你好，从今天开始我们就进入了数据挖掘的第一课。这一课时我将借助 5W1H 的思想来带你从整体上了解下数据挖掘，比如什么是数据挖掘、为什么要做数据挖掘、在哪些场景下用数据挖掘，以及怎么做 数据挖掘。在后面的课时里，我会从这条主线上逐渐细化，为这个 “骨架” 填充肌肉和血液，让它逐渐丰满起来。

什么是数据挖掘?

这个问题看似很简单，但似乎也很难有一个明确的答案。

如果非要给数据挖掘一个定义的话，那么我认为**数据挖掘就是寻找数据中隐含的知识并用于产生商业价值。**也就是说，它是我们在数据中（尤其是在大量的数据中）找到一些有价值，甚至是非常有价值的东西的一种手段。

为什么要做数据挖掘?

技术与商业就像一对双生子，在互相促进中不断演进发展，随之而来的就是各大公司业务突飞猛进，很多新模式也涌现出来，使得数据量激增。

面对数以千万甚至上亿，以及不同形式的数据，很难再用纯人工，或者纯统计的方法从成千上万的变量中找到其隐含的价值。

我们需要一种规范的解决方案，能够利用并且充分利用这些数据里的每一个部分，通过一些自动化的机器学习算法，从数据中自动提取价值。而数据挖掘就提供了这样一系列的框架、工具和方法，可以处理不同类型的大量数据，并且使用复杂的算法部署，去探索数据中的模式。

总之，数据挖掘的产生动因主要有以下 3 点。

- **海量数据。** 随着互联网技术的发展，数据的生产、收集和存储也越来越方便，海量数据因此产生。比如，我们常用的微信，每天要产生超过 380 亿条数据；今日头条每天要发布上百万的新文章；淘宝每天有上千万的包裹要发出。
- **维度众多。** 在一个多维度的数据中，每增加一个维度都会增加数据分析的复杂程度。比如点外卖事件涉及的维度就有：浏览饭店的菜品（形式有文字、图片或语言、视频等）、浏览时间、下单价 格、交易处理、分配配送员及 GPS 信息、完成订单后的评价等。
- **问题复杂。** 通常用数据挖掘解决的问题都比较复杂，很难用一些规则或者简单的统计给出结果。如果让开发者写一个微波炉的智能控制逻辑，我想难度不是很大，即便是有十几个，甚至几十个按 钮的控制中心也不过是多花费一点时间而已。但如果编写一段代码来区分某图片中是否有一只猫 咪，那要考虑的问题就太多了，使用传统的方法很难解决，而这恰恰是数据挖掘所擅长的。

以上是我们进行数据挖掘的初衷，在后续的课程中你也会看到，随着这些问题的出现，它们在数据挖掘中是如何被解决的。

数据挖掘有什么用处?

既然数据挖掘是一种方法，那就要用它去解决一些问题。下面我就来具体讲一下你最关心的，也是最实际的问题，数据挖掘到底有什么用处。

1. 分类问题

分类问题是最常见的问题。比如新闻网站，判断一条新闻是社会新闻还是时政新闻，是体育新闻还是娱乐新闻？这就是一个分类问题，也就是对已知类别的数据进行学习，为新的内容标注一个类别。

新浪导航栏图

2. 聚类问题

聚类与分类不同，聚类的类别预先是不清楚的，我们的目标就是要去发现这些类别。**聚类的算法比较适合一些不确定的类别场景。**

比如我们出去玩，捡了一大堆不同的树叶回来，你不知道这些树叶是从什么树上掉落的，但是你可以根据它们的大小、形状、纹路、边缘等特征给树叶进行划分，最后得到了三个较小的树叶堆，每一堆树叶都属于同一个种类。

3. 回归问题

简单来说，回归问题可以看作高中学过的解线性方程组。它的最大特点是，生成的结果是连续的，而不像分类和聚类生成的是一种离散的结果。

比如，使用回归的方法预测北京某个房子的总价（y），假设总价只跟房子的面积（x）有关，那么我们构建的方程式就是 ax+b=y。如何根据已知 x 和 y 的值解出 a 和 b 就是回归问题要解决的。回归方法是通过构建一个模型去拟合已知的数据（自变量），然后预测因变量结果。

4. 关联问题

关联问题最常见的一个场景就是**推荐**，比如，你在京东或者淘宝购物的时候，在选中一个商品之后，往往会给你推荐几种其他商品组合，这种功能就可以使用关联挖掘来实现。

京东组合购买推荐图

到这里，我们清楚了数据挖掘可以解决哪些问题，那具体应该怎么做呢？

数据挖掘怎么做？

数据挖掘，也是有方法论的。实际上，数据挖掘经过了数十年的发展和无数专家学者的研究，有很多人提出了完整的流程框架，这对于我们来说简直是福音。当然，如果你在使用的过程中觉得这些东西有问题，或者还有改进的空间，那也不要惧怕权威，尽信书则不如无书嘛。

在这里，我讲一个应用最多的 CRISP-DM（Cross-industry Standard Process for Data Mining，跨行业数据挖掘标准流程）方法论，不要被这么长的名字吓到，这里我们先简单地了解数据挖掘的操作步骤有哪些，后面我也会逐一详细讲解。

下面我们就来看一下，如何依照这 6 个步骤进行数据挖掘。

CRISP-DM 流程图

1. 业务理解（Business Understanding）

想象你在一个外贸公司上班，有一天，你的老板突然给你说：“小明啊，你能不能训练一个模型来预测一下明年公司的利润呢？”

这就是一个业务需求了，若要解决这个问题，首先要弄明白需求是什么，这就是业务理解，或者也可以叫作商业理解。比如，你要搞清楚什么是利润、利润的构成是什么样的、利润受什么影响，同时老板说的利润是净利润还是毛利润等问题。

业务理解，主旨是理解你的数据挖掘要解决什么业务问题。　任何公司启动数据挖掘，都是想为业务赋能，因此我们必须从商业或者从业务的角度去了解项目的要求和最终的目的，去分析整个问题涉及的资 源、局限、设想，甚至是风险、意外等情况。**从业务出发，到业务中去。**

2. 数据理解（Data Understanding）

明白了问题，还要明白解决问题需要什么数据。比如这个时候，你的老板又跟你说了：“小明啊，我想改改需求，能不能多做几个模型，把竞品公司明年的利润也都算算，我想对比一下。”然而 “巧妇难为无米之炊”，你根本就没有这个数据，这个需求也就无从完成了。

数据理解阶段始于数据的收集工作，但我认为重点是在业务理解的基础上，对我们所掌握的数据要有一个清晰、明确的认识，了解有哪些数据、哪些数据可能对目标有影响、哪些可能是冗余数据、哪些数据存在不足或缺失，等等。

需要注意的是，数据理解和业务理解是相辅相成的，因此你在制定数据挖掘计划的时候，不能只是单纯地谈需求，这也是大多数初入门的数据挖掘工程师容易忽略的。数据理解得不好，很可能会导致你对业务需求的错误评估，从而影响后续进度甚至是结果。

3. 数据准备（Data Preparation）

完成上面两个步骤后，我们就可以准备数据了。你需要找销售要销售数据，找采购要采购数据，找财务要各种收入、支出数据，然后整理所有需要用到的数据，想办法补全那些缺失的数据，计算各种统计值，等等。数据准备就是基于原始数据，去构建数据挖掘模型所需的数据集的所有工作，包括数据收集、数据清洗、数据补全、数据整合、数据转换、特征提取等一系列动作。

事实上，在大多数的数据挖掘项目中，数据准备是最困难、最艰巨的一步。如果你的数据足够干净和完整，那么在建模和评估阶段所付出的精力就越少，甚至都不必去使用什么复杂的模型就可以得到足够好的效果，所以这个阶段也是十分重要的。

4. 构建模型（Modeling）

也可以叫作训练模型，在这一阶段，我们会把准备好的数据喂给算法，所以这个阶段重点解决的是技术方面的问题，会选用各种各样的算法模型来处理数据，让模型学习数据的规律 ，并产出模型用于后续的工作。

对于同一个数据挖掘的问题类型，可以有多种方法选择使用。如果有多重技术要使用，那么在这一任务中，对于每一个要使用的技术要分别对待。一些建模方法对数据的形式有具体的要求，比如 SVM 算法只能输入数值型的数据，等等 。 因此，在这一阶段，重新回到数据准备阶段执行某些任务有时是非常必要的。

5. 评估模型（Evaluation）

在模型评估阶段，我们已经建立了一个或多个高质量的模型。但是模型的效果如何，能否满足我们的业务需求，就需要使用各种评估手段、评估指标甚至是让业务人员一起参与进来，彻底地评估模型，回顾在构建模型过程中所执行的每一个步骤，以确保这些模型达到了目标。在评估之后会有两种情况，一种是评估通过，进入到上线部署阶段；另一种是评估不通过，那么就要反过来再进行迭代更新了。

6. 模型部署（Deployment）

整理了数据，研究了算法模型，并通过了多方评估，终于到了部署阶段。此时可能还要解决一些实际的问题，比如长期运行的模型是否有足够的机器来支撑，数据量以及并发程度会不会造成我们部署的服务出现问题，等等。但是，关于数据挖掘的生命周期可能还远未结束，关于一些特殊情况的出现可能仍然无法处理，以及在后续的进程中，随着新数据的生产以及变化，我们的模型仍然会发生一些变化。所以部署是一个挖掘项目的结束，也是一个数据挖掘项目的开始。

总结

今天，我们主要来认识一下 “数据挖掘”，让它不再是一个抽象的名词，从“它解决什么问题” 和“怎么做” 的角度建立了全面而具体的认识。后面的篇幅，我也将围绕这些内容进行展开和扩充讲解。

另外，学习也是一个自主的过程，我建议你平时多找一些资料来学习、补充，任何关于数据挖掘的疑问、工作疑惑等，你都可以在留言区与我或专栏的所有用户一起沟通交流。