22 | 强化学习：让推荐系统像智能机器人一样自主学习

你好，我是王喆。这节课我们继续来讲深度推荐模型发展的前沿趋势，来学习强化学习（Reinforcement Learning）与深度推荐模型的结合。

强化学习也被称为增强学习，它在模型实时更新、用户行为快速反馈等方向上拥有巨大的优势。自从 2018 年开始，它就被大量应用在了推荐系统中，短短几年时间内，微软、美团、阿里等多家一线公司都已经有了强化学习的成功应用案例。

虽然，强化学习在推荐系统中的应用是一个很复杂的工程问题，我们自己很难在单机环境下模拟，但理解它在推荐系统中的应用方法，是我们进一步改进推荐系统的关键点之一，也是推荐系统发展的趋势之一。

所以这节课，我会带你重点学习这三点内容：一是强化学习的基本概念；二是，我会以微软的 DRN 模型为例，帮你厘清强化学习在推荐系统的应用细节；三是帮助你搞清楚深度学习和强化学习的结合点究竟在哪。

强化学习的基本概念

强化学习的基本原理，简单来说，就是**一个智能体通过与环境进行交互，不断学习强化自己的智力，来指导自己的下一步行动，以取得最大化的预期利益**。

事实上，任何一个有智力的个体，它的学习过程都遵循强化学习所描述的原理。比如说，婴儿学走路就是通过与环境交互，不断从失败中学习，来改进自己的下一步的动作才最终成功的。再比如说，在机器人领域，一个智能机器人控制机械臂来完成一个指定的任务，或者协调全身的动作来学习跑步，本质上都符合强化学习的过程。

为了把强化学习技术落地，只清楚它的基本原理显然是不够的，我们需要清晰地定义出强化学习中的每个关键变量，形成一套通用的技术框架。对于一个通用的强化学习框架来说，有这么六个元素是必须要有的：

1. **智能体（Agent）**：强化学习的主体也就是作出决定的“大脑”；
2. **环境（Environment）**：智能体所在的环境，智能体交互的对象；
3. **行动（Action）**：由智能体做出的行动；
4. **奖励（Reward）**：智能体作出行动后，该行动带来的奖励；
5. **状态（State）**：智能体自身当前所处的状态；
6. **目标（Objective）**：指智能体希望达成的目标。

为了方便记忆，我们可以用一段话把强化学习的六大要素串起来：一个**智能体**身处在不断变化的**环境**之中，为了达成某个**目标**，它需要不断作出**行动**，行动会带来好或者不好的**奖励**，智能体收集起这些奖励反馈进行自我学习，改变自己所处的**状态**，再进行下一步的行动，然后智能体会持续这个“**行动 - 奖励 - 更新状态**”的循环，不断优化自身，直到达成设定的目标。

这就是强化学习通用过程的描述，那么，对于推荐系统而言，我们能不能创造这样一个会自我学习、自我调整的智能体，为用户进行推荐呢？事实上，微软的 DRN 模型已经实现这个想法了。下面，我就以

DRN 模型为例，来给你讲一讲在推荐系统中，强化学习的六大要素都是什么，强化学习具体又是怎样应用在推荐系统中的。

强化学习推荐系统框架

强化学习推荐模型 DRN（Deep Reinforcement Learning Network，深度强化学习网络）是微软在 2018 年提出的，它被应用在了新闻推荐的场景上，下图 1 是 DRN 的框架图。事实上，它不仅是微软 DRN 的框架图，也是一个经典的强化学习推荐系统技术框图。

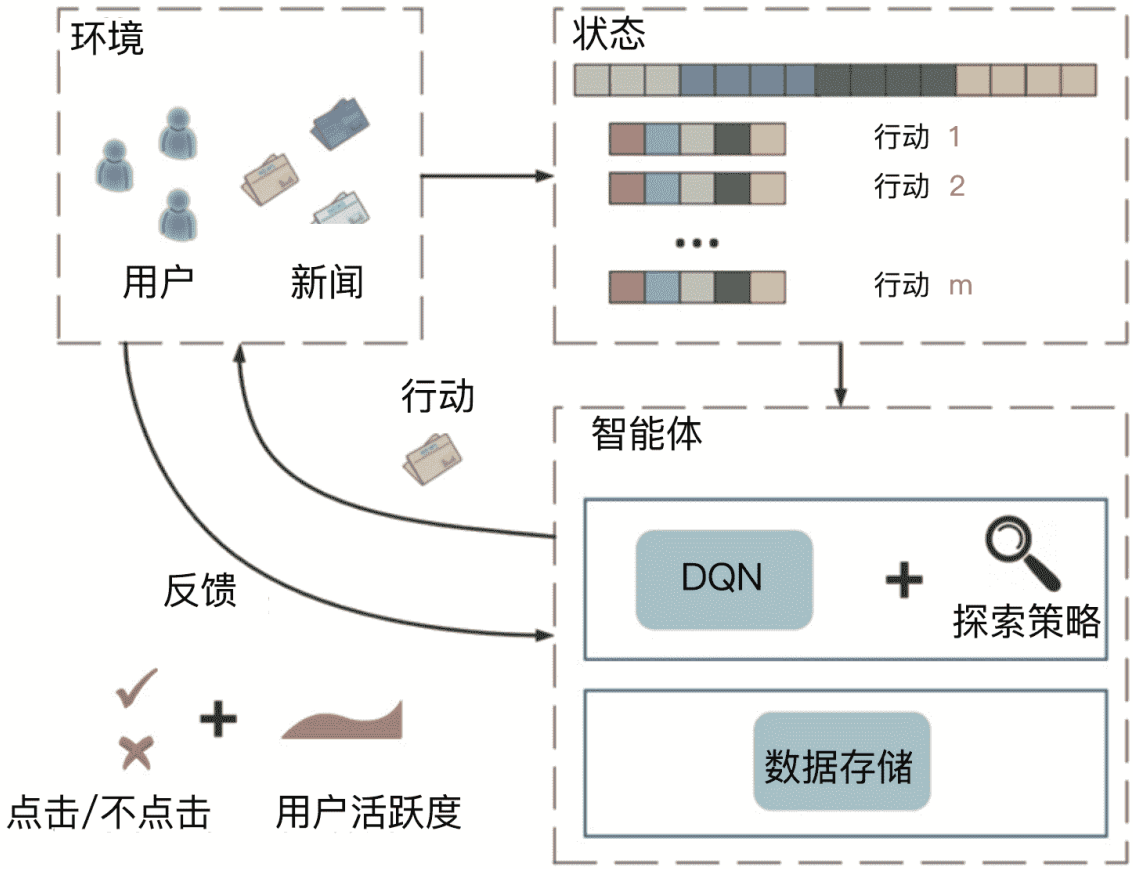

图1 深度强化学习推荐系统框架

从这个技术框图中，我们可以清楚地看到强化学习的六大要素。接下来，我就以 DRN 模型的学习过程串联起所有要素，来和你详细说说

这六大要素在推荐系统场景下分别指的是什么，以及每个要素的位置和作用。

在新闻的推荐系统场景下，DRN 模型的第一步是初始化推荐系统，主要初始化的是推荐模型，我们可以利用离线训练好的模型作为初始化模型，其他的还包括我们之前讲过的特征存储、推荐服务器等等。

接下来，推荐系统作为智能体会根据当前已收集的用户行为数据，也就是当前的状态，对新闻进行排序这样的行动，并在新闻网站或者 App 这些环境中推送给用户。

用户收到新闻推荐列表之后，可能会产生点击或者忽略推荐结果的反馈。这些反馈都会作为正向或者负向奖励再反馈给推荐系统。

推荐系统收到奖励之后，会根据它改变、更新当前的状态，并进行模型训练来更新模型。接着，就是推荐系统不断重复“排序 - 推送 - 反馈”的步骤，直到达成提高新闻的整体点击率或者用户留存等目的为止。

为了方便你进行对比，我也把这六大要素在推荐系统场景下的定义整理在了下面，你可以看一看。

六大要素	概念
智能体（Agent）	推荐系统本身，它包括基于深度学习的推荐模型， 与支持推荐模型的存储、计算等软硬件设备
环境 （Environment）	由新闻网站或App、用户组成的整个推荐系统外部环境。 在环境中，用户接收推荐的结果并做出相应反馈
行动（Action）	对一个新闻推荐系统来说，“行动”指的就是推荐系统推荐新闻给用户的动作
奖励（Reward）	用户收到推荐结果后，进行正向的或负向的奖励，反馈给推荐系统，例如点击行为被认为是一个典型的正向奖励，曝光未点击则是负奖励的信号。 此外，用户的活跃程度，用户打开应用的间隔时间也被认为是有价值的奖励信号
状态（State）	状态这里指的是推荐系统对环境及自身当前所处具体情况的刻画。 在新闻推荐场景中，状态可被认为是所有已经收集到的行动和相应奖励的总和
目标（Objective）	推荐系统要达成的目标，一般是提高整体点击率、增加用户留存等

到这里，你有没有发现强化学习推荐系统跟传统推荐系统相比，它的主要特点是什么？其实，就在于强化学习推荐系统始终在强调“持续学习”和“实时训练”。它不断利用新学到的知识更新自己，做出最及时的调整，这也正是将强化学习应用于推荐系统的收益所在。

我们现在已经熟悉了强化学习推荐系统的框架，但其中最关键的部分“智能体”到底长什么样呢？微软又是怎么实现“实时训练”的呢？接下来，就让我们深入 DRN 的细节中去看一看。

深度强化学习推荐模型 DRN

智能体是强化学习框架的核心，作为推荐系统这一智能体来说，推荐模型就是推荐系统的“大脑”。在 DRN 框架中，扮演“大脑”角色的是 Deep Q-Network (深度 Q 网络，DQN)。其中，Q 是 Quality 的简称，指通过对行动进行质量评估，得到行动的效用得分，来进行行动决策。

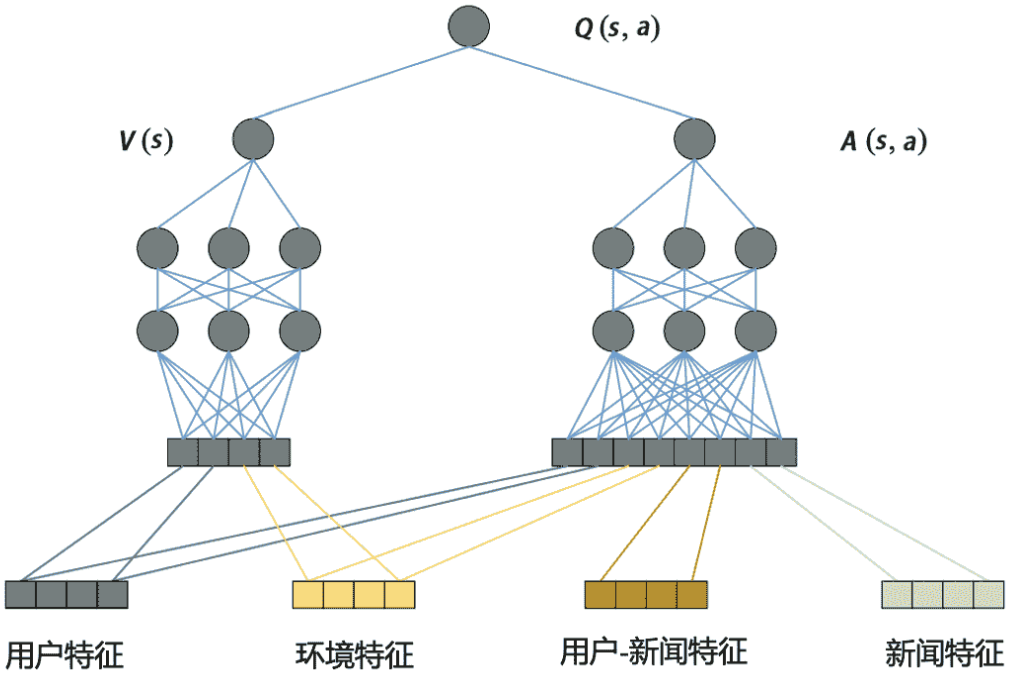

图2 DQN的模型架构图

DQN 的网络结构如图 2 所示，它就是一个典型的双塔结构。其中，用户塔的输入特征是用户特征和场景特征，物品塔的输入向量是所有的用户、环境、用户 - 新闻交叉特征和新闻特征。

在强化学习的框架下，用户塔特征向量因为代表了用户当前所处的状态，所以也可被视为**状态向量**。物品塔特征向量则代表了系统下一步要选择的新闻，我们刚才说了，这个选择新闻的过程就是智能体的“行动”，所以物品塔特征向量也被称为**行动向量**。

双塔模型通过对状态向量和行动向量分别进行 MLP 处理，再用互操作层生成了最终的行动质量得分 Q(s,a)，智能体正是通过这一得分的高低，来选择到底做出哪些行动，也就是推荐哪些新闻给用户的。

其实到这里为止，我们并没有看到强化学习的优势，貌似就是套用了强化学习的概念把深度推荐模型又解释了一遍。别着急，下面我要讲

的 DRN 学习过程才是强化学习的精髓。

DRN 的学习过程

DRN 的学习过程是整个强化学习推荐系统框架的重点，正是因为可以在线更新，才使得强化学习模型相比其他“静态”深度学习模型有了更多实时性上的优势。下面，我们就按照下图中从左至右的时间轴，来描绘一下 DRN 学习过程中的重要步骤。

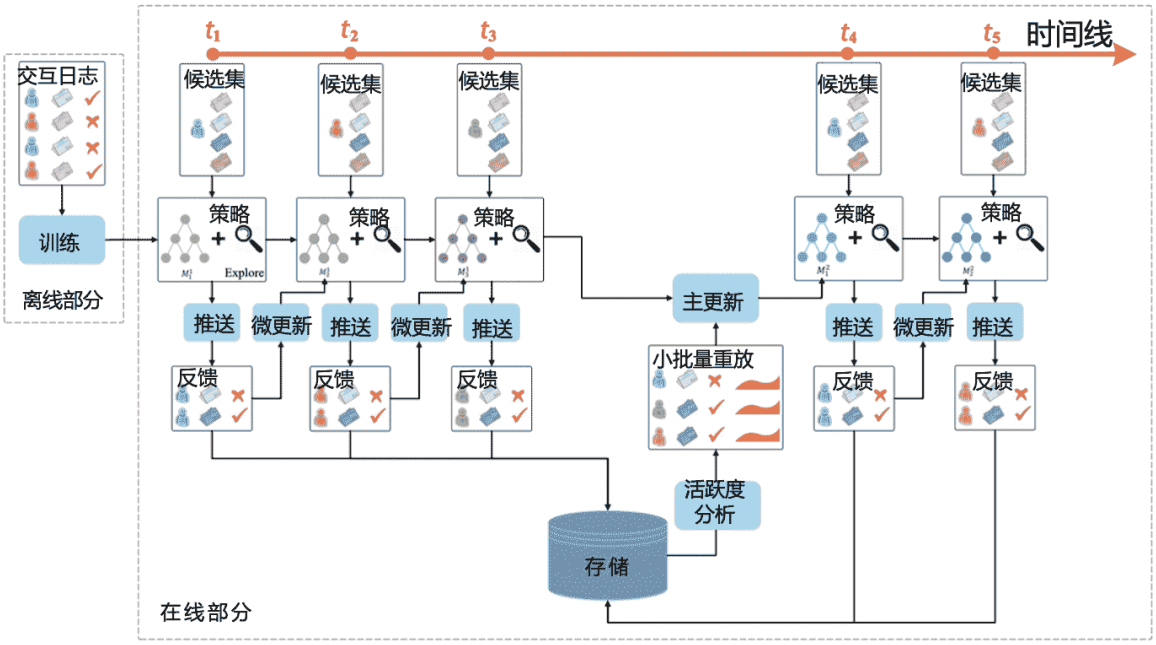

图3 DRN的学习过程

我们先来看离线部分。DRN 根据历史数据训练好 DQN 模型，作为智能体的初始化模型。

而在线部分根据模型更新的间隔分成 n 个时间段，这里以 t1 到 t5 时间段为例。首先在 t1 到 t2 阶段，DRN 利用初始化模型进行一段时间的推送服务，积累反馈数据。接着是在 t2 时间点，DRN 利用 t1 到 t2 阶段积累的用户点击数据，进行模型微更新（Minor update）。

最后在 t4 时间点，DRN 利用 t1 到 t4 阶段的用户点击数据及用户活跃度数据，进行模型的主更新（Major update）。时间线不断延长，我们就不断重复 t1 到 t4 这 3 个阶段的操作。

这其中，我要重点强调两个操作，一个是在 t4 的时间点出现的模型主更新操作，我们可以理解为利用历史数据的重新训练，用训练好的模型来替代现有模型。另一个是 t2、t3 时间点提到的模型微更新操作，想要搞清楚它到底是怎么回事，还真不容易，必须要牵扯到 DRN 使用的一种新的在线训练方法，Dueling Bandit Gradient Descent algorithm（竞争梯度下降算法）。

DRN 的在线学习方法：竞争梯度下降算法

我先把竞争梯度下降算法的流程图放下了下面。接下来，我就结合这个流程图，来给你详细讲讲它的过程和它会涉及的模型微更新操作。

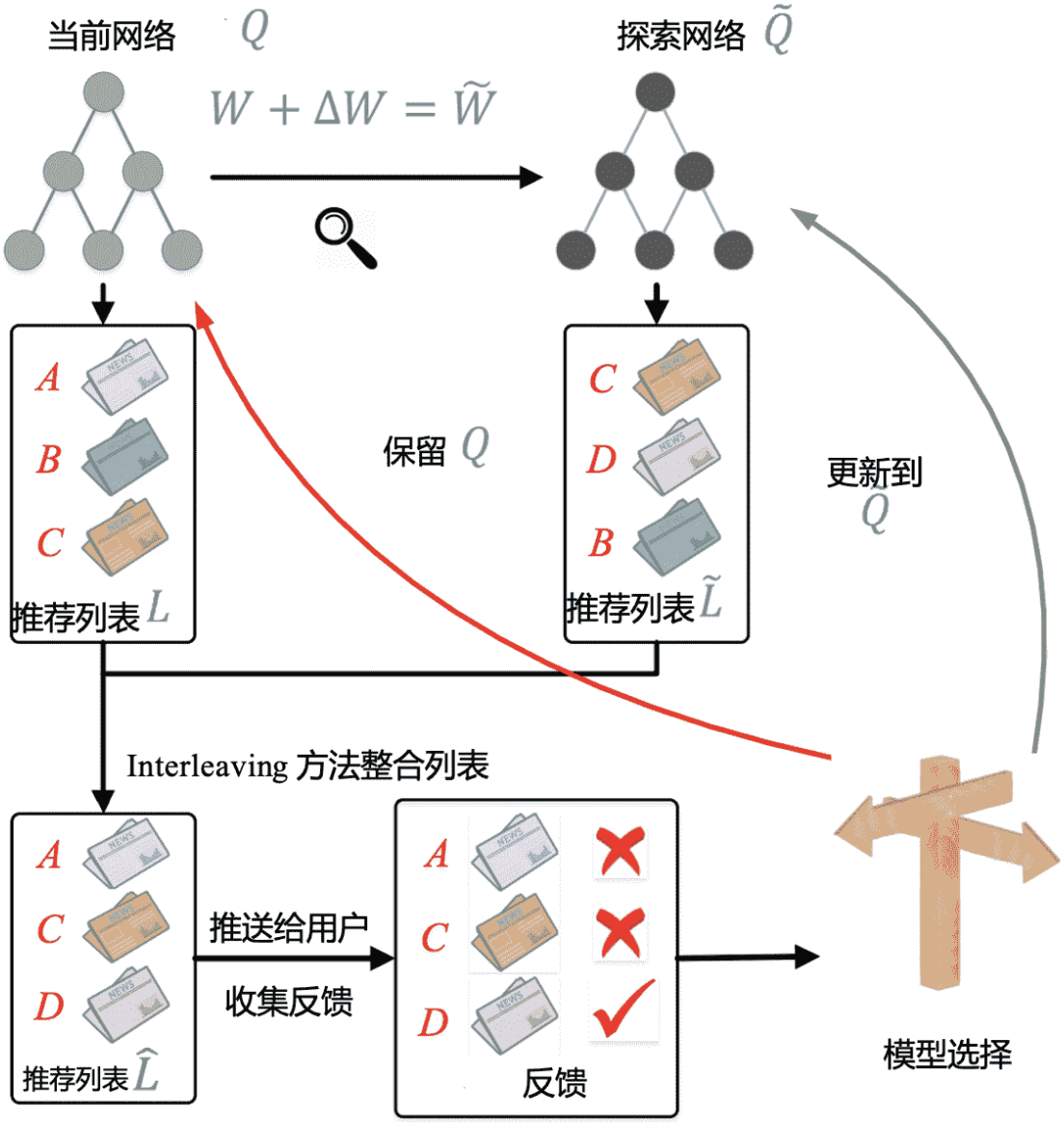

图4 DRN的在线学习过程

DRN 的在线学习过程主要包括三步，我带你一起来看一下。

第一步，对于已经训练好的当前网络 Q，对其模型参数 **W** 添加一个较小的随机扰动，得到一个新的模型参数，这里我们称对应的网络为探索网络 Q~。

在这一步中，由当前网络 Q 生成探索网络 ，产生随机扰动的公式 1 如下：

$$\Delta W = \alpha \cdot rand(-1, 1) \cdot W$$

其中，α 是一个探索因子，决定探索力度的大小。rand(-1,1) 产生的是一个[-1,1]之间的随机数。

第二步，对于当前网络 Q 和探索网络 Q~，分别生成推荐列表 L 和 L~，再将两个推荐列表用间隔穿插（Interleaving）的方式融合，组合成一个推荐列表后推送给用户。

最后一步是实时收集用户反馈。如果探索网络 Q ~ 生成内容的效果好于当前网络 Q，我们就用探索网络代替当前网络，进入下一轮迭代。反之，我们就保留当前网络。

总的来说，DRN 的在线学习过程利用了“探索”的思想，其调整模型的粒度可以精细到每次获得反馈之后，这一点很像随机梯度下降的思路：虽然一次样本的结果可能产生随机扰动，但只要总的下降趋势是正确的，我们就能够通过海量的尝试最终达到最优点。DRN 正是通过这种方式，让模型时刻与最“新鲜”的数据保持同步，实时地把最新的奖励信息融合进模型中。模型的每次“探索”和更新也就是我们之前提到的模型“微更新”。

到这里，我们就讲完了微软的深度强化学习模型 DRN。我们可以想这样一个问题：这个模型本质上到底改进了什么？从我的角度来说，它最大的改进就是把模型推断、模型更新、推荐系统工程整个一体化了，让整个模型学习的过程变得更高效，能根据用户的实时奖励学到

新知识，做出最实时的反馈。但同时，也正是因为工程和模型紧紧地耦合在一起，让强化学习在推荐系统中的落地并不容易。

既然，说到了强化学习的落地，这里我还想再多说几句。因为涉及到了模型训练、线上服务、数据收集、实时模型更新等几乎推荐系统的所有工程环节，所以强化学习整个落地过程的工程量非常大。这不像我们之前学过的深度学习模型，只要重新训练一下它，我们就可以改进一个模型结构，强化学习模型需要工程和研究部门通力合作才能实现。

在这个过程中，能不能有一个架构师一样的角色来通盘协调，就成为了整个落地过程的关键点。有一个环节出错，比如说模型在做完实时训练后，模型参数更新得不及时，那整个强化学习的流程就被打乱了，整体的效果就会受到影响。

所以对我们个人来说，掌握强化学习模型的框架，也就多了一个发展的方向。那对于团队来说，如果强化学习能够成功落地，也一定证明了这个团队有着极强的合作能力，在工程和研究方向上都有着过硬的实践能力。

小结

强化学习是近来在学术界和业界都很火的话题，它起源于机器人领域。这节课，我们要重点掌握强化学习的通用过程，以及它在深度学习中的应用细节。

简单来说，强化学习的通用过程就是训练一个智能体，让它通过与环境进行交互，不断学习强化自己的智力，并指导自己的下一步行动，以取得最大化的预期利益。这也让强化学习在模型实时更新，用户行为快速反馈等方向上拥有巨大的优势。

但强化学习的落地并不容易，整个落地过程的工程量非常大。现阶段，我们只需要以微软的 DRN 模型作为参考，重点掌握强化学习在推荐系统领域的应用细节就可以了。

一个是 DRN 构建了双塔模型作为深度推荐模型，来得出“行动得分”。第二个是 DRN 的更新方式，它利用“微更新”实时地学习用户的奖励反馈，更新推荐模型，再利用阶段性的“主更新”学习全量样本，更新模型。第三个是微更新时的方法，竞争梯度下降算法，它通过比较原网络和探索网络的实时效果，来更新模型的参数。

为了方便你复习，我们把这节课的重点知识总结在了下面的表格中，你可以看看。

知识点	关键描述
强化学习的典型过程	一个智能体通过与环境进行交互，不断学习强化自己的智力，并指导自己的下一步行动，以取得最大化的预期利益
强化学习六要素	智能体、环境、行动、奖励、状态、目标
DRN推荐模型的结构	双塔模型中用户塔的输入特征是用户特征和场景特征， 物品塔的输入向量是所有的用户、环境、 用户–新闻交叉特征和新闻特征
DRN更新的过程	微更新学习实时用户奖励数据 主更新进行全量数据训练和更新
DRN进行微更新的方法	竞争梯度下降算法

课后思考

DRN 的微更新用到了竞争梯度下降算法，你觉得这个算法有没有弊端？你还知道哪些可以进行模型增量更新或者实时更新的方法吗？

欢迎把你的思考和疑问写在留言区，如果你的朋友们也在关注强化学习在推荐系统上的发展，那不妨也把这节课转发给他们，我们下节课见！