好了，下边我们来看一下第六章，第六章呢？我们讲的时候呢？没有完全呢，按照教材来讲，所以说大家复习的时候呢？按照。课件儿的顺序内容来进行复习就OK了，那么第六章也重要。

大家不要觉得第六章我们学的时间比较晚，它就不重要。四五六这三张差不多，在考试所占的分值比重差不多的，大概都是十几20分儿吧，有的章可能20分儿还得多反正。都挺重要的，三四五六这这四张吧，大家要特别的注意啊，好。

那我们来看一下第六章机器学习。那么，前半部分呢？有一个重点就是学习系统的构成，那么一个学习系统由三个核心部分构成的这幅图，当时我在讲这部分内容的时候就给大家说了。大家要理解这幅图能够把它画出来，学习系统有三部分，三个基本的核心要素构成环境知识库和执行，还有呢。

这个箭头怎么走？大家要掌握环境向学习系统提供信息。学习部分利用这些信息修改知识库，增加系统的执行部分，执行部分根据知识库完成内容，把获得的信息反馈给学习部分，这就。三者共完整构成了一个学习系统，那么在具体的应用中，环境知识库和执行他们共同构成了学习系统的全部的要素。

好了，那么这儿呢？是学习的基本模型，大家要掌握下一个重点内容，就是机器学习的分类了，机器学习的分类，我们说了可以按照不同的方式来进行不同的分。但是呢，最常见的分法就是按照学习的方式来分，我们呢。

分成三大类监督学习，无监督学习和强化学习，对于这三大类，大家要深刻理解。不是说让你死记硬背的，你要深刻的理解，比如说最核心的这张表，那么其中监督学习它的特点是什么它？它需要有标签的数据，大家想想我们学的所有的监督学习算法。

我们学的分类算法，我们给大家讲的决策树，我们给大家讲的kn，它是。事先都要人为的，把我这个标签给打好，所以说呢，监督学习它呢，是需要有标签的数据，它直接来进行反馈。

它最终学习的目标就是分类算法和。回归预测算法，这就是监督学习，无监督学习，它是不需要有标签的数据，我们直接把数据喂给机器，让机器自主的去找寻这个数据内在的规律，因此它是无反馈的。那么，它主要运用在聚类和降维中。

那最后一个强化学习re for cement learning，它是一个激励的决策过程。我把计算机放在一个环境中。做对了给激励，做错了给惩罚，就是根据给他的不同的政府反馈，让机器来掌握这个相应的规则，那么这就是强化学习。三大学习他们的概念，他们的比较。

大家要掌握这是关于机器学习的分类。大家要掌握的内容。下边呢，就是具体的学习算法了，具体的学习算法呢，首先第一个就是我们给大家讲的这个视力学习这部分呢，一般了解就可以了，因为这部分。我们课本上虽然提了，但是我们课本上没有给大家具体的来讲例子。

我们给大家讲的这个例子呢，就是病态细胞的识别案例呢，主要呢是用在my中。其他的实例学习呢例子和这大差不差，但是呢，它还是有一点些许不同的，所以说呢，对于这部分的实例学习呢，大家能够大致的把例子给看明白就OK了。那么。

我们应该把焦距和重点呢放在后边的分类和聚类的算法上，尤其是分类，尤其是后边的分类。那么，首先我们来看一下。下呃，这个距离和相似度大家也一般的了解就行了，这也不是不重要，这非常重要，但是这属于我们给大家额外补充的内容。

就是给大家讲了四种距离和两种相似度的，具体的计算是属于额外补充的内容，所以说大家额外补充吧，考重要，但是我们考试的试卷不涉及。所以说呢，这个大家呢，稍微看看就可以，下边的重点呢，就是我们给大家讲的两种分类算法了。

尤其是。前一种就是决策树ID 3算法，大家务必要深入理解ID 3算法的计算过程。就是要掌握到什么程度呢？比如说给你一个系统，给你一个示例，让你把他的决策树画出来，不一定让你画完整的决策树可能。这个时间你不一定有那么长时间，但是呢。

让你计算某一个节点，比如说根节点，让你算那么它的这个BC I的值，然后呢MC的值，大家要会算。这一部分的关键点呢？因为这个我们才讲过啊，大家应该记忆犹新，关键点呢，就是怎么样利用信息商来通过计算得出它的最优决策数。

那么，信息商的计算呢？虽然是补充的内容，但是需要大家掌握的啊，因为你要不会信息商的计算，那你整个ID 3算法，你就你就不会了，所以说呢，大家呢，要掌握怎么样？

来计算这里边的核心值怎么样？来计算MC怎么样？来计算BC a注意书上这部分的内容呢？和我们讲的用的符号是不一样的。书上呢，采用的是传统的，就是代码式的符号，我们这里边呢，采用的是大部分教材里边呢，就是。

其他的大部分的教材里边采用的MC和BC I方法是一样，但是书上的符号表示是不一样的，那你用哪种都可以，比如说大家如果提前预习了教材，你用书上讲的那个代码的那种符号表示。它表示，我们这里边的MC和BC I也可以也可以呢，用我们课件里边直接讲的这个MC和BC I表示呢，也可以那这部分呢，大家要求重点掌握的。

好了，那么下一个呢？就是kn kn呢？大家呢？不需要让你掌握多么复杂的应用，但是呢，要掌握最基础的内容，比如说给你这幅图。然后你要知道kn是什么意思，距离最近的k个邻居那么。

然后看在这k个邻居中哪种类类型多，哪种类别的数目比较多。我们就选哪种那么这10 kn大家要掌握哪种都能把这个图看明白就OK了，至于像我们书课上讲的那个复杂的例子，比如说。怎么样来完成兴趣推荐呢？这个不做很深的要求，至于聚类，大家要掌握的第一个聚类是一种典型的无监督算法，它是无监督学习的代表。

这是第一个。第二个呢，就是大家能够把我们课上讲的这个例子，大概的看明白就OK了，更复杂的例子，我们不做要求这个呢，是聚类。对于整个第五，第六章呢，大家就掌握到这么一部分就可以了。

其中重点呢还是ID 3，大家应该把重点和焦距放在这上面。好了，这就是我们今天的第一个任务，我们差不多花了一节课，多吧把整个课程呢，给大家做一个回顾，再次强调。期末考试要重视考试不难，但是如果你不认真复习的话。

那你还真不一定能能能考多多高的分，因为。它每一章的内容都有涉及，我们课上给大家讲的所有的重点内容都有涉及，所以说大家在复习的时候呢，你不要复习过于偏题和怪题。比如说如果大家看我们的教材，课后的练习题有一些让你画那个语义网络的题目，或者有一些让你写位词的题目，那很难那种很偏很怪的。

大家可以先不看。大家呢，应该把重点放在我们课上，反复强调的，我们课上反复练习的，我们课后反复让你做作业的那些题目上，这才是我们整个课程的。重点和精髓再次强调，我们这门课是平台课，它最大的特殊点是期末考试所占的比重特别高。

是二八开。和我们其他的课程都不一样，在整个大学里边可能计算机专业可能也就这一门课是二八开吧，其他的课可能都不是。可能就这一门课是二八开，所以大家也一定要抓注意这个特殊点，它的期末的分值特别的关键。他不像别的课，平时分你要期末，万一考不好的话。

平时老师能救大家就救大就救大家，但是这门课老师的能力也有限，你那个期末的试卷才是关键。所以说大家一定要重视那个试卷，距离考试还有一周多的时间，时间足够，大家要认真复习，就怕什么你平时因为这样或那样的原因。课没有全来，可能关键的一些算法课没来。

然后平时你最最后考试的时候，你觉得很容易，你也没有认真复习，或者是某一个点疏漏了。那一个大题，他就十分儿这有有因为咱们看嘛，那基本上最最小的小题也是五分儿那大题一个大题就十分儿，那如果你要不会的话，那这十分儿一下子只要得扣个大半。打蛋吧。

或者是就没了，所以说大家在复习的时候一定要抓住重点好了，这是我们今天这次课的第一个任务啊。这个回头呢，我会把答疑的时间我个人还是建议线上答疑最方便，因为呢QQ我每不是时时都看，不像微信直直接就推送了。但是呢，我至少每天晚上之前会看一遍，所以说你在QQ上问我。

如果我当时没有回答的话，因为我就说明我当时没看，但是我看到了会在第一时间回复你，所以说我个人觉得呢QQ答疑更方便。但是呢，大家要想线下答的话，那我们就定在下周四的上午好了，那么这是我们今天的第一个任务，下边还有时间。我们把我们的第二个最后的两小节的内容给大家来介绍一下。

好，我们来看一下我们整个的机器学习，还有两章补，还有两节补充的内容，这两节其实挺有意思哈，挺有意思，而且是比较新的内容，但是呢。我们之前怕课时没够，所以说没有之前一开始就给大家讲。

现在还有差不多大半节课的时间，我们把剩下的内容呢给大家呢来讲一下。那么，首先第一个补充的内容呢，就是机器学习和大数据。就常见的大数据中用到了哪些机器学习的算法大数据呢？这两年非常的流行，因为本身嘛现在。有专门的大数据专业，大数据专业嘛。

现在也是计科的四大专业之一嘛，那大数据其实它里边学的课程呢？大部分的课程和计科和人工智能课是一样的，它的基础的课，核心的课都一样的，但是呢，他们在大三和大四阶段呢，像一些专业课，包括像一些实验课。那么。

和我们的传统的计算机课呢，有些区别，那其实就是我们这一节里边要讲的内容好，那我们来看一下这部分。要给大家讲的。等一下啊，这部分呢，要给大家讲的方面，那么首先呢，我们来看一下这里边几个概念。

他们是怎么样一步一步发展到现在的就是。统计学机器学习数据挖掘以及深度学习和大数据这些概念呢，他们其实呢，本身呢，就是相相关的概念。我们来看一下，这里面有张表，那么人工智能，包括人工智能有一个重要的分支，叫统计学习。

它本身呢，就是根据统。统计学里边的相关的概念和特征，然后呢，由统计学提供相应的特征和模型，然后形成相应的机器学习。其实大数据呢？在大数据概念发展之前呢？它其实早期的概念呢？叫数据挖掘。

我们就可以把数据挖掘呢？看成是当前大数据的一个前。前身所谓的数据挖掘呢，就是给你海量的数据，让你从海量的数据中来明确来分析这些数据之间的。内在的紧密的联系，形成相应的算法，所以说现在大数据的一些很多的经典的算法就来自于数据挖掘的经典的算法，比如说。像基于关联规则的大数据的算法。

像employee之类的，它其实最早呢就来自于数据挖掘，然后呢，在统计学和机器学习的基础上，形成了数据挖掘的概念。然后机器学习呢，它的其中的一个重要的分支就是deep learning，deep learning就是辛顿提出的深度神经网络，它的特点呢，就是在传统神经网络的基础上。

那么，构建参数更多的层数，最深的神经网络来实现对海量的数据和超越预期的分析的能力，这儿呢，是一张表。那么，他们的实现基础呢？就是有传统的概念，也有一些新的概念，比如说数据库。

数据仓库，这都是很传统的概念。再比如说分布式存储，比如说并行计算。并行计算呢，其实也是现代大数据的一个分支。像现在大数据专业，他们有一门课，专门就是来讲基于gpu的扩大的计算，因为很。

很多的时候，我们做这种大数据分析处理的时候呢，它的算法采用CPU是不行的，我们得基于这种gpu通，基于这种扩大来来完成，然后呢gpu加速等等这些概念，大家都理解。但是有一个概念呢，可能大家之前没有给大家没有听说过，给大家解释一下。

叫做流式计算，流式计算呢，这个概念呢，它主要呢是针对。这个批量计算，这两者呢，来相区别的那么什么是批量计算呢？就是咱们传统的对数据的处理，比如说。最核心的数据库对数据的处理就是采典型的批量计算。

那最典型的批量计算的方法就是我们先收集数据。然后呢，在这个基础上给他们形成二维的关系数据库，二维表放到数据库中，然后需要查询的时候，我们通过最后语句对它来进行查询。这就是这种传统的对数据来进行收集，存储并且处理的过程，那么我们把这种过程呢，叫做批量计算。

就是它字面上的意思，把数据存储起来，然后呢，按照批量来进行计算，那么这儿呢是批量计算的示意图，它所有的基于数据的批量计算呢？都是以二维表为核心的，我们的关系数据库说白了就是来学习怎么样，针对这些二维表来做一些日常的操作，那什么是流式计算呢？

流式计算呢，注意了和他的名字是一样的，它呢是对数据流来进行实时的计算，那么它不是更快的p计算。它是一种完全不同的思路，注意了和批量计算，那样慢慢积累不同流式计算呢，加大量的数据拼摊到每个时间点上。连续的进行小批量的传输数据，持续流动计算后就丢弃。

比如说我现在有大量的数据，那我和我批量计算一次，对这些数据来进行处理不一样。我把这些大的数据呢，给它分摊到不同的时间点，比如说我这一个小时，然后呢，第一分钟处理哪些，第二分钟处理哪些，把这些数据呢。

给它平摊。然后呢，就像数据流一样，连续的批量的进行计算，那计算以后呢，就进行丢弃注意了，批量计算呢，它的本质是二维表。它是对一张整个表格来实施的各种计算流式计算呢，注意正好相反。

它必须呢，先定义好计算逻辑。提交到流式计算系统中，这个计算逻辑一旦运行，它就没法更改。那么，这儿呢，是大数据中的流式计算的相应的过程。那么，在计算结果上。

批量计算呢？是对全部数据进行整体计算后，然后传输结果。而流式计算呢？不一样，它每次呢？小批量的计算。然后把结果直接投递到在线系统，做到实时化的展现，这是大数据处理的一个重要的分支流式计算。

给大家呢，简单的讲一下。然后下边呢，我们来看一下大数据和AI的相关的概念，所谓的大数据分析呢，它的本质就是构建更。各种数学模型从数据中来学习特征和规律，从中收获有用的知识，那么对于特征而言，它决定着我们数据对象中。

蕴含知识的关键属性，而规律呢，就是我们从特征的表达式中找出的一种相关的模式以及参数而学习。那我们所说的机器学习就是从样本数据集中，通过计算，通过统计得到的有规律的过程，这儿呢，是一个大数据分析建模的一个过程。那大数据同学专业的同学，他们第一年上他们的基础课的时候。

一开始见到的就是这张表，他的其实你看一下我们整个大数据分析的过程呢和我们。机器学习的很多的算法的过程是一模一样的，可以说大数据的核心的专业课和人工智能的很多专业课至少有七八成都是重复的那。那它基本上都是从原始数据出发，把数据呢分成两类训练数据和测试数据，那么训练数据呢，大家看，如果是采用的监督学习的话，那么那我们需要事先给它打标签。

那么，这里边都是典型的一些监督学习的算法，像贝叶斯决策树s vm kn等等，然后呢，如果是无监督学习的话呢，那么我们。从这个数据中呢，来提取相应的特征，找出相应的规律，然后呢。

来完成这个学习的目标，然后呢，下边我们在测试阶段呢，通过测试数据来。来测试我整个过程，整个的过程，但是这中间呢，它要处理的很多的一些数学问题，比如说最小二乘。梯度下降最大自然等等等。

这些具体的算法，这个我们课上就不给大家讲了，大家在。学真正的学机器学习那本书的时候，里边呢会有大量的介绍，那么下边呢，我们来看一下大数据中的机器学习，大数据中的机器学习呢，它的定义和我们课上的机器学习的定义呢是一样的。它同样也分成三类或四类。

那么监督学习无监督学习，强化学习，但是有的书上呢，还把这个在这个无监督学习和监督学习之间呢，再加一类叫半监督学习。它的本质呢？这种四分法呢？和三分法是一样的，我们的课本和教材中呢？采用的是三分法。

但是有些课本和教材呢？采用的是四分法，这个本质上是一样的，没有什么太大的差异。那么，监督学习呢？大家学过了它主要的特点是使用有标签的数据来进行学习，它的典型的场景就是分类和回归，而无监督学习呢？就是使用无标签的数据来进行学习。

把数据喂给机器，由机器自主的寻找数据的规律，那么典型的场景呢？是聚类半监督学习呢？注意了，它是监督学习和无监督学习的综合体。它使用的数据呢，一部分是有标签的，一部分是没有标签的，那么没有标签的数据量要远大于有标签的数据量。

那么这种学习的典型的应用场景呢，就是对。海量的数据来进行分类，而强化学习呢？这是一种强调的反馈。想想教机器来打游戏，那么典型的应用场景呢？是策略推理。这就是我们的讲的在大数据中的机器学习，以及他们典型的分类。

那下边具体的案例呢？咱们其实前边都已经给大家讲的差不多了。比如说前面的监督学习和无监督学习，我们都讲过了，那我们来看一个之前没有讲过的半监督学习吧，半监督学习呢，比如说我们举个例子。假设我现在有一个图库，然后呢？让你在这个图库中通过学习来自主的判别哪些图片儿是日食的图片儿。

哪些图片儿是？不是日食的图片，那么大家想想，对于这样的一个任务，监督学习怎么学那没有什么办法，我只能把一大堆图片中训练数据喂给机器。然后呢？日食的我们给他做了一个标注非日食的，我们给他做了一个标注，让机器通过算法自主的找出他们的规律。

这是监督学习，无监督学习呢，更简单直接把数据喂给他。然后让机器自主的来进行区分，当然了，它的准确率和效率是至少从现在来看还是远不如监督学习的那么半，监督学习是怎么做的呢？注意了。它相当于是监督学习和无监督学习的结合体。首先。

那么这里边的数据呢，有些是有标识的，有些是没标识的。什么意思？就是我这个数据呢？有少部分我已经事先给它标识好了，是带有日食的，有大部分绝大部分呢，是没有标识的，那么我们采用半监督学习呢。

就是这样。首先第一步就是对于那些有标识的数据，我们采用监督学习的做法，我们呢？给他。形成训练分类器。当然，你用不同的算法形成的分类器的原理是不一样的，但是本质上都是形成这个日食图片的分类器，这是第一步。

第二步，对于没有标识的数据，给它进行分类，然后呢，我们通过特定的算法，按照信任度从大到小来进行排列，我们把那种信任高的。图片排在前边，把那些信任度低的图片排在后边，那么下边步骤三将信任度最高的图片自动的加入标识项。

比如说有些图片。我明显的看它不可能是日食，比如说这是一个汽车的图片，那么这另外也是另外一个图片，它压根离这个日食差了十万八千里，那么我们就给它自动的加入标识项。那么告诉他，这不是日食，然后第四步重新训练分类器净重复步骤二和步骤四，直到把所有的这个数据全部都给尝试完毕。

这就是半监督学习。半年度学习呢，它兼有监督学习和无监督学习的特点，它是这两者的综合体，综合体好，那下边呢，我们来看一下数据挖掘的概念。数据挖掘这个概念呢，要远早于大数据，它其实就是大数据概念的前身。

大数据这个词呢，也就是最近这十几20年兴起的。但是如果你把时间跨度跨。跨到20年前，或者是更早的话，那么那个时候呢，人们研究数据库，或者是研究这种数据处理呢，基本上呢，都焦距于数据挖掘。

数据挖掘这个词呢，最早是在一九八九年的时候呢，提出的它的根源呢，可以追溯到经典的统计学，人工智能，机器学习这三个学科。然后呢？关于数据库的发展和互联网的这个广泛应用呢？两次推动了数据挖掘的发展，那么这儿呢？

是整个。大数据和数据挖掘的它的一个发展的一个历史的历程图，从上个世纪60年代开始，有了数据收集和数据库的概念。然后上个世纪70年代到80年代，那么有了关系数据库，关系模型的概念，这个大家学数据库模型rr dbms，这个大家应该非常非常的熟。然后到了90年代呢。

有了数据挖掘和数据仓库的概念，然后呢，往后衍生出了多媒体数据库和网络数据库的概念，然后再往后呢，那么有了流数据流的概念，有了这个XML技术。再往后，那么就是形成了我们现在的大数据的概念，这就是整个呢？这个数据学科的这样的一个发展和历程。

好，那下边呢？我们来看一下主流的和数据挖掘和大数据相关的算法有哪些呢？这些算法呢？其实就。就是我们所说的机器学习算法，只不过呢，我们把它呢用在了大数据领域呢，它呢可以分成这么六大类六大类，就是呢。

五大类针对六种不同的知识。分成五大类，哪六种不同的知识呢？就是广义型的数据，广义型知识，特征型知识，差异型知识，关联型知识预测和偏离型知识。然后呢，衍生出了和大数据或者说和数据挖掘相关的五大类的算法。

分别是分类。聚类这两者，咱们前面已经学过了，和机器学习的分类和聚类的概念是一样的，还有就是传统的关联规则，时序预测和异常检测，那我们分别来看一下。分类和聚类呢，不用说了，分类是一种典型的监督学习算法。

它的特点是按照分析对象的属性特征，建立不同的组类来描述事物。那么分类呢？它主要处理的是什么呢？主要处理这两类，大家想想咱们前边介绍的，或者是遇见的大量的分类算法。决策树svmknn，它都是针对广义性知识和差异性知识来进行分类的，大家想是不是什么是广义性知识？

反映同类事物共同特征的知识什么是差异性知识，反映不同事物属性差别的知识，那分类不就是这样吗把？这种特定的类别，从广义的这些实体中给它分辨出来，给它分别出来，那这不就是广义性知识和差异性知识吗？那同样我们来看一下聚类。聚类呢，是这样的。

我们把数据喂给机器，让机器自主的来完成类别的划分，那么它主要处理的是广义性知识和特征性知识。因为机器要自主的找寻我这个数据内部的联系，那么它当然反映的是特征特异性知识，然后关联规则，关联规则呢，是大数据的传统的算法。这个算法这一类算法早在上个世纪的七八十年代的时候就已经出现，并很流行了什么是关联规则。

就是它字面上的意思啊，我们来寻找这些数据和数据之间的联系。那么，通过关联某种事物发生时，其他事物会发生的这样的一种联系，这种是关联规则，时序预测是。持续预测呢，1度也非常的流行，那么不光是在我们这个领域。

在工科领域也非常的流行，什么叫持续预测就是把握分析对象发展的规律。对未来的趋势做出相应的预测，比如说我在以前看过很多的一些非计算机专业，他们的硕士，博士答辩，那有些同学他就用到这个时区预测的算法，比如说。预测我这个油田那根据我往年的这些油田的产量规律来预测未来几年的油田的产量和规律，这就是时序预测。

还有一个呢，就是异常检测。异常检测呢，它的特点呢，是对分析对象的少数极端的特征，然后呢，揭示内在的原因，这个呢，就是我们所说的异常检测，这些呢。

都是我们常见的大数据。和数据挖掘的算法，那前面的分类和聚类呢？我们都说过了，我们快速的来看一下，大家就是课上没讲的算法。第一个呢，就是关联规则。所谓的关联规则呢，我们不给大家详细的说了。

因为马上下课了。给大家呢，大概的说一下就可以，什么叫关联规则呢？就是发现有意义的事物之间同时存在的规规律。就是有的时候压根是两个不相干的数据，两个不相干的事件，但是他们内部呢，却彼此存在着联系。关联规则算法是一种典型的无监督算法。

注意啊，它是不需要实现做标签的，我直接把数据喂给机器，机器自主的来寻找这些数据内部的联系。那么，典型的关联规则算法有APP算法，这个算法历史非常悠久了，三四十年了，三四十年了，但是在曾经的历史上非常有用。

还有一个呢，就是后来出现的fp close算法。这就是关联规则。关于关联规则的算法呢，有很多经典的案例，可能很多大家听说过，在营销学领域中有一个非常著名的案例，叫啤酒和尿布的案例。它其实背后用到的就是关联规则的算法，本质上来说呢。

就是德国的一个超市，他把这个超市的日常的商品呢，给它放到数据仓库系统中，对它来进行。分析发现呢，有一些商品，它压根是风马牛不相及的两类完全不同的商品，但是人们在买的时候呢，经常放在一起来买，其中典型的就是啤酒和尿布。

一个是喝的啤酒，一个是婴儿用的尿布，两个商品本来不搭界儿，但是人们买的时候经常把这两个放在一起来卖，后来呢，那个超市呢，就用了这个系统之后发现了这个规律。把这两个商品啤酒和尿布放在一起捆绑销售，那么获得了很大的成功，这在营销学上是一个著名的叫啤酒和尿布的这个案例。

那么，为什么会出现这种案例呢？为什么会出现这种情况呢？后来呢？超市做过调查哦，发现当时正好处于世界杯的这个这个期间。德国男人一般喜欢一边喝啤酒一边看球，然后在买啤酒的时候，如果家里有孩子，妻子就会说你顺便把孩子尿布给买了吧。

然后就放在一起了，当然了。这个理由呢，很牵强，但这正是这种关联规则算法的，他们的魅力就是现实中我们明明没有关系的，表面上看压根儿马牛不相及的两个实体。但是他们在数据上却有一定的关联，那么对于大数据呢？大家注意了。

我们往往强调的是重在结果，而不是它的起因。我能够把这个结果把这个规律给它揭示出来，至于造成这规律的原因，我可以先不管，我有了结果之后，我就可以在这个基础上做出相应的决策，那么这就是。这种基于关联规则的算法，具体的算法。

它的数学原理我们不说了，那么这里边呢，典型的算法呢，有两个一个呢，是a，还有一个呢，是fp gross。这个的数学原理我们就不说了，这个时间关系没有时间说了，大家感兴趣的话呢。

去查一下相关的相应的算法，那么这种。关联规则的算法呢？虽然看似老，但现在仍然有着非常广泛的应用。比如说我们可以用我们。上次课讲的那个慢hot那个大数据大大数据的那个那个原理，然后呢，通过某个网站的两个月的日志信息，哎。

我们在上网上任何的网站。他在他的服务器上呢，都会留下所有用户的上网的记录，上网的痕迹，这个我们称之为网络的日志，那么通过慢hot对这个利用关联规则的算法对日志来进行模式挖掘。那么就可以得到相应的知信度的结果，然后呢，可以为这个网站未来的进一步的发展提供相应的决策，这就是关联规则。

下一类算法呢，叫做异常检测算法，异常检测算法呢，顾名思义就是我们把众多数据中的那些孤立点，异常点给它检测出来，我们收集到的很多数据。可能有些数据是带有噪声的数据，或者本身就是一些错误的数据，我们把这些数据呢，给它称之为孤立点。

或者是异常点，那么我们通过这种算法呢，给它检测出来。在数据库中，包含着少数特殊的数据对象，他们和数据的一般行为特征不一致。我们把这些数据对象呢称为孤立点或者是异常点。异常点的检测和分析也是我们进行数据处理中必不可少的一类算法，那我们呢？给它称之为异常点挖掘。

那么这类算法呢？大家注意了，它又包含很多。包含基于统计的异常点检测算法，基于距离的异常点检测算法，注意这里的距离就是我们前面给大家讲的那个距离和相似度计算的那个距离。和基于密度的异常点检测算法，这些呢，都是属于这种异常检测算法，那么具体的算法的。

实现我们也不说了，下一类大数据的算法呢，叫做时序预测算法，这个也是它字面上的意思，顾名思义呢，就是根据过去的变化趋势，未预测未来的发展。是整个大数据和数据挖掘中最重要的应用，那么这种基于时序预测的算法呢？也有很多我们大概的。

也不给大家多说了，那么这种时序预测呢？在现实中呢？也挺有用，比如说。我们可以通过这种数据的预测来预测某一个网站发展的未来的营收，某个公司发展的未来的营收，甚至现实中。某一个油油田或者是气田或者是煤矿，它未来的产量等等。

那么这些呢，都是属于这种时序，预测好了，这儿呢，就是给大家补充的，第一个节我们。在这里边呢，称为六点六节，其实是给大家补充的第一节，这一节呢。

用很简洁的关系给大家呢，介绍了大数据和。人工智能的相关的一些算法的结合的应用，那其实本质是大数据的算法，它就是人工智能的算法本身大数据。的应用和人工智能的主干的课程，它有着非常深的相似度和重复度。好，这是六点六节。六点七节呢。

就是deep learning呢，深度学习了，深度学习呢，我们真正要学的话，大家未来要学的话，比如说咱们在座的同学，未来如果你要读研的话。那深度学习是你的必修课啊，你未来研究生，你即使我们读这个信安专业的。

做这方面的研究，那机器学习深度学习是我们必须用的工具，要真正学的话呢，建议大家呢，还是找那本花书来看一看？就是我在这一章一开始给大家推荐的第二本书花书，它本身就是深度学习的缔造者，写的这本书我们这儿呢，时间关系呢，只是给大家。

大概的扫盲性的讲一下它的一些基础的内容吧，那么首先呢，有一个关于深度学习的怎么样，一步一步发展的一个帖子是csd n上的一个帖子。这个帖子呢，写的非常的浅显易懂，就是来讲这个深度学习怎么样，一步一步的发展到现在的，那么很值得一看，我把这个帖子的这个这个。

域名给他放在这上面了csd n，大家应该都上过吧，中国程序员网，你要是在编程领域的话，这个网站是必上的，那么大家可以去看一看这个帖子，我们课上呢就。就不再给大家重复的来看了。好，那下边我们来看一下深度学习的基本的背景。

机器学习的概念呢，要远早于深度学习，机器学习的概念呢，我们说了嘛，人工智能概念诞生没多久，机器学习的概念呢，就诞生了，但是机器学习呢它。它在发展的过程中呢，有很多问题呢。

没有得到很好的解决，比如说在深度学习，没有出现之前有很多问题呢，人们认为在短时间内。光靠机器学习是很难解决的，比如说精确的图像识别，比如说语音识别，比如说自然语言理解，比如说天气预测。比如说基因表达。

内容推荐等等，当然现在我们再来看这些问题，这都不是个事儿，但是在十几年前，这些问题个个都是大问题。还记不记得我们这个课程开始的时候给大家讲的那个例子就是三体二黑暗森林中那个时候零八年的时候，刘慈欣在写三体二的时候，那个时候还认为在科幻世界里边人脸识别是不可能实现的呢。那么短短的这十几年，那情况起到了翻天覆地的变化。

那这主要的背后都是深度学习的原因，那我们来看一下早期的这些问题是怎么样来处理的？上面的问题呢，其实大多呢，都可以把它归结成模式识别，比如说人脸识别，语音识别，那么都属于模式识别，那早期计算机是怎么样来处理模式识别呢？它采用了一种我们自认为很正确。

但是计算机却极难实现的一种方法。我把它称为特征提取法。就说什么叫特征提取法呢？早期计算机识别一个人脸非常的困难，我们认识一个人，只要认识他之后，我们再看到这个人，几秒钟我们在脑海中就能够把这张人这个人的人脸和我们脑海中记中这个人给他呢？打上这种等于号，但是计算机是怎么识别人脸的？

太难了，它采用的是这种方法，就是首先在低层次的水平上来进行感知。然后在这个基础上来进行预处理，然后呢，把人脸的特征给它提取出来，比如说这个人，他的眼部的特征是什么？鼻子的特征是什么？眉毛的特征是什么？

通过特征进行提取，然后进行特征选择，在这个基础上来进行识别和预测，那大家想想一个人可能两个人的人脸，他仅仅是有一些细微的差别。这些细微的差别，我们用眼睛来看的话，那么可以瞬间来进行发现，比如说我们人的眼，如果把它想象成一个数码相机的话。

它。它是一个精度非常高的，那么几十亿以上像素的这种数码相机，这种高的精度，所以说我们可以瞬间。在一大片的像素点中，找准确的找出我们想要定位的那些像素点，但是机器达不到的。机器是怎么样来存储图像的？它采用的是rgb 3基色的原理。

大家在学大一的时候，计算机基础入门的时候都知道。我们机器是按照一个又一个的像素点来存储图像的，每一个像素点我要记录rgb它三种颜色的程度r就是red嘛。g就是green嘛b就是blue嘛，这三种颜色的程度，然后呢？每种程度呢？用零到255，256种不同的程度来表示我们所说的16味色真彩色。

呃24位色真彩色就是二的24次方，那就是三三个二百五十六三个二的八次方那这。我们想准确的按照图像识别的模式来找出特征，不是不能，但是非常非常的难，这也是为什么。我们早期的时候想做到人脸识别，图像识别是非常非常困难的，比如说早期的时候他是怎么样来做呢？通过传感器像CMOS摄像头来获得图像数据。

然后通过预处理特征，提取特征选择，再经过推理，预测和识别最后一个部分，才是机器学习部分。绝大部分的工作呢，都是在最后一个部分机器学习部分来做的，那早期你把时间推到20年前，你去看那个时候的图像识别，人脸识别100%的论文都是焦距在这方面。

而中间最干，最关键的部分呢，就是怎么样来完成特征表达。比如说这个人的人脸。张三的脸和李四的脸最大的不同点就是张三脸的特征和李四脸的特征是不一样的，那我怎么样把这个特征表达来给它提取出来？这对算法起到非常关键的作用。那么，既这个系统的主要测试的工作都好在这一块儿，但是这一块儿呢。

实际是由人工来完成的，由人工来提取特征。那么，这就造成了一个问题，就是一般来说，我们给的特征越多，给出的信息越多。识别的准确率也就越高，这儿有一张图，这个有点反光。

大家可能看不清，它是早期的那种图像识别的算法中我给的特征。越多，那么它的准确率就越高，比如说我一开始只给了100个特征的话，那么我这个人脸识别的准确率极低。差不多只有50%到60%左右，但是我给到了八百一千二或者是1600个特征的话，那么人脸识别的准确率呢？就变高了。

但是早期最高。也就是70%多70%多，远没有像现在这种识别那么准确，所以说呢，早期的这种特征识别呢，主要是根据特征，但是问题来了。就是你给的特征多的话，准确率算法是和之前相比有了提高，但是造成的另外一个。

问题就是计算的复杂度增加，探索的空间大，用来训练的数据在每个特征上会变得更加的稀疏。所以说这是一个两难的境地，因此早期做特征识别的时候呢，人们得出的结论是什么特征？其实不是越多越好。它需要呢？有多个特征来进行学习和确定，所以说得出结论呢。

就是早期获得好的特征是识别的观念，那么存在着大量的基于人工的特征的不同。这都是早期的那种识别特征的，像那些算法手工选取费时费力，那么需要启发性的专业的知识。在很大程度上，他依靠着经验和运气，那么人们就想了，我们能不能自动的来进行学习特征，这就是后来。

深度学习为什么能够一药而红起来的原因？其实，深度学习那篇经典的文章在零六年提出的时候很长一段时间内，人们只是把它。当成是一个一般性的，一个科研的论文，但是真正让深度学起一炮而红的最重要的一个事件。就是一二年一三年的时候，那个image net大家都知道，那有一个非常著名的一个人脸识别网。

叫image net，那么这个网里边呢？存储了。现实中，大量的数以一万计的不同的这种图像，包括人脸图像，包括其他的图像，那么每年呢，都举办这个图像识别大赛。那以前的时候呢。

计算机在这方面是远不如人类的，它的识别率和人类相比，简直就不在一个等级上，但是呢，有了深度学习技术的特征之后。一夜之间，计算机的识别率一下子提高了10%几，提高了20%几，那么做的又快又准又好，一下子把所有的这种人类的顶尖儿的这种识别的冠军。

都抛在了后边，正是因为这个image net比赛，从这以后，人们才逐渐认识到深度神经网络的重要性。那么下边呢，就是深度神经网络的，它的基本的概念了，深度神经网络呢，它的主要的关键呢，就是能够自动的学习特征。

它的基础呢是在。传统神经网络的基础上呢，给它做大做强，注意了它其实是一个完全旧的概念deep learning这个词。在零六年之前。是没有的，是辛顿发明的这个词，但是呢，神经网络我们说了它的概念比。神经人工智能概念还要早。

最早的神经网络的原型在四一年，四二年的时候就出现了，但是我们前面说早期的神经网络呢？空就像空中楼阁一样，光有理论没有实现的土壤，但是到后来我们知道，现在随着大数据的兴起，随着gpu的兴起，神经网络既有数据又有并行计算的这种算力。所以说再加上辛顿和他的两个学生。

这一批人在这上面孜孜不断的研究，终于在零六年以后开花结果，神经网络它的特点深度神经网络和它的。和传统神经网络相比呢，它具有这么一个特点，第一个它的层数很深，那一开始的时候呢，新盾提出的深度神经网络呢，认为只要超过五层就行了，但是现在。

那么，五层那根本都都不好意思跟他打招呼，现在动辙大模型像这种神经网络，那么它动辙。千万层甚至更多的树木，它具有更强大的拟合能力，需要更有效的训练的方法。当然，它。它在CPU上是很难实现的。

它必须得在gpu的基础上来完成大规模的数据集，那么。深深度学习呢，起源于深人工智能和这个机器学习的研究，那前面这个我们不说了，马上就下课了，然后真正的起源呢就是。零六年的时候，辛顿和他的两个学生，然后发表的那个深度学习的标志着deep learning诞生的那篇文章。

所以说呢，零六年的时候呢，深度学习。被我们誉为深度学习的元年，那么这幅图非常有名啊，这个站在中间的是辛顿旁边的两个呢，是他的两个学生，这三个人呢，在后来共同。获得了因为深度学习方面的贡献呢。

共同获得了图灵奖，最右边这个人呢，文达当然也是现有的人工智能学界的非常著名的学者。那么，这是它的起源和发展，然后再往后就不说了，这是辛顿的介绍我们这个。在前边儿给大家讲人工智能发展史的时候呢，都说过了，然后呢。

这是新顿的两个学生，一个叫约束阿本吉奥，还有一个呢。叫叫那个呃，另外一个他一个加拿大人，一个法国人，现在仍然是活跃在人工，这个就是机器学习的。最呃最那个第一线，然后呢。

我们现在构成的深度深度网络呢，它动辙呢，都有几万亿个神经元呃。几千万以上的参数好了，那具体的那个我们就不说了，下边还有就是这是当年image net的时候，就是深度学习参加这个image net大。大赛的时候，一夜之间，他把这个大赛的准确率呢提高了十几个百分点。

这是使得深度学习。这个相应算法的威力呢，或者是相应技术的威力呢，也被人们被科学界呢所认知，这幅表呢，就是每一年参加这个e mag大赛的时候呢。他们的前几名的这个正确率，通过这个大家可以看出这个image net的威力，那最早的就是二零一二年的时候那么。它改发发生的相应的改变。

后边呢，这个卷积神经网络还有呢，深度神经元这个不说了，这个因为因为要说的话每一个。都得讲半天，这个错在课件里边呢，大家自己去看一下吧，还有最后呢，这部分呢，降了一下阿尔法狗。

这个我们时间关系呢，也不说了，回头大家呢，自己去看吧，好了。至此呢，整个给大家补充的内容呢，我们很快的吧，囫囵吞枣的给大家呢，讲完了讲。

讲到了最后。OK，最后呢，把我们课程做最后一个总结。我们整个课程呢，到现在为止就全部学完了，48个学时说短不短，但是呢，也不算太长。这门儿课程的名字叫人工智能。

基础确实呢，我们在课上呢，也只给大家讲了相应的基础，人工智能呢，它本身呢，现在是。当前，社会上最最热门的学科那么大家，热门的程度我们就不用说了吧。现在。

无论是在就业上还是在招生方面，那么。AI呢，都是热中之热，那本身作为计算机专业大类的，我们对这门儿，对相应的这门儿课程的理解和认知，我个人认为。光靠我们这一个48个学时的课程是远远不够的，大家呢？

未来呢？大部分同学呢？将会。进一步的深造或者是投入到相关的这个信安方面的工作，但是对AI的学习是不应该仅止步于这门课的。那么我个人觉得大家未来不管是就业也好，还是读研也好，那么关于机器学习，关于深度学习，包括现在热门的强化学习的。

很多的内容大家都有必要去深入的再学一本，再认真的去学一把，当然了，如果咱们读研的话呢，那很多课程是大家研究生的必修课。这我就不用说了，但即使大家是工作的话，那么呢，你可以不不成为AI的专家，但是呢。

对于AI的基本原理。像我们课上讲的，包括现在的一些算法的一些核心的内容，你应该至少有所了解，你不一定能够发明新的算法，但。但是未来，比如说让你利用深度神经网络构建来解决现实问题的话，你至少要能够学会怎么样来构建网络，怎么样来调参。

我觉得这是我们每个同学。都应该达到的好了，那最后呢？我们虽然课程结束了，但是大家呢？以后对我们这门课包括整个计算机方面啊。只要是我熟知的，我擅长的，大家有什么问题呢？也可以呢。

随时给我来联系，那我们今天的课呢，就上到这儿大家。大家注意啊，不要忘了好好的复习，还有呢，就是所有的课件我都发到群里了，大家要好好的看课件，那尤其是好多同学都没没我看都没课本。没课本更应该重点看课件。

尤其是课件里边打星号的，尤其是我们课堂上的练习和课后的那作业，那两次作业都是重中之重，大家认真的来进行复习。就可以了，好我们的课就上到这。