

候选集信息主要指各种打分

在电商任务中，候选集信息主要是候选集的全部打分。为说明这一点，我们不妨先把问题设定在最理想的环境下，有如下几点假设。

(1) 强化学习中的折扣系数为 0，单个流量最优化就是全流量最优化。

(2) 具备优化目标相关的全部因素，比如优化目标是 RPM，我们有每个广告的预估 CTR 和 BID。

(3) 所有的预估值都是准确的，例如 CTR 和 BID 完全准确。

(4) 从因素到优化目标的建模是准确的，例如输入三个广告的顺序和相应的预估 CTR、BID 值，建模能计算出准确的 RPM 收益（甚至已经考虑了三个广告的相互影响）。

在理想的环境下，我们不需要引入除候选集的全部打分外的任何信息，只需穷举广告三元组即可。

把假设条件稍微放松，如预估值或者优化目标建模有瑕疵，我们可以利用强化学习主动探索和对标真实奖赏的特性进行修正。

只有当一些假设严重失真的时候，我们才需要引入候选集的全部打分以外的信息，比如：

(1) 当折扣系数大于 0 时，这意味着单流量最优化，并非全流量最优化，而候选集的全部打分只能做到单流量最优化，所以必然引入额外信息；

(2) 优化目标相关因素不完备和部分预估值不准确，这二者其实有一定的重叠，它们都要求引入额外信息修正用户的点击、购买估计。

用一个简单的例子说明以上表述的道理。

开学初，老师说期末考题都在教材范围以内，熟练掌握教材就能得到满分。后来老师说，教材内容有错误，熟练掌握教材得 90 分还是有可能的，想得满分要同时参考教材勘误表。再后来老师又说，期末考题不限于教材范

围，只看教材最多考 70 分，想得满分要另外参考一本国外教材。

候选集的全部打分其实就是教材，教材(候选集的全部打分)是考试(决策)考高分(获得最优奖赏)的基础，其他资料(如用户最近的行为偏好)是教材的纠正或补充。

候选集打分具有自然转移的用户状态

我们已经说明候选集信息必须引入，且候选集全部打分就是用于决策进而优化奖赏的一类重要信息，那么它是一个转移的状态么？答案是肯定的。

候选集的商品构成是召回系统根据用户最近状态挑选出来的，候选集的打分是排序模块根据用户最近的状态和广告自身信息计算出来的，所以候选集的全部打分完全可以被视为对当前用户状态的一种凝练。对于用户相邻两次的到访业务场景，候选集的构成和打分都会发生变化，这就是一个自然的转移过程。

综上所述，我们主要从逻辑分析角度解释了对状态的选择，其实还可以从建模角度理解。我们可以将在动作网络中使用预估值作为状态类比成使用没有端到端训练的 Embedding，如图 6.2 所示。

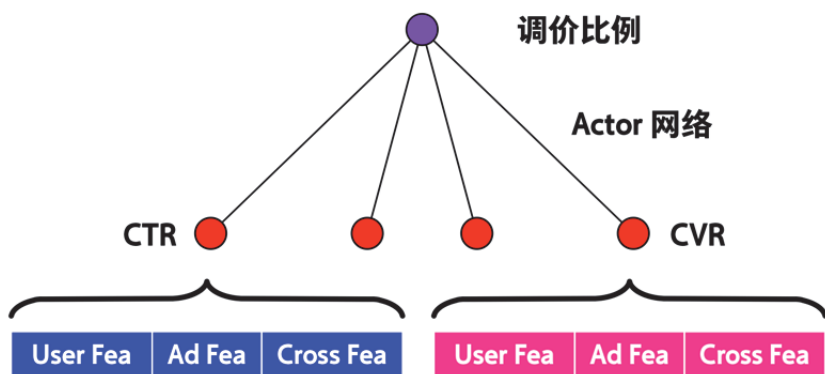


图 6.2 预估值是动作网络的 Embedding 输入

建模粒度

在建模粒度这个问题上，我们关心的是，应该将候选集的全部打分整体输入一个网络中，还是将每个广告的打分信息分别输入一个共享参数的网络中。我们称前者为 session 粒度建模，称后者为 ad 粒度建模。

对于这个问题我们并不陌生，ad 粒度建模就是在 CTR 预估中的一般做法。那么，为什么几乎没有见到过 session 粒度建模的预估方案呢？一个重要原因是，用监督学习直接建模组合优化问题，有很大的难度。

所谓组合优化问题，是指在电商环境下会面对的这样一类问题：**如何从一个大的候选集合中挑选一个子集，按照一定的顺序展现出来，实现流量指标（如 RPM、GMV）最优。**独立假设下的 ad 粒度预估排序是解决这类问题的一种高效、近似的手段。

基于上述认识，下面从强化学习的动作生成和值函数预估角度思考这两种建模粒度。

在 ad 粒度建模中，动作是每个 ad 的调价比例，和 CTR 模型输出预估值类似。

在 session 粒度建模中，动作的含义应该是**某种区分方式**，这种方式能够将 3 个最优广告和其他 397 个广告分开。这类动作包括但不限于以下几种可能：

（1）直接输出哪 3 个是最优广告。

（2）输出 400 维奖赏预估值向量，排序取最优的 3 个。

（3）输出 400 维调价比例向量，用调价比例和统一的公式给 400 个广告打分，排序取最优的 3 个。

（4）输出一个权重，用这个权重和统一的公式给 400 个广告打分，排序取最优的 3 个，如图 6.3 所示。

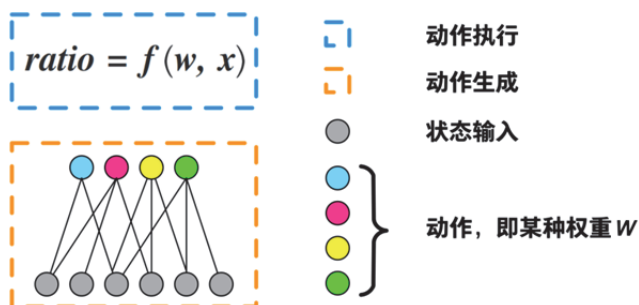


图 6.3 session 粒度建模举例

这四种动作网络的建模难度是一样大的，而第一种网络显然是想用监督学习直接建模组合优化问题。值函数预估上的分析与此类似，考虑到 session 粒度值函数预估实际上要完成两个任务：400 选 3 和 3 广告奖赏的预估。

除建模难度外，session 粒度建模还会导致动作生成（图 6.3 中橙色框）和动作执行（图 6.3 中蓝色框）两部分在优化上的耦合。强化学习并不会训练 f 函数，而 f 函数自身也有优化空间，如果我们调整了 f 或者 x ，那么动作生成部分的状态输入或网络设计也很可能要随之变化。

总体而言，我们面对的问题本质上是组合优化问题，组合优化问题自身的难度不会随着建模形式的变化而自动消除，它必然栖身于我们系统的某个环节中。session 粒度建模是将组合优化问题的难度放入监督学习建模（值函数预估与动作生成）中，而 ad 粒度建模则将其放入单广告调价（独立假设，近似组合优化）与强化学习探索（探索组合优化解空间）中。

其实在 ad 粒度和 session 粒度之间还有一个 ad 组合粒度，GroupCTR 之类的任务就是尝试在组合粒度上做预估，以弥补独立预估假设的不足。无论是 ad 粒度，还是 ad 组合粒度，本质上都是从监督学习中剥离组合优化的难度，只是两者对组合优化问题的近似程度不同。图 6.4 展示了在组合优化问题上的一种可持续迭代的方法论：一方面不断接近现有近似建模方案的效果天花板，另一方面让建模更接近原问题的组合优化本质。

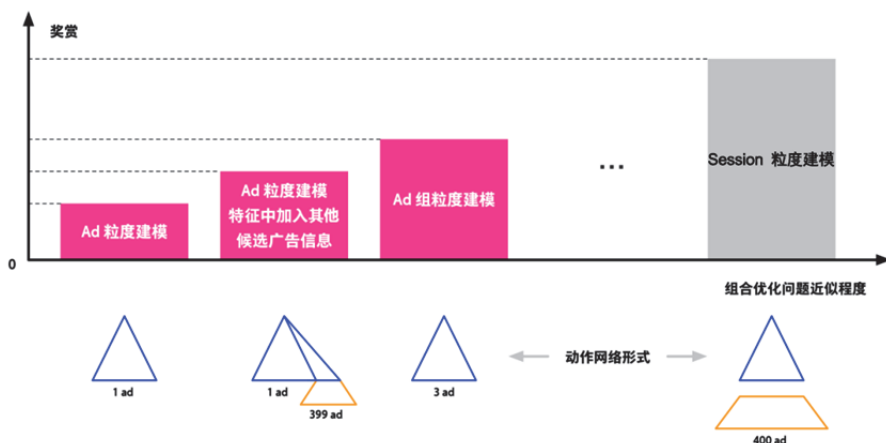


图 6.4 组合优化问题的可持续迭代方法论

最后，考虑到我们的动作是每个 ad 的调价比例，特别是最终要为每个 ad 单独计算 ECPM 进行排序，所以选择 ad 粒度建模。

6.3 模型选择

在模型选择方面，我们主要考虑三类方法。

- (1) 值函数估计方法，如 DQN^[89]。
- (2) Likelihood Ratio 策略梯度方法，如 Actor-Critic^[65]。
- (3) Pathwise Derivative 策略梯度方法，如 DDPG^[79]。

DQN

DQN 是一种最直观的、将现有监督学习预估方法和强化学习结合，解决组合优化问题的方法，图 6.1 的上半部分其实就是 DQN 方案。在这种方案中：

- (1) 强化学习的探索转为监督学习解决训练样本不充分的问题。