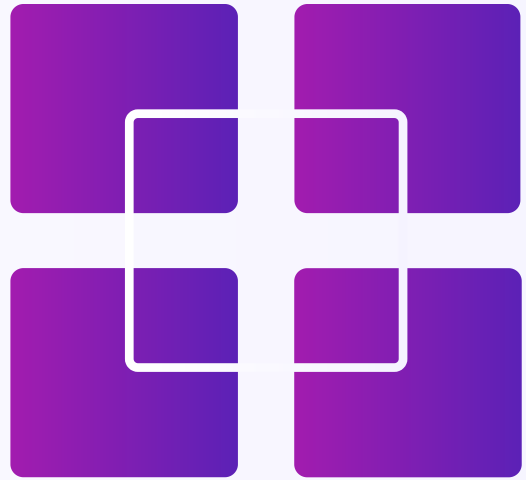
Modeling Task Relationships in Multi-task Learning with Multi-gate Mixture-of-Experts

KDD18,谷歌
乔梁 2022.11.25



一、摘要

CTR

背景：多任务学习已经成功地应用于许多现实世界的大规模应用中，如推荐系统。例如，在电影推荐中，除了向用户提供他们倾向于购买和观看的电影之外，系统还可以针对之后喜欢电影的用户进行优化。

之前的方法：对于多任务学习，我们的目标是建立一个单一的模型，同时学习这些多个目标和任务。然而，常用的多任务模型的预测质量往往对任务之间的关系敏感。因此，研究任务特定目标和任务间关系之间的建模权衡是重要的。

本文方法:本文提出了一种新的多任务学习方法——MMoE，该方法能够从数据中**显式地**学习任务关系模型。我们通过在所有任务之间共享专家子模型，同时还训练门控网络来优化每个任务，从而使MMoE结构适应多任务学习。

为了验证我们的方法，我们首先将其应用于合成数据集，其中我们控制了任务相关性。实验结果表明，在任务相关性较低时，该方法的性能优于基线方法。此外，我们还演示了MMoE在真实的任务中的性能改进，包括二进制分类基准测试和Google的大规模内容推荐系统。

二、引言

为什么需要引入多任务学习?

- 近年来，在推荐领域逐渐引入多任务学习来减轻一些使用单个模型指标可能带来的负面影响。
- 例如在视频推荐中，只考虑点击转化率时，会倾向推荐包含标题党、擦边海报的视频；只考虑完成度时，会倾向推荐时间比较短的视频等等。而这些倾向都会影响用户体验，并且可能导致业务长期目标的下降。
- 因此，大家开始尝试引入多个相互关联但又不一致的目标来进行综合考虑建模，并且实践表示，多任务学习在推荐系统中能够提升上下文推荐的效果。

之前的方法

- **难以测量真实任务差异**：之前的研究通过假设每项任务的特定数据生成过程，根据假设测量任务差异，然后根据任务的差异程度提出建议，研究了多任务学习中的任务差异。然而，由于真实的应用程序通常具有复杂得多的数据模式，因此通常很难衡量任务差异并使用这些方法。

- **参数过多**：最近的几项研究提出了新的建模技术，以处理多任务学习中的任务差异，而不依赖于明确的任务差异测量。然而，这些技术通常涉及为每个任务添加更多的模型参数，以适应任务差异。由于大规模推荐系统可能包含数百万或数十亿个参数，这些额外的参数通常是欠约束的，这可能影响模型质量。

本文方法

在这篇论文中，我们提出了一种MMoE多任务学习方法，该结构受到了MoE模型和MoE层的启发。MMoE明确地对任务关系进行建模，并学习特定于任务的功能以利用共享表示。它允许自动分配参数，以捕获共享任务信息或特定任务信息，从而避免为每个任务添加许多新参数。

MMoE的主干构建在最常用的共享底层多任务DNN结构之上。共享底层模型结构如图1（a）所示，其中输入层之后的几个底层在所有任务之间共享，然后每个任务在底层表示的顶部都有一个单独的网络"塔"。

底层：MMoE模型（如图1（c）所示）不是所有任务共享一个底层网络，而是有一组底层网络，每个底层网络被称为“专家”。在本文中，每个专家都是一个前馈网络。

门控：然后，为每个任务引入一个门控网络。门控网络采用输入特征和输出softmax门控，以不同权重组合专家，允许不同任务以不同方式利用专家。然后，集合专家的结果被传递到特定于任务的塔网络。

通过这种方式，不同任务的门控网络可以学习专家集合的不同混合模式，从而捕获任务关系。

本文贡献

- 首先，提出了MMoE，该模型对任务关系进行了显式建模。通过门控网络，MMoE模型自动调整共享信息建模和任务特定信息建模之间的参数化。
- 其次，对合成数据进行了控制实验。报告了任务相关性如何影响多任务学习中的训练，以及MMoE如何提高模型的表达性和可训练性。
- 最后，在真实的基准数据和一个拥有数亿用户和项目的大规模产品推荐系统上进行了实验。实验验证了我们提出的方法在真实环境中的效率和有效性。

图1: 模型比较

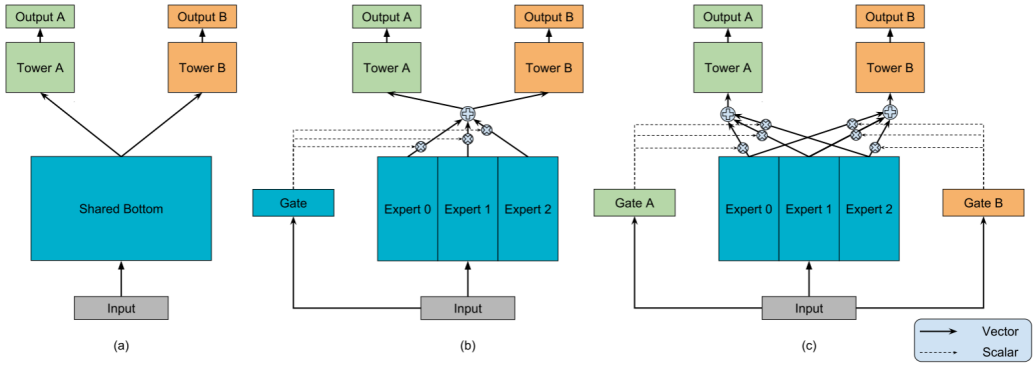


Figure 1: (a) Shared-Bottom model. (b) One-gate MoE model. (c) Multi-gate MoE model.

三、预备知识

Shared-bottom多任务模型

首先介绍图1（a）中的共享底层多任务模型，这是由Rich Caruana提出的框架，并在许多多任务学习应用中广泛采用。因此，将其视为多任务建模中的一种代表性基线方法。

给定K个任务，该模型由一个共享底部网络（表示为函数f）和K个塔网络h^k组成，其中$k = 1, 2, \cdots, K$分别为每项任务。共享底部网络跟随输入层，并且塔网络建立在共享底部的输出上。然后，每个任务的单独输出y_k遵循相应的任务特定塔。对于任务k，模型可以被公式化为，

$$y_k = h^k(f(x)) \tag{1}$$

四、建模方法

MoE

原始MoE模型可表述为：

$$y = \sum_{i=1}^{n} g(x)_i f_i(x) \tag{5}$$

其中$\sum_{i=1}^{n} g(x)_i = 1$，并且$g(x)_i$是$g(x)$的输出的第$i$个logit，表示专家$f_i$的概率。

这里，$f_i, i = 1, \cdots, n$是n个专家网络，并且g表示集合来自所有专家的结果的门控网络。更具体地，门控网络g基于输入在n个专家上产生分布，并且最终输出是所有专家的输出的加权和。

MMoE

我们提出了一种新的MoE模型，与shared-bottom多任务模型相比，该模型不需要更多的模型参数就可以捕获任务差异。新模型称MMoE模型，其中关键思想是用等式5中的MoE层代替等式1中的共享底部网络f。更重要的是，我们为每个任务k添加单独的门网络g^k。更准确地说，任务k的输出是

$$y_k = h^k\left(f^k(x)\right),$$

$$\text{其中 } f^k(x) = \sum_{i=1}^{n} g^k(x)_i f_i(x).$$

我们的实现由具有ReLU激活的相同MLP组成。选通网络是带有softmax层的输入的简单线性变换：

$$g^k(x) = \text{softmax}\left(W_{gk}x\right)$$

其中$W_{gk} \in \mathbb{R}^{n\times d}$是可训练矩阵。$N$是专家的数量，$d$是特征维数。

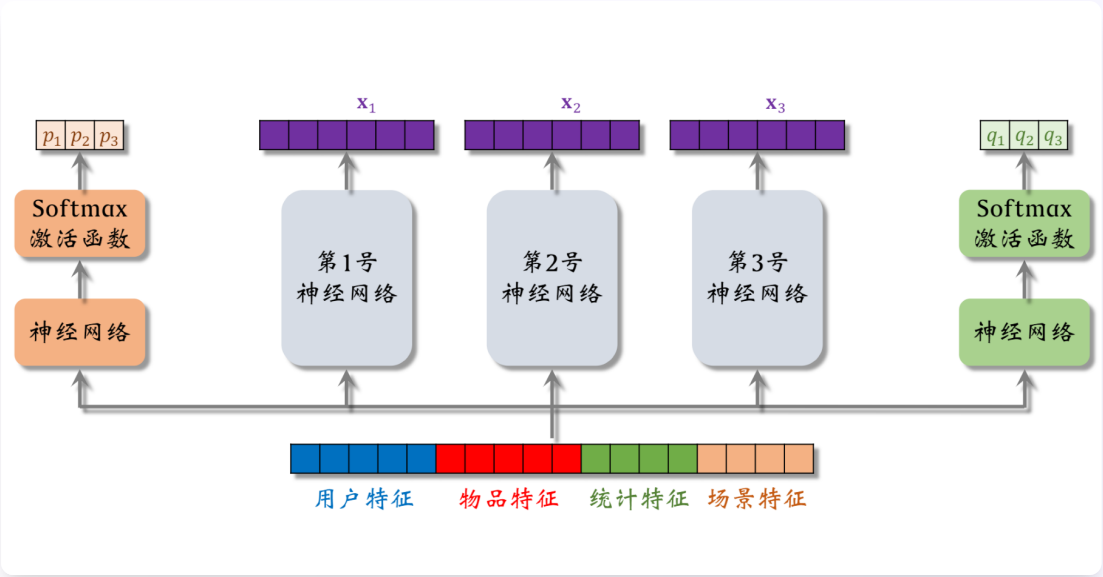
MMoE

每个门网络可以学习“选择”专家的子集以在输入示例的条件下使用。这对于在多任务学习情形中的灵活参数共享是令人满意的。

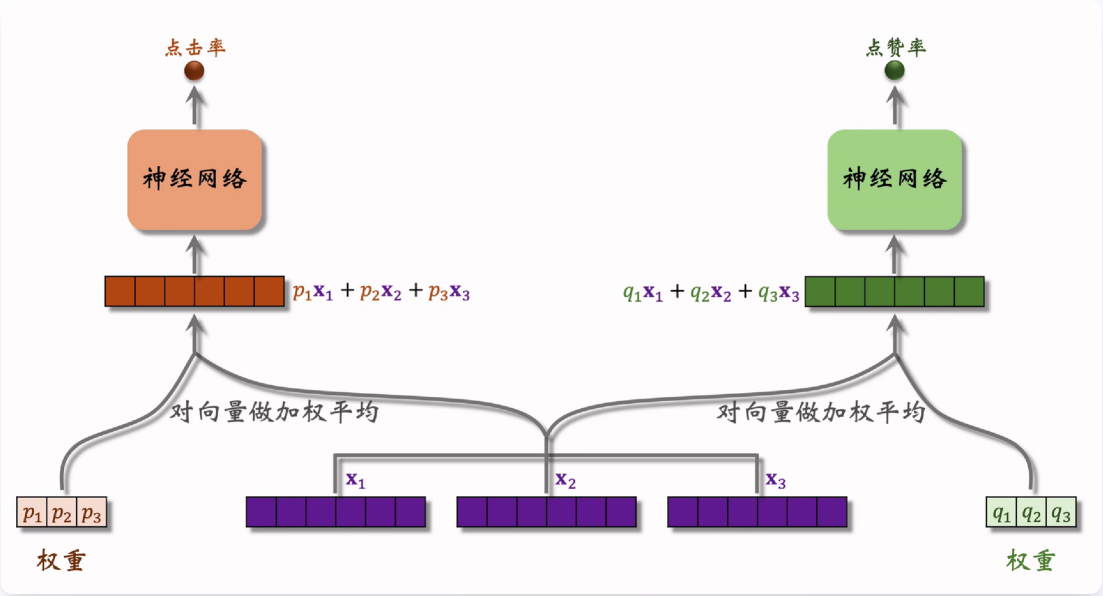
作为一种特殊情况，如果只选择一个具有最高门分数的专家，则每个门网络实际上将输入空间线性地分成n个区域，每个区域对应于一个专家。MMoE能够通过确定不同门导致的分离如何彼此重叠来以复杂的方式对任务关系进行建模。如果任务的相关性较低，则共享专家将受到惩罚，这些任务的门控网络将学习使用不同的专家。与shared-bottom模型相比，MMoE只增加了几个门控网络，门控网络中的模型参数可以忽略不计。因此，整个模型仍然尽可能地享有多任务学习中知识转移的好处。

为了理解为每个任务引入独立的门网络如何帮助模型学习任务特定的信息，我们将其与所有任务共享一个门的模型结构进行了比较。我们称之为OMoE。这是MoE层对shared-bottom多任务模型的直接适应。模型结构图示见图1（B）。

底层


p_1 p_2 p_3
Softmax
激活函数
神经网络
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_3$
第1号
神经网络
第2号
神经网络
第3号
神经网络
q_1 q_2 q_3
Softmax
激活函数
神经网络
用户特征
物品特征
统计特征
场景特征

门控


点击率
神经网络
$p_1\mathbf{x}_1 + p_2\mathbf{x}_2 + p_3\mathbf{x}_3$
对向量做加权平均
p_1 p_2 p_3
权重
点赞率
神经网络
$q_1\mathbf{x}_1 + q_2\mathbf{x}_2 + q_3\mathbf{x}_3$
对向量做加权平均
q_1 q_2 q_3
权重
$\mathbf{x}_1$
$\mathbf{x}_2$
$\mathbf{x}_3$

五、结论

结论

提出了一种新的多任务学习方法——MMoE，该方法从数据中显式地学习任务关系模型。

1. **更好地处理任务相关性较低的场景**。通过在合成数据上的控制实验表明，该方法能更好地处理任务相关性较低的场景。
2. **更易训练**。与基线方法相比，MMoE更容易训练。
3. **真实场景也取得了成功**。通过在基准数据集和真实的大规模推荐系统上的实验，证明了该方法在几种最新的基线多任务学习模型上的成功。
4. **计算效率**。模型的共享部分在服务时间节省了大量计算。所有三种最先进的基线模型都在丧失这种计算效率的情况下学习了任务关系。然而，MMoE模型在很大程度上保留了计算优势，因为门控网络通常是轻量级的，并且专家网络在所有任务之间共享。此外，通过将门控网络制成稀疏top-k门，该模型有可能实现更高的计算效率。希望这项工作能启发其他研究人员使用这些方法进一步研究多任务建模。

六、学习收获

学习收获

- 在点击率预估任务中引入多任务学习。

Thank you