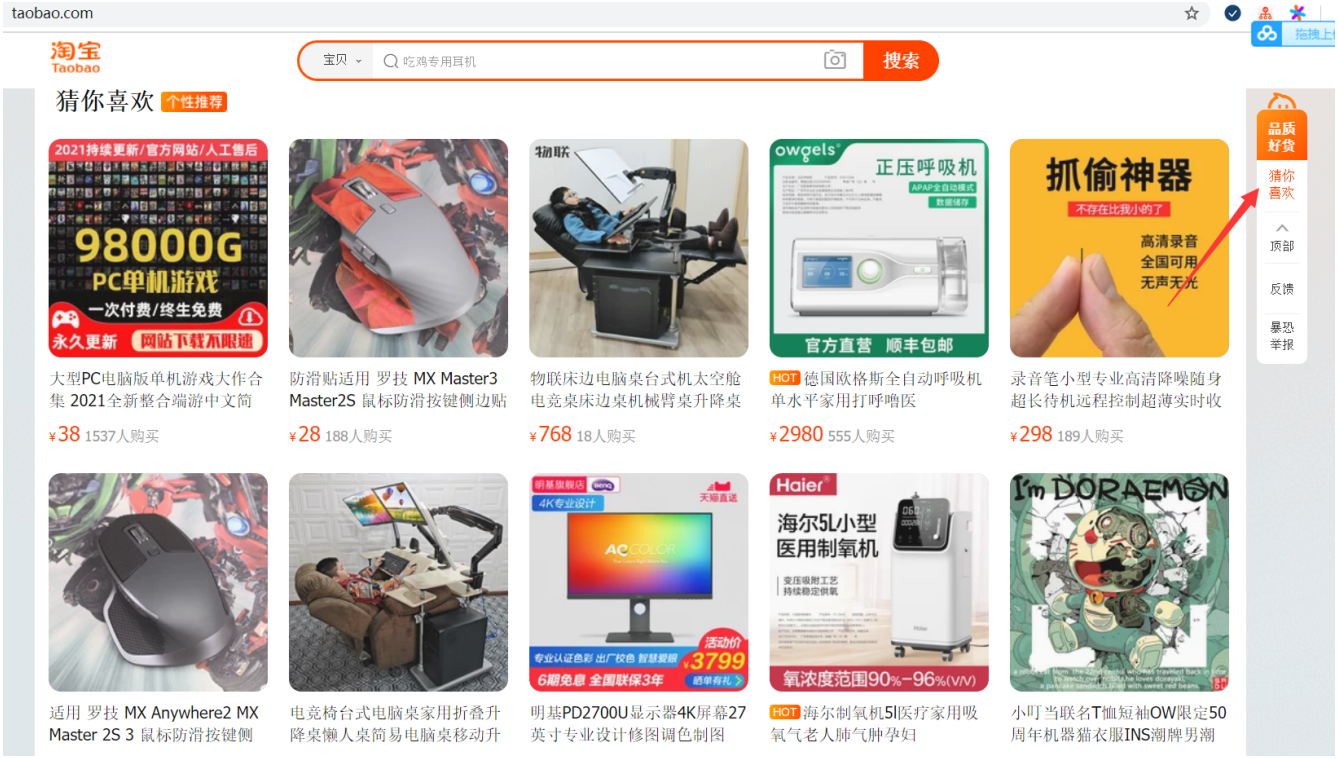
电商推荐系统理论基础

图灵：楼兰

一、推荐系统功能及作用介绍

1、关于推荐系统

我们整个电商项目的实战课程，到目前为止，已经参照互联网的最新潮流技术，构建出了一个功能比较全面、性能也非常强的电商网站。但是，相比于其他互联网大厂的各大电商平台，还缺少一个很大的功能模块，那就是推荐系统。比如像淘宝首页的猜你喜欢：



淘宝单品页的"看了又看"，"买了又买"。

看了又看

¥78.00

适用 罗技 Craft MXKEYS 键盘...

¥39.00

防滑贴 适用 罗技 G102 G304 ...

¥39.00

火线竞技 夹铺路爪 野蜂 3号 ...

还有像京东、唯品会、亚马逊等等各大电商，都有自己的推荐系统。而随着互联网的技术发展，还出现了很多跨平台的推荐系统。例如大家有没有过这样的经历，在淘宝浏览了一个商品后，去百度首页上逛一逛就会发现百度经常会给你推荐一些你查看过或者购买过的商品。所以，现在的推荐系统，都是一个稳定的推荐核心加上上层大量的业务渠道组成的一个庞大的系统。我们这几天来跟大家讨论的，就是这个后台的核心推荐系统。

那大家现在想一下，如果现在要你来设计一个推荐系统，你会怎么去推荐？你的第一感觉是不是这东西很简单嘛，只要随便推荐出几个商品就可以了。例如就到订单表中找销售量最靠前的产品。像这样：

```
select product_id,count(1) count from oms_order_item group by product_id order by count desc limit 5
```

然后给所有人展示销售排名前5名的这几个商品，这样也能实现一个推荐系统。

这么做确实是没错，这也确实就算是一个推荐系统了。但是是不是直觉就会觉得这样的推荐系统不太好？为什么？因为没有考虑到用户的喜好对吧。如果是我们一个程序员来逛这个电商网站，结果每次看到的都是推荐卫生巾，这样是不是会有点尴尬？

要分析这个问题，肯定不能只是从直觉出发来进行分析。

首先，我们需要了解推荐系统的作用。

谈到推荐系统，就不得不提一个互联网经常提到的问题"**信息过载**"。互联网的兴起，使我们获取信息的途径得到了极大的扩展。但是，随着互联网的迅速发展，网上的信息量呈现爆炸式的增长。过多的信息，反而使得用户很难获得对自己真正有用的那部分信息，信息的利用率反而降低了，这就是所谓的信息过载问题。

而解决信息过载的问题一个非常有潜力的方法就是推荐系统。他是根据用户的信息需求、兴趣爱好等，将用户感兴趣的信息、产品等推荐给用户的个性化信息推荐系统。一个好的推荐系统，对于用户，能够让用户更快更好的获取到自己需要的信息；对于信息，可以让信息能够更快更好的推动到喜欢它的用户手中；而对于网络平台，可以更有效的保留用户资源，提升用户粘性。他是一个让用户、信息、平台三方合力共赢的产品。

事实上，我们从身边的很多互联网产品中看到推荐系统的价值。比如，大家可以看到，我比较喜欢鼠标，淘宝也经常给我推荐鼠标。那很多时候，我上淘宝就只是随意逛逛，并不是就要去买鼠标。而有了这个推荐系统之后，我就会不经意间就会点进去，看到喜欢的就购买一个。这样，对于淘宝，就带来了非常多的浏览和购买记录。这些好处，从个人方面来看，或许还收益有限，因为我毕竟还是看的时候多，买的时候少。但是如果从整体宏观来考虑，从最早的亚马逊，再到后来的雅虎，再到现在京东、天猫、唯品会，推荐系统都给他们带来了实实在在的巨大收益。 而在2011年9月，百度世界大会2011上，李彦宏更是将推荐引擎与云计算、搜索引擎并列为未来互联网重要战略规划以及发展方向。现在的百度首页也在逐步实现个性化，智能地推荐出用户喜欢的网站和经常使用的APP。

所以，推荐系统的作用不是简单的选出几个商品就可以了的。事实上，推荐系统是一个门槛比较低，但是技术深度非常深的一个领域。

然后，回到我们这个简单的推荐系统实现。

站在电商网站的角度，只是从销量一个方面进行推荐，这样的推荐系统肯定起不到吸引客户消费的目的。所以，好的推荐系统必须要考虑到用户的喜好以及产品的特点。用户经常浏览以及购买些什么商品？购买产品最多的是年轻人还是老年人？那电商网站怎么知道用户的喜好以及产品的受众呢？这就必须要基于大量用户以及产品的数据。所以，我们这次的课程就是带大家一起来收集、处理、计算这些大量的用户以及产品的数据，通过机器学习的方式，形成一个有业务价值的推荐系统。

2、关于本课程：

我们这个课程重点是在带大家了解推荐系统的设计实现思想，并最终以机器学习的技术，带大家实现一个推荐系统。课程中介绍到的推荐系统，相比简单的SQL数据分析，会更灵活一点，因为他已经具备了根据数据进行学习，进化的能力。但是相比目前最前沿的一些推荐系统，肯定还是有点差距。

这个课程既然是用机器学习，就还是需要涉及到一些大数据方面以及机器学习方面的基础知识和案例。在这次的课程上，会尽量简化这一方面的内容，只是使用Spark来进行示例讲解。有兴趣的同学可以了解下图灵官网上的大数据录播课程。并且，后续我们也会针对大数据方向的每一个组件和技术，整理新的优化课程。

在高级语言方面，机器学习其实是一个重思想而不重语言的技术方向，各种主流语言，像Java、Python、R、C以及Scala等，都有针对机器学习的实现框架。而在所有的实现框架中，python中的sklearn框架是封装得最全面，技术门槛最低的。所以在课程中，会适当引入部分的Python样例。对于这一部分样例，大家不用担心不懂python，我们课上重点关注的是计算的思想，而不是计算的过程和结果。如果有兴趣了解更多细节的同学，也可以关注我们图灵学院的python相关课程。

另外给大家推荐一个语言 Jython，作为知识的扩展。官方网站 https://www.jython.org/ ，这是Python的一个纯JAVA实现，兼容了JAVA和Python。可以在java中直接运行python代码，也可以在python中运行java代码，然后都通过JVM执行。这个项目非常有趣，java语言的强大高效，结合python庞大的第三方类库，可以产生非常多的精彩使用场景。像HBase就专门提供了基于Jython的客户端实现。有兴趣的同学可以去了解下。

3、课程基础环境搭建

本课程需要使用到python的sklearn库，以及spark。

python环境搭建：采用python3版本，推荐安装集成开发环境anaconda，一键安装，省时省力。包含常见库，以及一些常用的集成开发环境。 开发IDE，建议使用PyCharm，设置使用anaconda中的python替代内置的python。工程化项目代码管理。

spark: 采用Spark 3.1.1版本。Spark是基于scala语言开发，在服务器上运行，只需要安装JDK就可以了。我们课程中会要基于源码进行简单调试，还需要在本地安装scala的语言包， 采用对应的2.12版本。

二、推荐系统核心问题分析

1、到底什么是推荐系统?

关于推荐系统，百度百科的解释是：利用电子商务网站向客户提供商品信息和建议，帮助用户决定应该购买什么产品，模拟销售人员帮助客户完成购买过程。个性化推荐是根据用户的兴趣特点和购买行为，向用户推荐用户感兴趣的信息和商品。其实对于推荐系统最直接也是最简单的理解，就是电商网站向用户推荐商品。

随着互联网的不断发展，其实推荐系统也会逐渐扩展出非常多的使用场景。比如说百度搜索的第一页，往往是百度最为推荐的搜索结果，这是不是也是一个推荐系统？还有像抖音，我们每次随意刷到的视频，其实也就是抖音根据用户的兴趣爱好的内容。还有像Linked In 这样的交友平台，都会主动给你推荐感兴趣的好友，这也是这些平台根据我们的目的推荐的对象。提到这些，你或许会意识到，不太起眼的推荐系统，往往是这些互联网产品最重要的收入渠道。例如大家应该听说过，抖音包括tiktok，最为值钱的，其实就是他的推荐算法。

推荐系统在我们实际生活中，可以有很多的衍生场景。那这些推荐系统的核心到底是什么呢？其实推荐系统的真正核心，可以抽象为一个矩阵求解的数学问题。

推荐系统首先要有数据。推荐系统的数据要如何组织呢?

比如，电商网站向用户推荐商品，往往要基于用户以往的浏览记录或者评价记录。而这些历史记录就可以抽象为(userid,productId,score)这样的一个向量结构，userId表示一个用户，productId表示一个商品或者内容，score表示用户与商品之间的关系。这个score可以是一个任意的数字，比如在这里表示用户的浏览次数。也可以是一个0或1的值，表示用户与商品之间是否建立了关系，比如是否浏览过商品，是否购买过商品。而这样的一些数据往往是比较零散的，当我们需要将这些数据整体进行梳理，就会以userId为一列，productId为一行，整理成这样一个矩阵。

ProductId

UserId	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
u_1	1			7		2		
u_2			3		6		9	
u_3	2	5		7		10		4
u_4		3	7		9		2	12
u_5	5		2	8		2		11
u_6	12	4		6		9	10	

一个向量数据，就代表了矩阵中的一个点。在这个矩阵中，数据通常是比较稀疏的，称为**稀疏矩阵**。而推荐算法要做的，就是将这些矩阵中的空白点，以某一种方法进行部分填充或者全部填充。每填充一个点，就代表向用户推荐这个产品的一个推荐指数。而以后我们要做的这些"看了又看"这样的推荐功能，就只需要从这些填充的指数中寻找排名比较靠前的就可以了。

关于推荐系统的数据要如何梳理，后续还会有说明。所以，关于推荐系统，可以有各种各样，千变万化的应用形式和场景，但是本质其实就是二维矩阵的补全问题。

2、如何衡量一个推荐系统好不好?

现在有了数据，我们就可以把推荐系统这么一个抽象的理论问题，转成了一个具体的数学问题。但是这个问题有点奇怪，他并没有一个完整的解。你可以随便往里面填充数据。

我们之前说的，就按照产品的销售量给用户推荐，其实也可以映射到这个矩阵当中，只不过标注的方式就是按照产品的销量排序，将对应产品所在的那一列全部填充为产品的销量。甚至，我们在这个矩阵中随意的填上几个数字，也都可以说这就是实现了一个推荐系统。而互联网上像天猫，京东实现这些好的推荐系统，也可以认为是对这个矩阵填数字的过程。

ProductId

UserId	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8
u_1	1	100	200	7		2		
u_2		100	3	300	6		9	
u_3	2	5	200	7		10		4
u_4		3	7	300	9		2	12
u_5	5	100	2	8		2		11
u_6	12	4	200	6		9	10	

但是，从直觉上我们就会觉得，这样的推荐系统不太好，而天猫、京东的推荐系统更能抓住我们的兴趣。甚至像抖音、头条这类的应用，我们甚至会觉得他们推送的内容总是能抓住我们的胃口，我们自己都不太知道我想要看什么内容，但是抖音、头条他们懂，比我们自己更懂自己。**不同的推荐系统虽然没有明确的分数来表示他的好或者坏，但是，最终他们还是会体现出不同的好坏层次。**

那为什么会这样呢？只是归根于这种玄学总归是不靠谱的。从数学的角度来看这个问题，就是因为我们自己设计的推荐系统没有很好的利用已有的数据，没有从已有数据中"学习"到内在的规律。而好的推荐系统则是通过机器学习很好的挖掘出了已有数据之间的一些内在规律。这些规律可以体现为每个用户的兴趣爱好，每个产品的最佳受众等等很多规律，甚至是很多人类无法描述的规律。

比如最经典的啤酒和尿不湿要放在一起售卖的问题，这就是一个机器学习中的一个经典故事，你可以强行做一些解释，但是总是很难接触到本质。所以，对于推荐系统好坏的评价，就不能像我们之前学习的J2EE项目，简单的从实现效果，压测结果等方面来分辨出好坏。 对于推荐系统的衡量通常需要基于非常多的维度进行综合评价。大致可以分为以下几类：

1、基于常识的评判标准

从上面我们已经知道了，推荐系统其实只是一个数字游戏，但是我们的业务不可能是简单的数字游戏。有一些推荐的结果，我们是可以从业务上直接判断是好是坏的。其中坏的推荐的比例，就可以作为对推荐系统一个评判的标准。

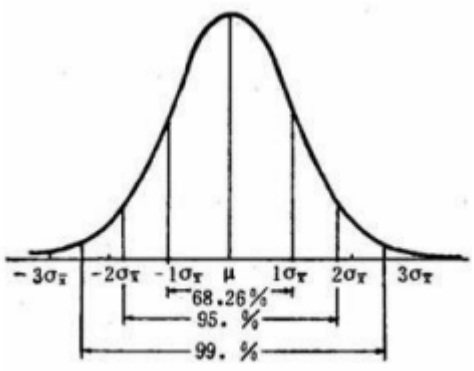
比如，对于天猫、淘宝这类电商网站，给用户推荐他已经购买过的商品，往往就不是一个比较好的结果。但是推荐已购买过的商品的周边商品，这个结果就会比较好。就好像用户购买过一个鼠标，但是再给用户推荐一次这个鼠标，用户再去购买的可能性就会非常低。但是推荐一个配套的鼠标贴，或者鼠标垫，用户去购买的可能性明显就会更高。而对于一个音乐类的内容推荐系统，如果一个用户是某一个歌星的粉丝，那再去给他推荐这个歌星其他的歌或者专辑，通常来说就意义不大了，因为这些内容用户自己通常会去主动搜索。再比如对于一个新闻类的内容推荐系统，如果给用户推荐的新闻或者文章包含了很多"过时"的内容，这包括很久前发布的内容和拿很久以前的内容重新发布的灌水帖，那显然这也是一个不好的推荐。

2、基于指标的评判标准

通常对于推荐系统的评价，还是会要回归到业务本身。所以对于推荐系统最为常见的评判标准，还是通过一定的业务指标来衡量。例如常见的PV，UV，用户留存率、转化率等等，通过比对上推荐系统之前和之后的指标数据，来衡量一个推荐系统是不是有效。或者拿推荐系统改进前和改进后的指标数据来比较，衡量推荐系统是不是越变越好了。这些普遍性的业务参数，不光对推荐系统管用，实际上，对所有的功能模块都可以用来进行衡量。比如对于类似于猫眼这样的电影购票系统，最直接的衡量标准就是推荐系统带来的流量和收入的增长。

另外，对于推荐系统，还有一个非常重要的指标就是推荐产品的覆盖率，也就是推荐出来的产品应该要越丰富越好。这是为什么呢？这就涉及到了电商的一个根本性的理论模型-"**长尾经济模型**"。商品的交易行为，通常都会遵循一个普遍性的2-8理论，即80%的利润出自于20%的商品。比如我们去超市购物，通常也都比较喜欢购买最热门的，品牌印象力大的商品。所以当我们以产品为X轴，产品带来的利润为Y轴，经过整理通常都能得到一个这样的正态分

布图



这是一个普遍的商业规律，大量的商业价值都集中在少数的几种商品。对于传统的购物方式，如商店、超市等，要提高自己的收入，就必须遵循这样的商业规律，将有限的货架留给最有价值的少数几种商品，而大量销售量比较少的商品，则只能放弃。这也造成了在产品、商铺、超市之间，热门商品的竞争更加激烈，传统商业模式一段时间内增长有限。

而基于互联网的现代电商，虽然跟传动商铺有很多形式上的差异，但是最为根本的差距，就在于电商上的货架成本是非常低的，并且基本上可以说是无穷无尽的。所以电商产品就有可能抓住这一部分处于长尾的，比主流市场还要大的小商品市场。其实我们生活中应该也有这样的感觉。早几年电商还在起步时，电商网站上的商品相对比较少。而随着电商的逐渐火爆，如今淘宝上的商品的丰富程度，已经发展到了一个令人无法想象的程度。例如早些年，DDD还只是一个默默无闻的软件理论，在他提出的很长一段时间内，基本无人问津，因此基本不可能在传统的书店购买到。但是放到电商网站上售卖，就基本不会有什么成本压力。随着这几年微服务的程序，DDD开始呈爆炸式的推广。而相关书籍带来的收益，就只有电商网站能够最先吃得到。 其实关于电商与实体经济的矛盾，一直争论不休，但是，其实电商与实体经济也可以是相辅相成的。比如，随着电商逐渐火爆，超市的发展速度实际上也是更快了，很明显可以感受到，超市的规模越来越大了，购买的商品和配套的服务也越来越多了。这就是电商的长尾效益带来的变化，也可以说，长尾经济效益就是电商的立身之本。

而基于电商的推荐系统，就承担了向用户推荐长尾产品的重任。这些长尾部分的小商品，天生注定不太可能得到自有的流量，也不太可能通过传统的广告等方式来推广。比如之前提到的鼠标贴，基本不会有人主动去搜索，也更不会有商家为这几毛钱的商品去打广告。所以隐藏在人们焦点之外的这些小商品，就需要通过推荐系统来重见天日。而对于头条、抖音这类内容平台，也正是因为推荐系统，众多默默无闻的自媒体，才有了发展的空间。

对应我们之前的商品矩阵，也就意味着，矩阵越大，矩阵中的空白点填得越多，推荐系统的算法也就越好。所以电商网站通常也会拿一些数据来模拟进行推荐，通过统计推荐算法的商品覆盖率，来评判推荐系统的好坏。这也是推荐系统比较有特色的一个指标。

3、基于机器学习的评判标准

通常推荐系统需要结合大量的业务数据，通过对历史数据的挖掘、分析，归纳出用户与产品之间的一些关系。这些关系通常过于隐晦，有些是能够进行解释的特征，比如用户的爱好、产品的受众特点等。但是往往还有很多隐藏的关系是无法用简单的常理来解释的。这些关系就要通过机器学习的算法来进行深入挖掘。最终通过这些分析，你是不是又会觉得这个推荐系统没那么简单了？

所以现在业界普遍的推荐系统都是基于机器学习算法来完成的。而每一个机器学习的算法，都会有他自己的评判指标和优化方式。这个我们会在后面的课程中逐步带大家深入了解。

总结

最后，大家可以看出，推荐系统没有绝对的好坏之分，对推荐系统的评测更是一个浮动的评判标准。但是一个普遍的规律是，好的推荐系统往往是从坏的推荐系统基础上改进演化而来的。这个过程，其实很像我们人类学习的过程。我们对很多事物的理解，不是简单的用懂或者不懂来衡量的。往往是基于自己的经验，开始某一项知识的入门学习，而随着我们的学习"经验"越来越丰富，我们对知识的把握也越来越准确深刻。而这，其实也就是机器学习的处理过程。通过对历史数据(经验)的学习，逐渐加深对数据整体的理解(规律)。 下一节，就会先带大家来了解一下机器学习的基础。

> 机器学习其实是门槛比较高的，需要有大量的高等数学以及统计学的基础知识。但是别害怕，机器学习很复杂，但是我们的入门教程会很简单，会带大家尽量绕过那些复杂的理论知识，从工程化的角度来接触一下机器学习。但是，入门之后还是需要大家自己继续深入学习才行。关于机器学习以及深度学习的内容，可以关注图灵官网的大数据系列视频，我们后续也会再像我们之前的课程一样，再准备更全面的学习课程。

三、推荐系统-机器学习基础

1、什么是机器学习?

机器学习其实是属于人工智能的一个研究范畴。说到人工智能，大家应该都很熟悉了。 像AlphaGo大战柯蓝，已经是一个耳熟能详的故事了。关于人工智能的起源，最早诞生于一个可以跟下简单的跳棋的小程序，但是当时人们也都没意识到这小程序有什么用，只不过是一个简单的消遣。只到1956年8月，在美国达特茅斯学院中，约翰·麦卡锡（John McCarthy，LISP语言创始人）、马文·闵斯基（Marvin Minsky，人工智能与认知学专家）、克劳德·香农（Claude Shannon，信息论的创始人）、艾伦·纽厄尔（Allen Newell，计算机科学家）、赫伯特·西蒙（Herbert Simon，诺贝尔经济学奖得主）等科学家聚在一起，讨论起一个不食人间烟火的主题：**用机器来模仿人类学习以及其他方面的智能**。这个会议开了两个月的时间，虽然大家没有达成普遍的共识，但是却为会议讨论的内容起了一个名字：人工智能。 这一年也被普遍认为是人工智能的元年。

其实从这个会议也能看出，两个多月的时间，这些大神级的人物一起最终也只讨论出了一个笼统的目标。具体要怎么去做，做成什么样子也完全没有结论。从现在来看，人工智能的实现方式大致分为了两个学派：

一个是**符号主义或者也称为逻辑主义**(Symbolism)，他强调人对事物的认知是有一定的推理过程的。所以只要通过计算机，把逻辑推理的过程模拟出来，就能实现人工智能。像纽厄尔、西蒙包括后来的尼尔逊，都是这个学派的代表人物。

另一个是**连接主义**(Connectionism)，他强调的是从仿生学的角度，来模拟生物体的结构，其中重点就是人脑的结构。这个学派认为只要用计算机硬件模拟出人脑神经网络的结构，最终计算机系统就能够模拟出人的认知以及决策的过程。

另外还有一个学派，**行为主义**(Actionism)，他强调的是从行为角度来用计算机模拟生物，主要是昆虫，的行为。他们早期的研究工作重点是模拟人在控制过程中的智能行为和作用，比如自寻优、自适应、自镇定、自组织、自学习等控制论系统的研究，并进行"控制论动物"的研究。但是这个学派是从20世纪末期才出现的新面孔，到目前为止，无论是影响力，还是研究成果，都与前面两个学派差距还比较大。

到20世纪80年代，一些搞数学与统计学的研究人员，实现了一些基于统计学的机器学习算法。强调从已有数据集中基于统计学的概率来学习数据中的经验，最典型的应用就是筛选垃圾邮件。所以机器学习也可以认为是符号主义的一种实现方式。

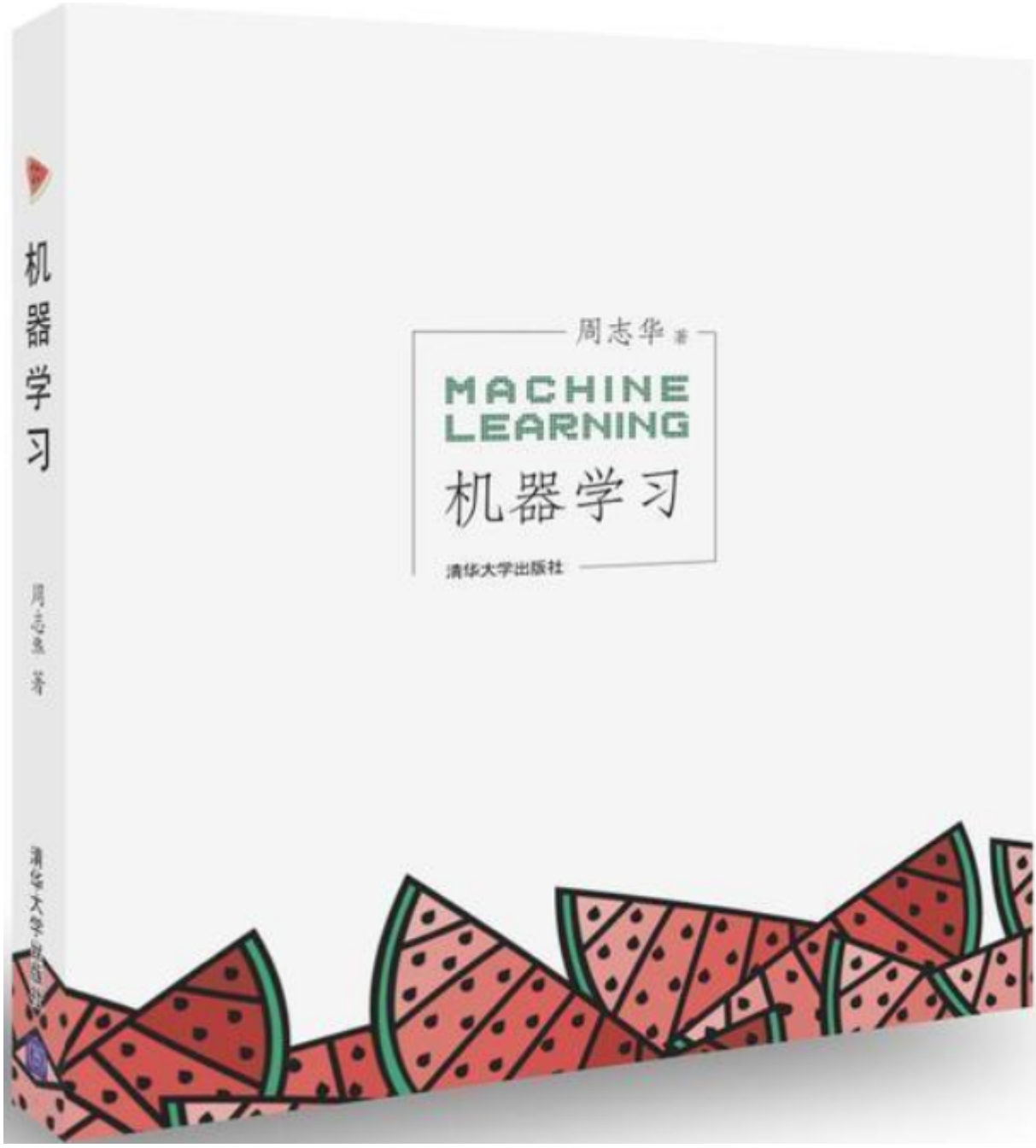
而直到2010年往后，在一些图像识别领域的竞赛中，出现了一些基于神经网络的深度学习算法，在比赛中脱颖而出。于是深度学习就开始在各个领域迅速火爆起来，最典型的应用就是图像识别，自然语言处理。关于深度学习，他是以模拟人脑神经元的结构来构建的，可以认为是连接主义的一种实现。但是其实深度学习的实现方式很多地方都借鉴了机器学习，所以深度学习更多时候认为是机器学习的一种方法。

机器学习的应用领域： 机器学习的应用领域是非常多的，大体上，可以分为三个主要的方向

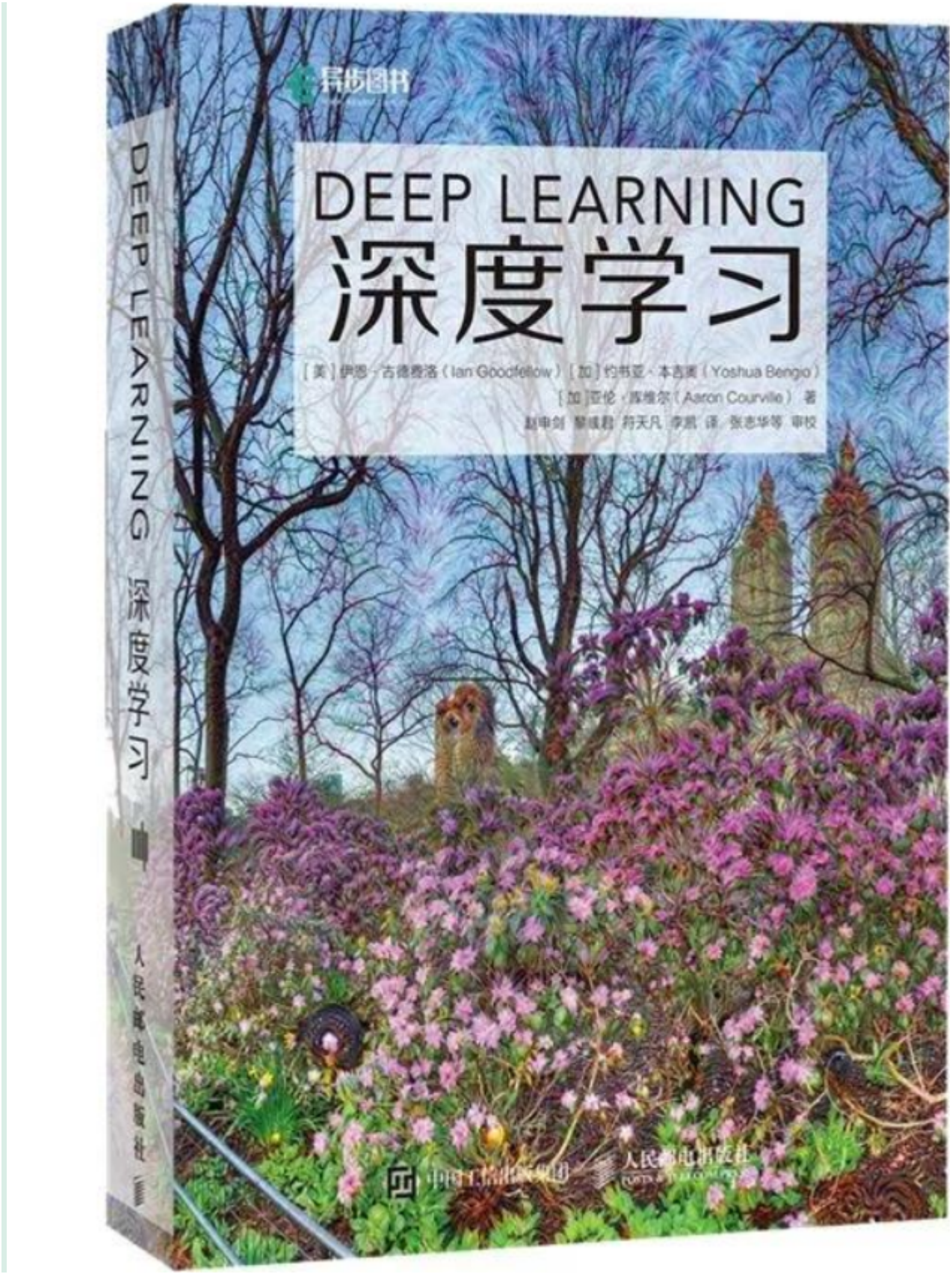
- 传统预测：主要用在数据挖掘，预测领域。典型的应用场景： 店铺销量预测、房价预测、垃圾邮件安全监测等。包括我们这个课程的推荐系统，其实大体上也可以分到这一类。当然，这也并不是绝对的。基于神经网络的推荐系统也是有很多落地实现的
- 图像识别：典型应用场景：自动驾驶、人脸识别、涉黄图片视频过滤等
- 自然语言处理：典型应用场景：文本分类、聊天机器人、智能客服、文本翻译等。其实我们能感觉到，早期的百度中英文翻译就非常难懂，语法非常混乱。但是现在百度中英文翻译就相当人性化了，语法也非常自然。这其中就有深度学习参与其中。

所以我们这节课的重点，就是带大家掌握一些基础的机器学习的算法和技巧，能够从某些简单的业务场景切入去解决一些实际的问题。

> 关于机器学习，有一本非常经典的入门资料，就是周志华的《机器学习》，俗称为西瓜书。因封面有很多西瓜，并且全篇很多问题都从西瓜谈起而得名。

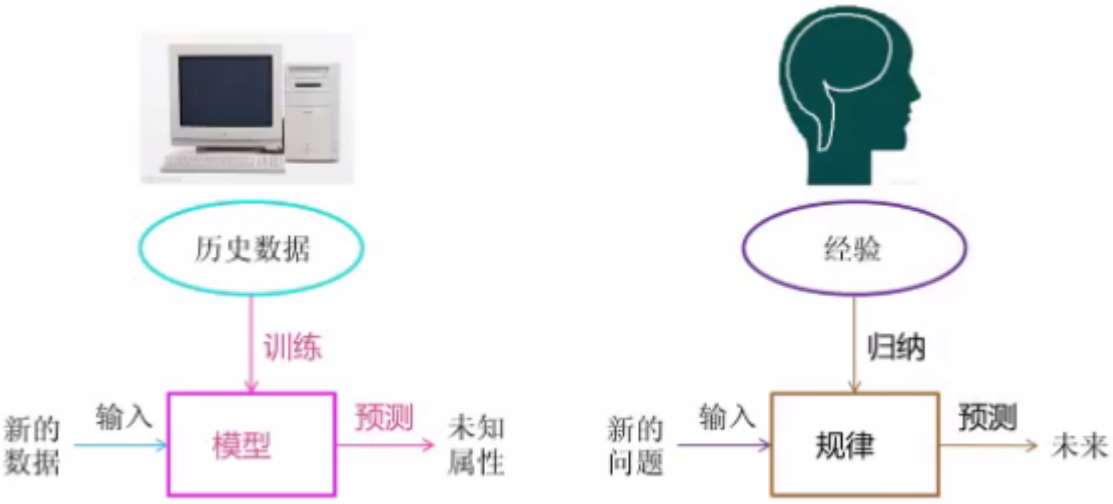


而关于深度学习，也有一本非常经典的资料，名字就叫做《Deep leaning》，深度学习。俗称为花书，因封面有非常多的花而得名。



2、机器学习数据形式

机器学习有三个关键词： 数据，模型，预测。**机器学习强调从历史数据中自动学习，对数据之间的规律进行归纳，形成模型，然后用模型来对实际问题进行预测**。这个过程跟人类理解一个事物的过程是很类似的。回想一下，人类去分辨猫和狗、或者预测房价未来的走势，其实也是这样一个过程。人需要从大量的日常生活经验中归纳出一系列的规律，然后在面临具体问题时，就可以从众多规律中找到最优化的规律，来解决日常问题。



机器学习的数据集构成：机器学习能够处理的数据由特征值+目标值组成。

	房子面积	房子位置	房子楼层	房子朝向	目标值
数据1	80	9	3	0	80
数据2	100	9	5	1	120
数据3	80	10	3	0	100

比如像这样一个房价数据，每一行数据称之为一个样本。多个样本就构成了一个数据集。而在每个样本当中，前面部分房子的各个属性构成了特征值，代表样本的各个数据特征。而最后的目标值相当于是样本的结果。而机器学习的过程就是要从已有的房价数据中学习到房价之间的规律，然后以后再来一个房子，我们就可以根据房子的这些属性，预测他的房价是多少。而在数据集中，特征值是必不可少的，一般就是原始数据。而目标值有可能需要通过对数据进行处理来获得，但是有些数据集是可以没有目标值的。

> 推荐系统的数据要如何梳理?

3、机器学习的分类

机器学习涉及到非常多的数学算法。对这些数学算法，通常会根据目标值的类型进行简单分类。

- 分类算法： 这一类问题的目标值是有限的几个离散值。例如我们对动物进行分类。常用的算法有： k近邻算法、贝叶斯算法、决策树与随机森林、逻辑回归等。
- 回归算法： 这一类问题的目标值是一组连续值。例如对房价的预测。常用的算法有：线性回归、岭回归等。
- 无监督学习：这一类问题没有目标值。也就是说，没有一个固定的目标去监督机器学习的过程。例如我们常说的人以类聚，物以群分。我们通常会需要将所有客户区分成一个个具有相似特征的客户群，但是我们也不知道要把客户分成哪些群比较合适。这个时候，就可以用无监督学习，让机器学习去找出最具有区分度的划分方式。常用的算法有 k-Means分类算法。

与无监督学习对应的，分类算法和回归算法都是有目标值，也就是有具体目标的机器学习算法，他们就统称为监督学习。

> 推荐系统是属于哪种算法?

4、怎么获取数据集?

从上面已经知道，在机器学习中，数据集是根本。而在机器学习的工作过程中，其实很大一部分的资源就是消耗在数据集的处理过程中。一方面，数据集要尽量全面。越庞大，越全面的数据集，机器学习出来的效果越有说服力。例如我们要分析不同年龄段用户在电商网站上的兴趣爱好，从每个年龄段选一个用户出来，分析他的行为，也能形成一个模型，但是不用我说，你们也会觉得，这样的分析没有说服力。而反过来，如果我们拿到的是整个淘宝所有的用户行为数据，甚至是全球所有电商网站的用户行为数据，那这个说服力就不同了。另一方面，之前提到过，对于监督学习的目标值，在很多情况下都是需要对数据进行处理的，这需要很多的工作量。例如我们要分析用户的行为，需要将

所有用户按年龄区分为年轻人、中年人、老年人，这就是一个数据打标的过程。而像对猫、狗图片集的正确分类，还有网络涉黄图片、视频的鉴别，工作量就更大了。在大数据量下，如何快速对数据进行正确的区分，也是大数据领域非常重要的一个工作方向。

所以，大家如果对于大数据工作不了解的话，也不需要有惧怕的心理。在机器学习方向，算法最复杂，但是也是最难创新的，往往在平均一百个从业者当中，也就寥寥一个不到的大神级科学家是负责设计核心算法的。剩下来大概四个左右的算法工程师是负责各种编程语言的算法实现的。另外，绝大部分的从业者，包括市面上招聘的绝大部分算法工程师，也就是围绕这些核心算法，匹配业务场景，然后对数据进行收集、清洗、标注等这些工作的。搞不好，你填写一个图形验证码，也就是一个算法工程师了。至于计算过程，也就是挑几个常用的算法，调整参数，计算出来再比较一下效果。至于为什么这个算法效果就比另外的算法好？这个问题其实很难，但是你不知道，大部分的算法工程师也同样不知道。

在实际业务中，这些有用的数据集成本巨大，甚至可能包含了很多核心的商业机密。那在学习阶段，我们要怎么去获得有价值的数据集呢？主要还是通过直接使用别人维护好的数据集。

行业内有很多科研人员都维护了很多质量非常高的开源数据集。例如python的sklearn框架就集成了一部分常用的数据集。

> 参见pycharm中的Demo: sklearn_datasets.py，加载sklearn本地的Iris鸢尾花数据集，还有也加载了Boston波斯顿房价数据集。这两个数据集是机器学习领域最为经典的数据集。Iris就是分类问题数据集， Boston则是回归问题数据集。
>
> 而在java领域，可以使用Spark的mllib包来做机器学习。也可以将这些csv文件读到spark当中。参见SparkDemo中的LoadDataDemo。

UCI： http://archive.ics.uci.edu/ml/ 这个网站上维护了很多经典的数据集。

kaggle： https://www.kaggle.com/ 一个综合性的机器学习竞赛平台。上面会开放很多数据集，开展很多机器学习的竞赛。有很多都是一些公司自己处理不了的实际数据，数据集的质量通常都是比较高的。同时也有很多别人分享的基础教程以及算法分享，也都是非常不错的学习资料。

总结

这一章节介绍了一下机器学习的概貌，大家有兴趣可以去kaggle上简单看一看。体会一下机器学习解决哪些问题以及他解决问题的方式。上面不光有竞赛，还有很多别人发布的计算过程以及专业的教程。接下来的两个章节，将给大家介绍一些典型的机器学习工具和算法，让大家能够在kaggle上入门，至少看懂这些大神们在干什么事情。