泛，比如在推荐系统中对用户推荐物品时，不仅要推荐用户感兴趣的物品，还要尽可能地促进转化和购买，因此要对用户评分和购买两种目标同时建模。阿里之前提出的 ESSM 模型属于同时对点击率和转换率进行建模，提出的模型是典型的 shared-bottom 结构。多任务学习中有个问题就是如果子任务差异很大，往往导致多任务模型效果不佳。今天要介绍的这篇文章是谷歌的一个内容推荐团队考虑了多任务之间的区别提出了 **MMoE** 模型，并取得了不错的效果。

一、Motivation

多任务模型通过学习不同任务的联系和差异，可提高每个任务的学习效率和质量。多任务学习的的框架广泛采用 shared-bottom 的结构，不同任务间共用底部的隐层。这种结构本质上可以减少过拟合的风险，但是效果上可能受到任务差异和数据分布带来的影响。也有一些其他结构，比如两个任务的参数不共用，但是通过对不同任务的参数增加 L2 范数的限制；也有一些对每个任务分别学习一套隐层然后学习所有隐层的组合。和 shared-bottom 结构相比，这些模型对增加了针对任务的特定参数，在任务差异会影响公共参数的情况下对最终效果有提升。缺点就是模型增加了参数量所以需要更大的数据量来训练模型，而且模型更复杂并不利于在真实

生产环境中实际部署使用。

因此，论文中提出了一个 Multi-gate Mixture-of-Experts（ MMoE ）的多任务学习结构。MMoE 模型刻画了任务相关性，基于共享表示来学习特定任务的函数，避免了明显增加参数的缺点。

二、模型介绍

MMoE 模型的结构（ 下图 c ）**基于广泛使用的 Shared-Bottom 结构（ 下图 a ）和 MoE 结构**，其中图（ b ）是图（ c ）的一种特殊情况，下面依次介绍。

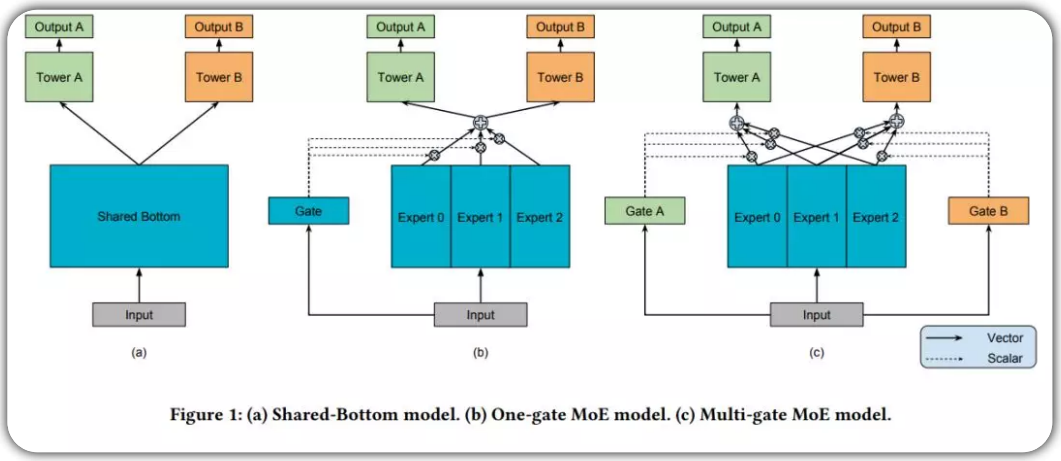

Figure 1: (a) Shared-Bottom model. (b) One-gate MoE model. (c) Multi-gate MoE model.

- **Shared-Bottom Multi-task Model**

如上图 a 所示，shared-bottom 网络（ 表示为函数 f ）位于底部，多个任务共用这一层。往上，K 个子任务分别对应一个 tower network（ 表示为 h^k ），每个子任务的输出 $y_k = h^k(f(x))$ 。

- **Mixture-of-Experts**

MoE 模型可以形式化表示为：

$$y = \sum_{i=1}^{n} g(x)_i f_i(x)$$

其中 $\sum_{i=1}^{n} g(x)_i = 1$ ， $f_i, i = 1, \cdots, n$ 是 n 个 expert network（ expert network 可认为是一个神经网络 ）。g 是组合 experts 结果的 gating network ， 具体来说 g 产生 n 个 experts 上的概率分布，最终的输出是**所有 experts 的带权加和**。显然，MoE 可看做基于多个独立模型的集成方法。这里注意 MoE 并不对应上图中的 b 部分。

后面有些文章将 MoE 作为一个基本的组成单元，将多个 MoE 结构堆叠在一个大网络中。比如一个 MoE 层可以接受上一层 MoE 层的输出作为输入，其输出作为下一层的输入使用。

- **Multi-gate Mixture-of-Experts**

文章提出的模型（简称 MMoE）目的就是**相对于 shared-bottom 结构不明显增加模型参数的要求下捕捉任务的不同**。其核心思想是将 **shared-bottom 网络中的函数 f 替换成 MoE 层**，如上图 c 所示，形式化表达为：

$$y_k = h^k(f^k(x)), f^k(x) = \sum_{i=1}^{n} g^k(x)_i f_i(x)$$

其中 $g^k(x) = softmax(W_{gk}x)$ ，输入就是 input feature ，输出是所有 experts 上的权重。

一方面，因为 gating networks 通常是轻量级的，而且 expert networks 是所有任务共用，所以相对于论文中提到的一些 baseline 方法在计算量和参数量上具有优势。

另一方面，相对于所有任务公共一个门控网络（One-gate MoE model ，如上图 b），这里 MMoE（上图 c）中每个任务使用单独的 gating networks 。每个任务的 gating networks 通过最终输出权重不同实现对 experts 的选择性利用。不同任务的 gating networks 可以学习到不同的组合 experts 的模式，因此模型考虑到了捕捉到任务的相关性和区别。

三、总结

整体来看，这篇文章是对多任务学习的一个扩展，**通过门控网络的机制来平衡多任务**的做法在真实业务场景中具有借鉴意义。下面补充介绍文中的一个数据集设置的做法和实验结果中对不同模型的相互对比分析。

- **人工构造数据集**

在真实数据集中我们无法改变任务之间的相关性，所以不太方便进行研究任务相关性对多任务模型的影响。轮文中人工构建了两个回归任务的数据集，然后通过两个任务的标签的 Pearson 相关系数来作为任务相关性的度量。在工业界中**通过人工构造的数据集来验证自己的假设**是个有意思的做法。

- **模型的可训练性**

模型的**可训练性，就是模型对于超参数和初始化是否足够鲁棒**。作者在人工合成数据集上进行了实验，观察不同随机种子和模型初始化方法对 loss 的影响。这里简单介绍下两个现象：第一，Shared-Bottom models 的效果方差要明显大于基于 MoE 的方法，说明 Shared-Bottom 模型有很多偏差的局部最小点；第二，如果任务相关度非常高，则 OMoE 和 MMoE 的效果近似，但是如果任务相关度很低，则 OMoE 的效果相对于 MMoE 明显下降，说明 **MMoE 中的 multi-gate 的结构对于任务差异带来的冲突有一定的缓解**作用。

作者介绍：

杨镒铭，阿里妈妈算法专家。硕士毕业于中国科学技术大学，记录广告、推荐等方面的模型积累@知乎专栏作者。

——END——

「 更多干货，更多收获 」

如何将知识图谱特征学习应用到推荐系统?

今日头条推荐系统原理

深度学习与推荐系统完结篇（知识、论文、源码、数据集与行业应用）

智能推荐之：什么是A/B测试（定义、步骤、应用场景及作用）

个性化推荐研究人点之用户画像

京东购物在微信等场景下的智能推荐算法应用与实践

feed流设计：那些谋杀你时间的APP

【推荐算法】基于关联规则的推荐算法及业务实践

【推荐算法】基于用户和产品的协同过滤算法

【推荐算法】基于内容的推荐算法

【推荐算法】基于流行度的推荐算法

【推荐算法实践】音乐歌单智能推荐

从零开始搭建创业公司后台技术栈

如何搭建一套个性化推荐系统

> **关注我们**