本课时，我将为你介绍数据挖掘操作流程的第二个环节，准备数据。

在对业务和数据有了清醒的认识之后，你就要开始收集、处理数据了。这个环节看起来好像是一个非核心环节，实际上在整个过程中却是最重要、最耗时的环节。

就如 2008 年北京奥运会的成功离不开城市规划、场馆建设、志愿者招募等一系列准备工作一样，数据准备在数据挖掘中同样也承担着这样一个重要的角色。原始的数据通常不可能跟你的算法所适配，而且其本身也存在着各种各样的问题，如不够准确、格式多样、部分特征缺失、标准不统一、特殊数据、错误数据等，这些问题都将在一定程度上影响你后续算法模型的训练和实施。

为了避免上述麻烦，我将带你一步步避坑，准备出合适模型的数据。

找到数据

在一个公司中，数据往往会有很多的存在形式，比如它们所属的业务部门不一样，使用的数据库类型就可能不一样，存储数据的方式也有可能不一样等问题。所以，对于你要做的项目来说，就可能需要很多不同来源的数据。你要知道每个项目需要什么数据，并从哪里获取。尽管在一些大的公司存在数据平台部门、数据仓库部门，但这仍然不能保证你所需要的数据只用一种方法就能获取到。所以在这一步，可能需要你掌握一些数据库的使用技巧，如常用的关系型数据库 MySQL、大数据使用的 Hbase、Hive、搜索引擎数据库 ES、内存数据库 Redis，还有图数据库，如 Neo4j 或者 JanusGraph 等，甚至还要跟各种业务部门沟通协商以获取数据。数据库的内容我就不在本课时中一一介绍了，如果感兴趣你可以去官网深入了解。

当你从各种地方收集到所需要的数据之后，最好是能够把它们进行简单的整理，如用统一的 id 把数据整合在一起等，以便后面查询和使用。

准备好需要的数据后，就要对它们进行一系列的加工，从而达到后期训练模型的要求。

数据探索

在该阶段，为了尽可能获得足够多的特征，你要对数据进行分析、预处理以及转换等基础工作，以构建出更加贴合你所要预测结果的特征，这使得数据维度大量扩展，所以我把这个环节叫作把数据变多或者数据升维。

假设你要做一个给新闻内容分类的项目，已经从数据仓获取了新闻内容、新闻标题、新闻发布时间等数据，并从运营部门获得了运营给这些新闻标注的分类数据。这时候你要做的就是把数据变多，可以进行如下操作：

- 把内容进行分词，这样就获得了一个分词后的字段；
- 把分词后的内容进行词语的统计，看看哪个词出现得更多；
- 同样地把标题进行分词，进行词语的统计；
- 还可以对词语的词性进行标注，获得一份词性数据；
- 你可以找到一些特殊的词，比如名人的名字、机构的名字、地点的名字等一些信息。

通过这些处理，可以看到你的数据是否存在问题，比如异常值、数据的偏差、缺失，等等。如果是数值型的数据，还可以通过计算均值、方差、中位数、标准差、最大值、最小值等去探索、扩展。

有了足够多的数据，接下来就要对其进行整理，提取对项目最有用的部分。

数据清洗

终于讲到了这个，在整个数据准备，甚至是整个数据挖掘过程中，最烦琐、最头疼的步骤——数据清洗。如同你打扫卫生时，会把不需要的东西扔掉，需要的东西留下来摆放整齐一样，数据清洗步骤就是要做这样一个工作，处理扩展后的数据、解决所发现的问题，同时又要顾及处理后的数据是否适合应用于下一个步骤，所以我也把这一步骤称作把数据变少。

数据清洗主要包含如下 5 个方面的内容：

1. 缺失值的处理

在美好的童话世界中，我们的数据都是完完整整的，拿来即用。实际上，在工作中最常见的一个问题就是数据的缺失，比如一条新闻可能只有正文没有标题、发布地点、发布时间等任意数据。你需要区分这些数据缺失的情况，因为有些是业务所允许的缺失，而有些则是错误情况导致的。通过分析，了解数据缺失的原因以及数据缺失的影响范围，这会关系到你后面如何来处理缺失值。

关于缺失值的处理，一般就 3 种情况：**删掉有缺失值的数据**；**补充缺失值**；**不做处理。**当然这些处理方式也依赖于数据是否可以被补充、缺失值是否重要，以及你所选用的算法能否处理缺失的情况等因素。

2. 异常值的处理

异常值通常说的是那些与样本空间中绝大多数数据分布差距过大的数据，这些数据的产生通常有 2 种情况：

- **错误的情况**，比如医院录入病人病历的时候，忘了给数字输入小数点，导致某个人的身高显示为 173 米，等等；
- **正常的情况**，就需要重视了。比如在平均充值为 100 元的游戏中，有人充了 100 万元，这是一个真实的结果，但是如果直接使用到模型中可能会影响到平均值的计算，影响模型训练的效果；再比如只有 1000 万在线用户的 App ，突然拥有十亿的在线用户，这就有可能是应用网络受到了攻击，等等。

不同情况的异常值有不同的处理办法：

- 数据本身的错误，需要对数据进行修正，或者直接丢弃；
- 数据是正确的，需要根据你的业务需求进行处理。如果你的目标就是发现异常情况，那么这种异常值就需要保留下来，甚至需要特别关照。如果你的目标跟这些异常值没有关系，那么可以对这些异常值做一些修正，比如限定最大值和最小值的标准等，从而防止这些数据影响你后面模型的效果。

3. 数据偏差的处理

这也是非常常见的问题。没有什么数据是非常对等和均衡的，越是天然的数据越是符合正态分布的规律。比如 UGC 内容（User Generated Content，用户生成内容）的质量，质量较差的内容占大多数，质量好的占少数，质量非常好的是少之又少。这是一种正常的现象，但是对于算法模型来说，有些算法会倾向于预测占比较大的数据，比如说质量好的内容只占 2%，而质量差的内容占到了 98%，模型倾向于给出质量差的结果。如果给出所有的结果都是 “差”，那该模型的准确率也能达到 98% 了，可这并不是我们想要的结果。

数据偏差可能导致后面训练的模型**过拟合或者欠拟合**，所以处理数据偏差问题也是你在数据清洗阶段需要考虑的。如果你需要比较均衡的样本，那么通常可以考虑丢弃较多的数据，或者补充较少的数据。在补充较少的数据时，又可以考虑使用现有数据去合成一些数据，或者直接复制一些数据从而增加样本数量。当然了，每一种方案都有它的优点和缺点，具体的情况还是要根据目标来决定，哪个对目标结果的影响较小就采取哪种方案。

4. 数据标准化

在处理完数据的问题之后，你就该对数据的标准进行整理了，这可以防止某个维度的数据因为数值的差异，而对结果产生较大的影响。在有些算法中，每一个维度的数据标准都需要进行统一；而在另外一些算法中，则需要统一数据的类型。比如在预测一个地区的房价时，房屋的面积可能是几十到几百的数值范围，房屋的房间数可能是个位数，而地区平均单价可能是以万为单位的。一个处理方法是你把这些维

度的数据都进行标准化，比如把这些数据都规范到 0~1 的区间，这样使用不同的单位来衡量的数据就变得一致了。

5. 特征选择

特征选择就是尽可能留下较少的数据维度，而又可以不降低模型训练的效果。一个项目中，数据的维度可能会有成百上千，比如一个文本中，每一个词或者每一个字都是一个维度，那么要用一个向量去表示一篇文章，这个向量可能需要有上万个维度，所以你要排除掉那些不重要的部分，把重要的部分保留下来。

也许你会认为数据的维度越多越好，但实际上，**维度越多，数据就会越稀疏，模型的可解释性就会变差、可信度降低。**过多维度还会造成运算的缓慢，尤其是一些运算量较大的算法，同时那些多余的维度可能会对模型的结果产生不好的影响，如某个维度的数据跟结果实际上并没有什么关系，数据也呈现出一种随机的情况，如果没有把这部分数据排除掉，就可能会对某些算法产生影响。

这个时候就需要用到特征选择的技巧，比如自然语言处理里的关键词提取，或者去掉屏蔽词，以减少不必要的数据维度。对于数值型的数据，可以使用主成分分析等算法来进行特征选择，如果你对这部分内容有兴趣，可以在网上找一些资料进行更深入的学习。

构建训练集与测试集

在数据进入模型之前，你还需要对其进行数据采样处理。如果说前面的部分是为了给模型提供一个好的学习内容，那么数据采样环节则是为了评估模型的学习效果。

在训练之前，你要把数据分成训练集和测试集，有些还会有验证集。

- 如果是**均衡的数据**，即各个分类的数据量基本一致，可以直接随机抽取一定比例的数据作为训练样本，另外一部分作为测试样本。
- 如果是**非均衡的数据**，比如在风控型挖掘项目中，风险类数据一般远远少于普通型数据，这时候使用分层抽样以保障每种类型的数据都可以出现在训练集和测试集中。

当然，训练集和测试集的构建也是有方法的，比如：

- **留出法**，就是直接把整个数据集划分为两个互斥的部分，使得训练集和测试集互不干扰，这个是最简单的方法，适合大多数场景；
- **交叉验证法**，先把数据集划分成 n 个小的数据集，每次使用 n-1 个数据集作为训练集，剩下的作为测试集进行 n 次训练，这种方法主要是为了训练多个模型以降低单个模型的随机性；
- **自助法**，通过重复抽样构建数据集，通常在小数据集的情况下非常适用。

思想准备

准备数据可能是数据挖掘所有环节中，**最苦、最累、耗时最长**的一环了。由于实际生产中的数据，会存在各种各样的问题，一如我上面说的，数据缺失、异常、偏差等，而对于数据的准备其实没有一个统一的标准算法去解决。所以在这个环节，一定要保持认真仔细的态度以及平和的心态，做好数据准备工作是获得一个好结果的必由之路。

准备数据不是独立存在的过程。不是说，你一次性做完数据准备工作就结束了，后面的模型训练和模型评估环节与数据准备紧密相关，当你的模型出现错误，结果达不到预期，往往需要重新回到数据准备环节进行处理，反复迭代几次最终才能达到你期望的目标。

总结

写到这里呢，关于数据准备的工作已经进行得七七八八了，不知道你看完之后是否对准备数据有了一个比较全面的认识呢？在该环节，我们将走下象牙塔，走进实际的工作当中，处理在现实中数据存在的各种问题以使得数据达到我们模型算法的要求。

通过这些步骤，可以说数据准备已经比较充分了，数据挖掘中最困难、最烦琐的一个步骤已经结束，接下来我们就要进入到模型训练的环节了。在这里呢，我想给大家布置一个问题，你可以观察一下你所能够获得的数据，仔细查看里面会有什么样的问题呢？欢迎你在留言区写下你的问题，与我和其他同学一起讨论。