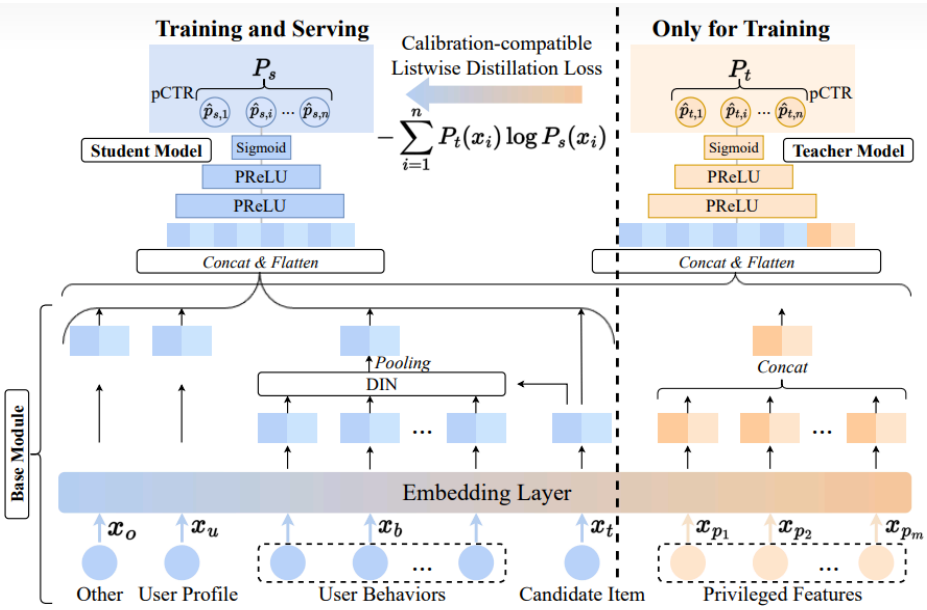
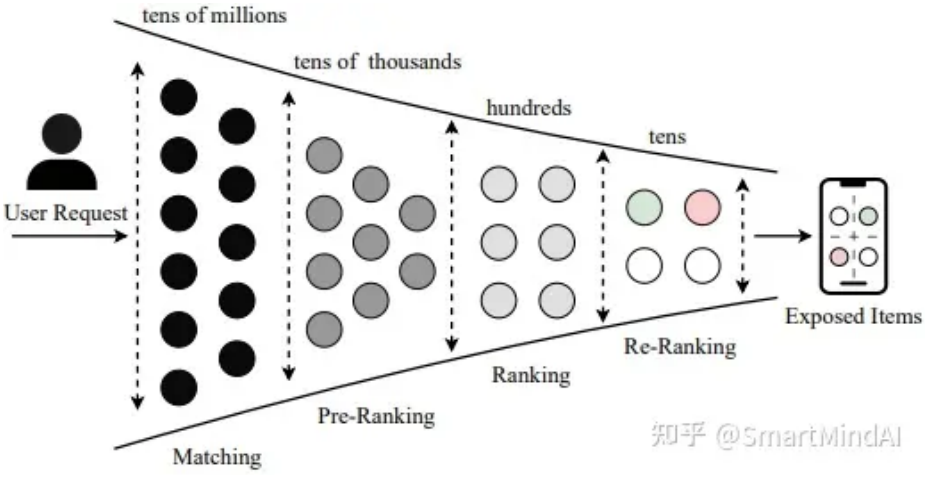


阿里最新研究成果：2023年提出的基于Listwise校准的CTR模型蒸馏技术深度解读与实践探讨

SmartMindAI
专注搜索、广告、推荐、大模型和人工智能最新技术，欢迎关注我
已关注

Introduction

点击率（CTR）预测是在线推荐*系统中的重要组成部分。在接收用户请求时，推荐系统召回一组候选项，并按其排序后向用户显示。在排名阶段，CTR预测模型通常以用户的特征和候选项的特征作为输入。然后预测用户点击候选项的概率。在大多数情况下，在离线训练和在线服务中使用相同的特征集来保证模型一致性。然而，先前的研究也发现，尽管在离线训练期间有用，但有些特征并不容易用于在线服务。例如，在预测用户点击商品后购买的可能性时，详细的商品页面停留时间是一个有用的事件后的特征，只有在离线训练数据中存在，因为在线模型在用户离开页面之前无法知道这个时间。为了便于区分，我们称只在训练期间存在的特征为"特权特征"，而同时在训练和服务期间存在的特征为"非特权特征"。CTR预测任务中有许多具有信息性的特权特征。



在排名阶段，CTR模型从数百个候选项中召回并生成数十个将在重新排名阶段进行精炼并将暴露给用户的物品。因此，排名模型（即排名阶段的CTR模型）在预测用户点击候选物品的可能性之前不能观察到曝光的物品。也就是说，关于与候选物品一起出现在同一页面的物品的特征对排名模型来说是特权特征。在接下来的内容中，我们将这些物品及其特征称为"上下文物品"和"上下文特征"。上下文物品显著影响候选物品的用户点击倾向性。

因此，恰当地利用特权上下文特征可以极大地提高CTR预测性能。请注意，上下文特征对于排名阶段的模型是非特权特征，在那里不存在训练-预测不一致的问题 。为了利用特权特征，先前的研究

力。为了在保留离线-在线一致性的同时融入特权特征，Xu等人提出了特权特征蒸馏（PFD）框架。PFD引入了学生和教师模型，其中教师模型使用了特权和非特权特征以获得更好的训练性能。相反，学生模型仅对非特权特征进行训练，并作为最终的模型用于服务。从教师模型学到的知识（即教师模型的预测）被用作软标签来指导学生模型。

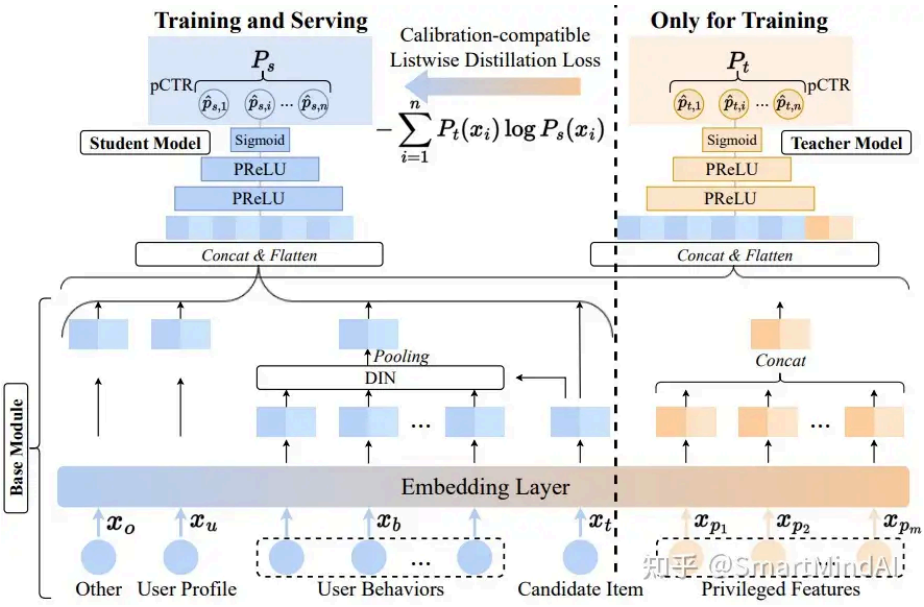
现有PFD方法的知识蒸馏通常通过pointwise损失实现。尽管有效，但pointwise损失分别对待每项基于独立且相同的假设，这与现实情况不符，因为推荐系统中的项目通常是列表格式显示的，点击某项的概率受同一页面上其他项目的直接影响。因此，pointwise损失没有考虑教师模型预测的pCTR（点击概率）排名对于在同一页面上的项目的信息，使得它无法有效地传递教师模型的排序能力。

为了解决上述问题，一种直接的方法是将PFD框架扩展到列表损失作为分馏损失。列表损失对待每个列表作为优化实例，并自然地考虑到物品的pCTR的非独立同分布*性质。实际上，我们实验证明了与点法相比，基于listwise的PFD方法在排序性能上有所提高。然而，当用作分馏损失时，通常使用的列表损失会破坏学生模型预测的概率意义作为pCTR，降低学生的校准能力。对于CTR预测任务，校准能力（即预测点击概率是否与实际点击率对齐）是衡量模型性能的另一个重要因素。例如，在在线广告的 CPC系统中，候选广告（ad）通过有效的每千次印象成本（eCPM）策略进行排名和收费，该策略计算为每个广告被点击的次数除以展示给用户的总次数（impressions）。然而，eCPM模型往往需要大量的数据来捕捉复杂的用户行为模式，并且在某些情况下可能会过拟合。因此，我们提出了一种新的 eCPM 训练框架，即调整优化过程中的惩罚项，使其可以更好地适应小样本场景。实验表明，我们的模型在点击率、eCPM 和 ad 利润等方面都取得了优秀的结果。为了同时提高模型的排名能力和保持其在listwise distillation损失下的校准能力，我们提出了一种适用于CTR预测的Calibration-compatibleLIstwise Distillation (CLID)方法。

具体来说，受Learning-To-Rank（LTR）领域中的calibration-compatible列表式损失的启发，我们首先定义了在PFD语境下distillation损失的*calibration-compatible*特性：使用教师模型的软标签时，distillation loss可以与学生和教师模型的损失同时达到全局最小值。我们展示出广泛使用的列表损失（例如ListNet 和 ListMLE）不能应用于distillation loss，因为它们不是calibration-compatible。

接着，为了实现calibration-compatible的listwise distillation loss，我们仔细设计了损失，通过利用交叉熵*来衡量在同一页面上两个相对项目顺序之间的差异，这两个顺序分别由学生和教师模型生成的归一化pCTR编码。最后，我们提供了CLID中设计的tailored listwise loss确实是*calibration-compatible*的理论证明。

System Overview



图展示了解释性增强网络（CLID）的整体架构。该架构由三个部分组成：基础模块、学生模型和教师模型。基础模块首先将离散的特征ID转换为低维嵌入。然后这些嵌入被聚合以获得固定长度的向量。教师模型使用来自非特权和特权特征的向量作为输入，并输出$\hat{p}_t$。学生模型与教师模型具

知乎

Model Structure

一次点击率预测任务的目标是预测每个用户-物品对点击的概率。在这种情况下，CTR预测模型通常使用非特权特征，包括用户行为序列x_b、用户profilex_u、候选物品x_t和其他特征x_o作为输入，并输出pCTR。

Base Module.

如下图所示，基础模块首先使用嵌入层将高维特征x转换为低维嵌入$e(x)$。这些特征包括非特权特征和特权特征，不同特权特征用m表示为

$$x_{p_1}, x_{p_2}, \ldots, x_{p_m}$$

对于用户行为特征，池化层[+]被应用到嵌入向量列表，将其转换为固定长度的向量。在池化之前，DIN 被用来更好地建模特征交互。通过连接层，我们可以获得非特权表示向量$\mathbf{V}_r$作为学生模型的输入。

$$\mathbf{V}_r = Concat\Big(e(x_o), e(x_u), e(x_t), Pool(\mathrm{DIN}(e(x_b), e(x_t)))\Big)$$

我们将m个特权特征的低维度嵌入相加以获取特权表示向量。

$$\mathbf{V}_p = Concat\Big(e(x_{p_1}), e(x_{p_2}), \ldots, e(x_{p_m})\Big)$$

由于特权特征在训练期间可用，但在服务时不可用，因此我们将其添加到教师模型以引导学生模型的学习。教师模型的输入向量$\mathbf{V}_t$由非特权向量$\mathbf{V}_r$和特权向量$\mathbf{V}_p$组成，如下所示：

$$\mathbf{V}_t = Concat(\mathbf{V}_r, \mathbf{V}_p)$$

Teacher and Student Model.

给定一个样本x_i，点击标签y_i∈0,1$\mathbf{V}_{r,i}$和$\mathbf{V}_{t,i}$分别代表学生模型和教师模型的输入向量。在训练过程中$\mathbf{V}_{t,i}$会被馈送到教师模型中来得到样本x_i的logit$s_{t,i}$，然后采用sigmoid激活函数[+]$\sigma(\cdot)$来计算pCTR$\hat{p}_{t,i}$如下：

$$\hat{p}_{t,i} = \sigma(s_{t,i})$$

为了获得样本x_i的良好可调预测概率，我们采用标准pointwise交叉熵损失（PointCE）。

$$L_t = -y_i \log(\hat{p}_{t,i}) - (1 - y_i) \log(1 - \hat{p}_{t,i})$$

与教师模型不同的是，学生模型在线服务，因此只能使用非特权向量$\mathbf{V}_{r,i}$作为输入。学生模型的输出$\hat{p}_{s,i}$如下：

$$\hat{p}_{s,i} = \sigma(s_{s,i})$$

其中$s_{s,i}$是样本x_i的logit[+]，它是在学生模型之前通过sigmoid激活函数得到的。 我们采用了两种损失函数进行训练。第一种损失类似于教师模型，在其中PointCE损失被用于具有地面真实标签，另一种损失则是简单二分类交叉熵损失[+]（Softmax BCE）。

$$L_{s,CE} = -y_i \log(\hat{p}_{s,i}) - (1 - y_i) \log(1 - \hat{p}_{s,i})$$

第二种损失是一种知识蒸馏损失

$$L_d(\hat{p}_{t,i}, \hat{p}_{s,i})$$

旨在从教师模型中提取知识。原因是教师模型包含特权特征，并且比学生模型表现更好。学生模型的最终损失可以写为如下：

$$L_s = \alpha L_{s,CE}(y_i, \hat{p}_{s,i}) + (1-\alpha) L_d(\hat{p}_{t,i}, \hat{p}_{s,i})$$

其中α是平衡pointwise损失和知识蒸馏损失的重要参数。知识蒸馏损失将在下一部分详细讨论。

Listwise Privileged Features Distillation

Challenges

当我们设计学生模型的知识蒸馏损失时，遇到了两个挑战。

对于知识蒸馏损失，通常会使用pointwise损失，即Teacher-Student Pointwise Cross Entropy (T-S-PCE)。首先，由于不同模型的精度不同，T-S-PCE会导致模型之间的差距较大；其次，pointwise损失在小数据集上的性能不佳，因为它的计算量随着样本数量的增加而线性增加。

为了解决这两个问题，我们提出了一种新的知识蒸馏损失------Anchor Distance Regularization (ADR)。

$$L_d = -\sigma(\frac{s_{t,i}}{\tau}) \log(\sigma(\frac{s_{s,i}}{\tau})) - (1-\sigma(\frac{s_{t,i}}{\tau})) \log(1-\sigma(\frac{s_{s,i}}{\tau}))$$

其中$\tau > 0$表示温度参数。τ越大，预测分布越软。当$\tau \to +\infty$并且对数its的零均值假设时L_d可以等效于直接匹配对数its的MSE损失。 第一项挑战是pointwise损失在从教师模型中提取排序能力时不是最优的。 具体来说，pointwise损失不考虑同一列表中的物品相对顺序，并根据独立同分布假设对待每个项目。但是，推荐任务的数据是非独立同分布的，即一个项目的点击概率受到同一列表中其他项目的影响。因此，仅仅依靠pointwise损失来从教师模型学习CTR预测是不够的。

一些基于listwise的损失已被提出，以便以每组列表作为学习实例优化模型，从而自然地考虑了相对项目顺序的信息。然而，尽管直接利用典型的列表损失可以在pointwise损失的基础上提高模型的排名能力，但它会使学生的预测失去概率意义作为pCTR，破坏模型的校准能力。对于CTR预测来说，这是一个不可接受的问题，从而提出了第二项挑战。

Calibration-compatible Listwise Distillation

如果一个distillation损失L_d能够实现全局最小化，那么当学生和教师模型的PointCE损失都达到全局最小值时，对于任意候选物品x_i都能实现全局最小化。 我们可以轻易证明pointwise损失是可校准的。具体来说，对于每一个列表q，设$P_i = \mathbb{E}[y_i|q, x_i]$是样本$x_i$的真实点击概率。假定我们从其真实标签分布$Y_i$中随机抽取$n_i$个样本$x_i^k$，其中$y_i^k$是第$k$个样本的标签。我们可以看到，PointCE损失在

$$\hat{p}_{s,i} = \sum_{k=1}^{n_i} y_i^k / n_i, \hat{p}_{t,i} = \sum_{k=1}^{n_i} y_i^k / n_i$$

时被最小化，其中

$$\sum_{k=1}^{n_i} y_i^k / n_i = \mathbb{E}[y_i|q, x_i]$$

且$n_i \to +\infty$。因此，学生和教师模型的PointCE损失都是校准的，它们总是能在同时实现全局最小值时达成这一点

$$\hat{p}_{s,i} = \hat{p}_{t,i} = P_i$$

为了证明pointwise损失并不具备可校准性，我们以常用的列表损失ListNet为例。通过推导，我们发现当ListNet作为distillation loss时，它并不满足可校准性。 具体来说，在pointwise损失中，当

$$\hat{p}_{s,i} = \hat{p}_{t,i}$$

时，该损失被最小化。然而，在 ListNet 损失中

$$\hat{p}_{s,i} \neq \hat{p}_{t,i}$$

知乎

$$L_d^{ListNet} = -\sum_{i=1}^{n} \frac{p_{t,i}}{\sum_{j=1}^{n} \hat{p}_{t,j}} \log \frac{\exp(s_{s,i})}{\sum_{j=1}^{n} \exp(s_{s,j})}$$

根据微分规则，我们知道当 ListNet 迁移学习损失达到全局极小值时，其最优解为

$$\hat{p}_{s,i} = \frac{1}{n} \sum_{j=1}^{n} p_{s,j}$$

接下来，我们将分析这个最优解是否满足可校准性。首先，我们需要将损失函数表示为拉普拉斯形式，即

$$\mathcal{L}(\hat{p}) = -\log\left(\prod_{i=1}^{m} (1-\hat{p}_{s,i})^{x_i}\right) + \log\left(\prod_{i=1}^{m} (1-\hat{p}_{t,i})^{y_i}\right)$$

然后，我们可以看到在ListNet中

$$\hat{p}_{s,i} \neq \hat{p}_{t,i}$$

因此即使优化到全局极小值，也不能满足可校准性要求。

$$\frac{\exp(s_{s,i})}{\sum_{j=1}^{n} \exp(s_{s,j})} = \frac{\hat{p}_{t,i}}{\sum_{j=1}^{n} \hat{p}_{t,j}}$$

从公式可以看出，在学生和教师模型的 PointCE 损失都达到全局极小时$L_d^{ListNet}$不是最优（即$\exp(s_{s,i}) \neq P_i$，对于$i = 1 \dots n$）。这表明 ListMLE 不具备可校准性，即使在 ListNet 中应用也不符合预期。

另外，我们注意到另一个常见的列表损失 ListMLE 被广泛应用于 teacher-student learning models。但是，当使用 ListMLE 作为 distillation loss 时，它并不具备可校准性。这是因为在ListMLE中，学生的预测结果$\hat{P}$只是接近教师的预测结果T而非相等。这意味着即使学生和教师的学习目标完全相同，ListMLE仍然不能确保学生的学习成果与教师的期望相匹配。

$$L_d^{ListMLE} = -\sum_{i=1}^{n} \log \frac{\exp(s_{s,\pi_i})}{\sum_{j=i}^{n} \exp(s_{s,\pi_j})}$$

作者提出了一种新的特权特征迁移框架------可校准式列表损失Calibrated Listwise Distillation (CLID)，该框架能够在保持学生模型的可校准能力的同时，利用教师模型的排名能力进行转移。

本文详细介绍了CLID的实现步骤如下：

首先，将学生的预测分数$\hat{p}_{s,i}$与教师的预测分数$\hat{p}_{t,i}$空间投影到概率单纯形上，形成原分布$P_t(x_i)$和得分分布$P_s(x_i)$，具体公式如下：

$$P_t(x_i) = \frac{\exp(-x_i)}{\sum_{j=1}^{n} \exp(-x_j)}, x_j \in -\infty, x_i \leq x_j$$

然后，通过最大化$P_s(x_i)$的对数似然函数[+]来调整学生模型的参数，使学生的预测结果更接近教师的预测结果。同时，为了保证学生模型的可校准性，还需要考虑到在概率单纯形上的约束条件，通过引入权重矩阵W来限制学生模型的参数变化范围。

最后，通过对训练数据进行蒸馏，将教师模型的知识传递给学生模型。蒸馏过程主要包括两个部分：一个是提取教师模型的关键知识，将其编码为一组隐含向量；另一个是重构教师模型的知识，使其能够适应不同的任务。蒸馏过程的具体公式如下

$$P_t(x_i) = \frac{\hat{p}_{t,i}}{\sum_{j=1}^{n} \hat{p}_{t,j}}, P_s(x_i) = \frac{\hat{p}_{s,i}}{\sum_{j=1}^{n} \hat{p}_{s,j}}$$

作者提到了一种叫做"可校准式列表损失"（Calibrated Listwise Distillation，CLID）的损失函数。这个函数用于衡量两个分布之间的差异：一个是来自教师模型的预测分布$P_t(x_i)$，另一个是来自学生模型的预测分布$P_s(x_i)$。 这个损失函数的形式如下

$$L_d^{CLID} = -\sum_{i=1}^{n} P_t(x_i) \log P_s(x_i)$$

$P_t(x_i) = P_s(x_i)$这表明，在最优情况下，学生的模型预测结果与教师模型预测结果完全相同。这是一个理想的目标，但在实际应用中可能很难达到。

作者提出了一种方法，该方法能够同时提高学生模型的排名能力和学生模型的校准能力。为此，他们首先定义了两个损失函数

Experiment Setup

Datasets.

作者进行了大量的实验来展示 CLID 的一般化能力。为了进行实验，他们在两个广为接受的公共数据集和一个生产数据集上进行了大量实验。公共数据集包括 Web30K 和 Istella-S，标签范围从 0（无关）到 4（完美相关），并进行了二进制化处理（等级 1、2、3、4 分别为 1，等级 0 为 0）。

由于公共数据集缺乏位置特征，他们使用上下文特征作为特权特征。生产数据集用于训练生产级别的排名模型，并且采用了同样的上下文特征作为特权特征。为了评估模型的性能，他们在测试集上进行了实验，并报告了平均指标和95%置信区间*。为了保持一致性，作者在所有实验中都使用上下文特征作为特权特征。这部分介绍了两个公共数据集：《Istella-S》和《生产》。《Istella-S》是一个LTR数据集，包含33,018个查询和220个特征，每个查询文档对都有大约103个候选文档关联。经过二值化处理后，标签为0和1的比例分别为82.1%和17.9%。《生产》是生产数据集，是从阿里巴巴在线广告系统的印象日志中采样的。该数据集包含数十亿个样本和数百个特征，用于训练和测试模型。

Baselines.

我们进行了六种基准方法的比较，包括：非PFD方法（PAL和PriDropOut），主流点基PFD方法和我们的扩展列表基PFD方法，以及常用的列表损失（ListMLE和ListNet）作为蒸馏损失的最先进的PFD方法。基准的详细信息如下：

(i) base：使用非特权特征作为输入，并由PointCE损失优化。这是部署在生产系统中的方法。

(ii) PriDropOut：构造了一个浅层塔，使用特权特征作为输入。应用一个丢弃层后，其logits被添加到主要塔的logits中，用于计算pCTR在训练期间。在测试期间，浅层塔被丢弃，pCTR通过主要塔的logits计算。

(iii) PAL：在训练时，通过乘以其概率得分，将其浅层塔的概率分数乘以其非特权特征的主塔概率分数来计算pCTR。在测试期间，只有主塔的概率分数作为pCTR。 (iv) base+点基 ：采用PFD框架提高模型性能，其中教师模型接受非特权和特权特征作为输入。

base+ListMLE：它利用PFD框架从教师模型中提取特权特征，蒸馏损失是ListMLE损失。

(vi)base+ListNet：它利用PFD框架从教师模型中提取特权特征，蒸馏损失ListNet损失。

Evaluation Metrics.

在之前的研究中，我们使用NDCG\@10作为对公共数据集上排名性能的评估标准。NDCG\@10值越高，排名性能就越好。 对于生产数据集，我们通过计算GAUC来评估排名性能。GAUC已被证明与在线性能更一致，并且是我们生产系统的顶级指标。GAUC可以通过公式$GAUC$来计算，其中U表示用户数量$\#$impression(u)和AUC_u分别表示第u个用户的印象次数和AUC。

$$\mathrm{GAUC}=\frac{\sum_{u=1}^{U}\#impressions(u)\times\mathrm{AUC}_u}{\sum_{u=1}^{U}\#impressions(u)}$$

为了衡量校准性能，我们使用了公开和生产数据集上广泛使用的平均LogLoss。按照指示进行计算。LogLoss测量样本级别的校准误差，定义在公式[$LogLoss$中，其中N是样本的数量。

此外，我们还比较了公开数据集上的期望校准误差(ECE)。根据指示，为了计算ECE，我们将文档按模型预测排序并将其分为K个桶，每个桶包含大约相同数量的文档。ECE在公式[ECE_pub]中定义，其中Q是列表的数量n_q和$n_{q,k}$是列表q和列表q的第k个桶中的样本数量$y_{q,k,i}$和$\hat{p}_{q,k,i}$代表列表q中的第i个样本在列表q中的第k个桶中的真实标签和pCTR，分别。在这里，我们设置了$K=10$。 LogLoss和ECE值越低，校准性能越好。

$$\mathrm{ECE} = \frac{1}{Q}\sum_{q=1}^{Q}\frac{1}{n_q}\sum_{k=1}^{K}\left|\sum_{i=1}^{n_{q,k}}(y_{q,k,i} - \hat{p}_{q,k,i})\right.$$

Implementation Details.

我们在PT-Ranking库上进行实验，该库包含了所有公共数据集的代码。我们使用的神经排名模型是三个层的全连接网络，其隐藏层的维度分别为1024、512和256。对于PriDropOut和PAL，浅塔则包含一个具有256个隐藏单元的层。我们对输入进行了变换处理：

$$\log 1p(x) = sign(x)\log(1 + |x|)$$

并应用了批标准化、权重衰减和丢弃。在训练过程中，我们设置了丢弃率为0.5，权重衰减为0.001。在每个方法的验证集上，我们会根据结果调整学习率。在基线和点式方法中，我们也会搜索τ的值范围为

$$\{0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000\}$$

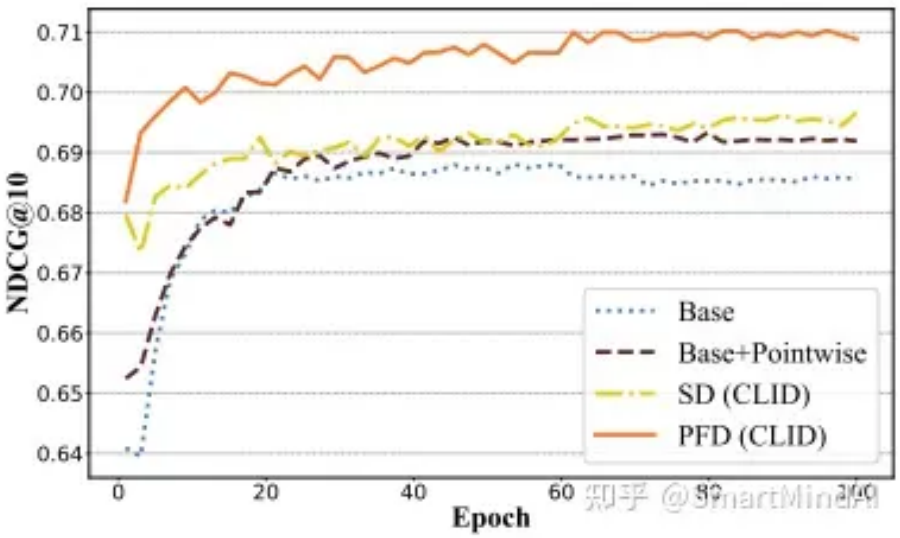
为了平衡排名和校准，我们会在Web30K数据集上将Base+ListMLE的权比设为0.01，在Istella-S数据集上将其设为10。我们还会将Base+ListNet的权比设为100在Web30K数据集上，在Istella-S数据集上将其设为10。对于CLID的数据集，我们将其权比设为1在Web30K数据集上，在Istella-S数据集上将其设为100。对于基线+点式方法，我们将其权比和τ分别设置为10和1在Web30K数据集上，以及1和1在Istella-S数据集上。在生产数据集上，我们使用 DIN 作为所有方法的神经网络结构。对于在公共数据集上的蒸馏方法，我们首先训练教师模型，然后固定收敛的教师模型来指导学生模型。对于生产场景，由于大量数据的存在，教师和学生模型同时更新，并且带有一定的蒸馏损失。

Performance on Public Datasets

Datasets	Web30K			Istella-S		
Metrics	NDCG@10↑	LogLoss↓	ECE↓	NDCG@10↑	LogLoss↓	ECE↓
Base	0.4478 (±0.0004) •	0.6101 (±0.0003) •	0.1629 (±0.0003)	0.6862 (±0.0015) •	0.1174 (±0.0010)	0.0481 (±0.0010)
PriDropOut	0.4472 (±0.0001) •	0.6204 (±0.0016) •	0.1741 (±0.0009) •	0.6922 (±0.0025) •	0.1314 (±0.0028) •	0.0534 (±0.0027) •
PAL	0.4491 (±0.0003)	0.6631 (±0.0028) •	0.1840 (±0.0011) •	0.7070 (±0.0025)	0.1212 (±0.0027) •	0.0465 (±0.0013) ◦
Base+Pointwise	0.4483 (±0.0006) •	0.6095 (±0.0003) •	0.1627 (±0.0003)	0.6911 (±0.0022) •	**0.1129 (±0.0009)** ◦	**0.0454 (±0.0006)** ◦
Base+ListMLE	0.4484 (±0.0010)	0.6134 (±0.0003) •	0.1627 (±0.0003)	0.7021 (±0.0006) •	0.1343 (±0.0003) •	0.0588 (±0.0002) •
Base+ListNet	0.4491 (±0.0004)	0.6095 (±0.0001) •	0.1674 (±0.0002) •	0.6989 (±0.0006) •	0.1177 (±0.0003) •	0.0493 (±0.0012) •
CLID	**0.4495 (±0.0007)**	**0.6090 (±0.0002)**	**0.1626 (±0.0003)**	**0.7084 (±0.0019)**	0.1175 (±0.0008)	0.0489 (±0.0005)

Ablation Study

CLID的排名改进来自于列表式损失、特权特征的蒸馏和知识的蒸馏。

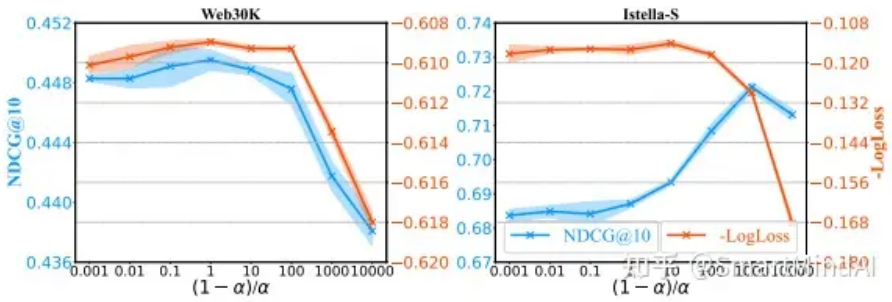


Impact of Weight Ratio

知乎

$\{0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000\}$

范围内的变化。



结果表明，随着权重比例的增加，列表式损失的主导性也增强。我们评估了每个权重比例在NDCG\@10和-LogLoss等度量上的表现，并发现度量值越高，性能越好。然而，我们也注意到，在Web30K和Istella-S数据集上，排名和校准性能逐渐上升然后下降。这可能是因为权重比例过高会削弱PointCE损失的作用，该损失使用点击标签作为真实标签，从而对校准贡献显著；另一方面，这也可能导致学生模型学习了一些来自教师模型的噪声，因此无法达到最优的排名性能。总的来说，当权重比例在1到100范围内时，CLID可以在排名和校准性能之间找到适当的折衷。

Metrics	GAUC↑	LogLoss↓
PriDropOut	+0.32%	+3.15%
PAL	+0.00%	+2.05%
Base+Pointwise	+0.04%	**+0.00%**
Base+ListMLE	+0.37%	+0.78%
Base+ListNet	**+0.38%**	+0.17%
CLID	**+0.38%**	+0.02%

Performance on Production Dataset

Overall Performance.

(i) 列式化PFD方法总是优于点式化PFD方法，进一步证明了列式化距离损失在从教师模型学习中的优势。 (ii) 基于listwise的PFD方法会导致显著的校准退化，只有CLID例外。这是因为只有CLID的列式化距离损失是其中的校准兼容，因此保持了校准能力。CLID的微小增加可能是由于添加的列式化损失需要更多的训练步骤来收敛于Base。 (iii) PriDropOut和PAL损害了模型的泛化能力。具体来说，PriDropOut和PAL在训练和提供服务期间由于数据分布不一致导致了大量的校准退化。PriDropOut表现良好，但是Pal没有效果，因为排名改进缺乏理论保证。这些结果再次证明了CLID可以显著提高模型的排序能力，同时保持其校准能力。

原文《Calibration-compatible Listwise Distillation of Privileged Features for CTR Prediction》

发布于 2024-03-05 10:42 · IP 属地北京

ctr预估　知识蒸馏