《python数据分析与挖掘实战》笔记第1章

[《python数据分析与挖掘实战》笔记第1章 - 腾讯云开发者社区-腾讯云 (tencent.com)](https://cloud.tencent.com/developer/article/1796252)

**文章目录**

* + [第1章：数据挖掘基础](https://cloud.tencent.com/developer?from=10680)
    - [1.3、数据挖掘的基本任务](https://cloud.tencent.com/developer?from=10680)
    - [1.4、数据挖掘建模过程](https://cloud.tencent.com/developer?from=10680)
      * [1.4.1、 定义挖掘目标](https://cloud.tencent.com/developer?from=10680)
      * [1.4.2、 数据取样](https://cloud.tencent.com/developer?from=10680)
      * [1.4.3、数据探索](https://cloud.tencent.com/developer?from=10680)
      * [1.4.4、数据预处理](https://cloud.tencent.com/developer?from=10680)
      * [1.4.5、挖掘建模](https://cloud.tencent.com/developer?from=10680)
      * [1.4.6、模型评价](https://cloud.tencent.com/developer?from=10680)
    - [1.5、常用的数据挖掘建模工具](https://cloud.tencent.com/developer?from=10680)

**第1章：数据挖掘基础**

**数据挖掘**是从大量数据（包括文本）中挖掘出隐含的、先前未知的、对决策有潜在价值的关系、模式和趋势，并用这些知识和规则建立用于决策支持的模型，提供预测性决策支持的方法、工具和过程。

**1.3、数据挖掘的基本任务**

**数据挖掘的基本任务**包括利用分类与预测、聚类分析、关联规则、时序模式、偏差检测、智能推荐等方法，帮助企业提取数据中的商业价值，提高企业的竞争力。

**1.4、数据挖掘建模过程**

**1.4.1、 定义挖掘目标**

理解挖掘任务，定义挖掘目标

**1.4.2、 数据取样**

抽取数据的标准，一是相关性，二是可靠性，三是有效性，而不是动用全部企业数据。通过对数据样本的精选，不仅能减少数据处理量，节省系统资源，还可以使我们想要寻找的规律性更加凸显出来。

**衡量数据取样的质量标准：** 1）资料完整无缺，各类指标项齐全。 2）数据准确无误，反应的都是正常（而不是异常）状态下的水平。 对获取的数据，可从中进行抽样操作。抽样的方式是多种多样的，常见的方式如下：

* 随机抽样
* 等距抽样
* 分层抽样
* 从起始顺序抽样
* 分类抽样

**1.4.3、数据探索**

数据探索主要包括：异常值分析，缺失值分析，相关分析和周期性分析等。

**1.4.4、数据预处理**

数据预处理主要包括：数据筛选、数据变量转换、数据缺失值处理、坏数据处理、数据标准化、主成分分析、属性选择、数据规约等。

**1.4.5、挖掘建模**

样本抽取完成并经预处理之后，接下来要考虑的问题是：本次建模属于数据挖掘应用中的哪类问题（分类、聚类、关联规则、时序模式或者智能推荐），选用哪种算法进行模型构建？这一步是数据挖掘工作中的核心环节。

**1.4.6、模型评价**

模型评价的目的之一就是从这些模型中自动找出一个最好的模型，另外就是要根据业务对模型进行解释和应用。

**1.5、常用的数据挖掘建模工具**

数据挖掘是一个反复探索的过程，只有将数据挖掘工具提供的技术和实施经验与企业的 业务逻辑和需求紧密结合，并在实施过程中不断地磨合，才能取得好的效果。下面简单介绍 几种常用的数据挖掘建模工具。

* SAS Enterprise Miner，集成数据挖掘系统
* IBM SPSS Modeler，
* [SQL Server](https://cloud.tencent.com/product/sqlserver?from=10680)，
* Python，面向对象的解释型高级编程语言
* WEKA，知名度较高的机器徐文琪和数据挖掘软件
* KNIME，可以扩展使用WEKA中的挖掘算法
* RapidMiner，
* TipDM，数据挖掘建模平台

(1) SAS Enterprise Miner Enterprise Miner ( EM)是SAS推出的一个集成的数据挖掘系统，允许使用和比较不 同的技术，同时还集成了复杂的[数据库](https://cloud.tencent.com/solution/database?from=10680)管理软件。它的运行方式是通过在一个工作空间 (workspace)中按照一定的顺序添加各种可以实现不同功能的节点，然后对不同节点进行相 应的设置，最后运行整个工作流程(workflow),便可以得到相应的结果。

(2 ) IBM SPSS Modeler IBM SPSS Modeler原名Clementine, 2009年被IBM公司收购后对产品的性能和功能 进行了大幅度改进和提升。它封装了最先进的统计学和数据挖掘技术来获得预测知识，并 将相应的决策方案部署到现有的业务系统和业务过程中，从而提髙企业的效益。IBM SPSS Modeler拥有直观的操作界面、自动化的数据准备和成熟的预测分析模型，结合商业技术可 以快速建立预测性模型。

(3 ) SQL Server Microsoft 公司的 SQL Server 中集成了数据挖掘组件---- Analysis Servers,借助 SQL Server 的[数据库管理](https://cloud.tencent.com/product/dbbrain?from=10680)功能，可以无缝地集成在SQL Server数据库中。在SQL Server 2008中提供了 决策树算法、聚类分析算法、Naive Bayes算法、关联规则算法、时序算法、神经网络算法、 线性回归算法等9种常用的数据挖掘算法。但是，预测建模的实现是基于SQL Server平台 的，平台移植性相对较差。

(4) Python Python ( Matrix Laboratory,矩阵实验室)是美国Mathworks公司开发的应用软件，具 备强大的科学及工程计算能力，它不但具有以矩阵计算为基础的强大数学计算能力和分析功 能，而且还具有丰富的可视化图形表现功能和方便的程序设计能力。Python并不提供一个专 门的数据挖掘环境，但它提供非常多的相关算法的实现函数，是学习和开发数据挖掘算法的 很好选择。

(5 ) WEKA WEKA ( Waikato Environment for Knowledge Analysis)是一款知名度较高的开源机器学 习和数据挖掘软件。高级用户可以通过Java编程和命令行来调用其分析组件。同时，WEKA 也为普通用户提供了图形化界面，称为WEKA Knowledge Flow Environment和WEKA Explorer, 可以实现预处理、分类、聚类、关联规则、文本挖掘、可视化等。

(6 ) KNIME KNIME ( Konstanz InformationMiner, http://www.knime.org)是基于 Java 开发的，可以 扩展使用Weka中的挖掘算法。KNIME釆用类似数据流(data flow)的方式来建立分析挖掘 流程。挖掘流程由一系列功能节点组成，每个节点有输入/输出端口，用于接收数据或模型、 导出结果。

（7 ） RapidMiner RapidMiner 也称为 YALE （ Yet Another Learning Environment, https://rapidminer.com）,提 供图形化界面，釆用类似Windows资源管理器中的树状结构来组织分析组件，树上每个节点 表示不同的运算符（operator）o YALE中提供了大量的运算符，包括数据处理、变换、探索、 建模、评估等各个环节。YALE是用Java开发的，基于Weka来构建，可以调用Weka中的 各种分析组件。RapidMiner有拓展的套件Radoop,可以和Hadoop集成起来，在Hadoop集 群上运行任务。

（8） TipDM TipDM （顶尖数据挖掘平台）使用Java语言开发，能从各种数据源获取数据，建立多种 数据挖掘模型。TipDM目前已集成数十种预测算法和分析技术，基本覆盖了国外主流挖掘系 统支持的算法。TipDM支持数据挖掘流程所需的主要过程：数据探索（相关性分析、主成分 分析、周期性分析）；数据预处理（属性选择、特征提取、坏数据处理、空值处理）；预测建 模（参数设置、交叉验证、模型训练、模型验证、模型预测）；聚类分析、关联规则挖掘等一 系列功能。

Python 数据分析与挖掘实战（张良均）git相关： https://github.com/apachecn/python\_data\_analysis\_and\_mining\_action https://github.com/keefecn/python\_practice\_of\_data\_analysis\_and\_mining https://github.com/hfl15/MachineLearning/tree/master/book\_python\_practice\_of\_data\_analysis\_and\_mining#01 https://blog.csdn.net/besmarterbestronger/article/details/105409448

github搜索：https://github.com/search?o=desc&p=1&q=%E6%95%B0%E6%8D%AE%E6%8C%96%E6%8E%98&s=stars&type=Repositories

github 数据挖掘相关： https://github.com/luanshiyinyang/DataMining https://github.com/fighting41love/funNLP https://github.com/YouChouNoBB/data-mining-introduction https://github.com/julycoding/The-Art-Of-Programming-By-July

本文参与[腾讯云自媒体分享计划](https://cloud.tencent.com/developer/support-plan) ，欢迎热爱写作的你一起参与！

本文分享自作者个人站点/博客：https://blog.csdn.net/xc\_zhou/[复制](javascript:void(0);)

如有侵权，请联系 cloudcommunity@tencent.com 删除。

[编程算法](https://cloud.tencent.com/developer/tag/10663?entry=article)[GitHub](https://cloud.tencent.com/developer/tag/10284?entry=article)[https](https://cloud.tencent.com/developer/tag/10813?entry=article)[Git](https://cloud.tencent.com/developer/tag/10283?entry=article)[开源](https://cloud.tencent.com/developer/tag/10667?entry=article)

[举报](javascript:;)

点赞 2分享

[**登录**](javascript:;) 后参与评论

0 条评论

相关文章

* 《python数据分析与挖掘实战》笔记第5章

分类和预测是预测问题的两种主要类型，分类主要是预测分类标号（离散属性），而预测 主要是建立连续值函数模型，预测给定自变量对应的因变量的值。

[周小董](https://cloud.tencent.com/developer/user/4908043)

* 《python数据分析与挖掘实战》笔记第4章

数据预处理一方面是要提高数据的质量，另一方面是要让 数据更好地适应特定的挖掘技术或工具。统计发现，在数据挖掘的过程中，数据预处理工作量占到了整个过程的60%。

[周小董](https://cloud.tencent.com/developer/user/4908043)

* 《python数据分析与挖掘实战》笔记第2章

python有4个内建的数据结构–list（列表）、tuple（元组）、dictionary（字典）以及set（集合），它们可以统称为容器。

[周小董](https://cloud.tencent.com/developer/user/4908043)

* 《python数据分析与挖掘实战》笔记第3章

数据质量分析是数据挖掘中数据准备过程的重要一环，是数据预处理的前提，也是数据挖掘分析结论有效性和准确性的基础，没有可信的数据，数据挖掘构建的模型将是空中楼阁。

[周小董](https://cloud.tencent.com/developer/user/4908043)

* Python数据分析与实战挖掘

基础篇 书推荐：《用python做科学计算》 ? 扩展库 简介 Numpy数组支持，以及相应的高效处理函数 Scipy矩阵支持，以及相应的矩阵数值计算模块 Ma...

[企鹅号小编](https://cloud.tencent.com/developer/user/1008345)

* Python数据分析与挖掘实战.pdf

[学一学大数据](https://cloud.tencent.com/developer/user/1685822)

* 关于《Python数据挖掘入门与实战》读书笔记二（亲和性分析）

首先确保最新的版本吧，python最让人头痛和无解的就是版本兼容性问题，不过既然选择了就不要怕麻烦，不断的纠错不断的尝试，在频繁的错误中提高自己。

[python与大数据分析](https://cloud.tencent.com/developer/user/7190337)

* 关于《Python数据挖掘入门与实战》读书笔记七（主成分分析二）

主成分分析算法（Principal Component Analysis，PCA）的目的是找到能用较少信息描述数据集的特征组合。它意在发现彼此之间没有相关性、能...

[python与大数据分析](https://cloud.tencent.com/developer/user/7190337)

* 关于《Python数据挖掘入门与实战》读书笔记六（主成分分析一）

特征抽取是数据挖掘任务最为重要的一个环节，一般而言，它对最终结果的影响要高过数据挖掘算法本身。但怎样选取好的特征，还没有严格、快捷的规则可循，这也是数据挖掘科学...

[python与大数据分析](https://cloud.tencent.com/developer/user/7190337)

* Python业务分析实战｜共享单车数据挖掘

自行车共享系统是传统自行车租赁的新一代，从注册会员、租赁到归还的整个过程都是自动化的。通过这些系统，用户可以很容易地从一个特定的位置租用自行车，并在另一个位置归...

[数据STUDIO](https://cloud.tencent.com/developer/user/8756457)

* Python数据分析与挖掘实战 (大数据技术丛书) - 张良均 等著

10余位数据挖掘领域资深专家和科研人员，10余年大数据挖掘咨询与实施经验结晶。从数据挖掘的应用出发，以电力、航空、医疗、互联网、生产制造以及公共服务等行业真实案...

[用户3157710](https://cloud.tencent.com/developer/user/3157710)

* Python 爬取淘宝商品数据挖掘分析实战

? 作者 孙方辉 本文为CDA志愿者投稿作品，转载需授权 项目内容 本案例选择>> 商品类目：沙发； 数量：共100页 4400个商品； 筛选条件：天猫...

[CDA数据分析师](https://cloud.tencent.com/developer/user/1283928)

* 关于《Python数据挖掘入门与实战》读书笔记一（数据挖掘、机器学习、深度学习、人工智能）

Python数据挖掘、Python机器学习、Python深度学习的书籍买了不少本了，但真正读下来的却很少，为何？涉及高等数学，图论、离散数学等等，对我来说和天文...

[python与大数据分析](https://cloud.tencent.com/developer/user/7190337)

* 慕课网数据分析与挖掘实战（一）-数据获取

其中，urllib，urllib2，requests，scrapy主要用于数据抓取 xpath主要用于分析提取有价值的内容 phantomjs，beauti...

[Meet相识](https://cloud.tencent.com/developer/user/3090887)

* 《数据挖掘与数据化运营实战 思路、方法、技巧与应用》—— 读书笔记

[用户1154259](https://cloud.tencent.com/developer/user/1154259)

* 关于《Python数据挖掘入门与实战》读书笔记三（估计器和近邻算法）

Scikit-learn库，实现了一系列数据挖掘算法，提供通用编程接口、标准化的测试和调参工具，便于用户尝试不同算法对其进行充分测试和查找优参数值。

[python与大数据分析](https://cloud.tencent.com/developer/user/7190337)

[更多文章](https://cloud.tencent.com/developer/column)