面向程序员的数据挖掘指南

作者：Ron Zacharski

欢迎辞

这是一本用于学习基本数据挖掘知识的书籍。大部分关于数据挖掘的书籍都着重于讲解理论知识，难以理解，让人望而却步。不要误会，这些理论知识还是非常重要的。但如果你是一名程序员，想对数据挖掘做一些了解，一定会需要一本面向初学者的入门书籍。这就是撰写本书的初衷。

这本指南采用“边学边做”的方式编写，因此在阅读本书时，我强烈建议您动手实践每一章结束提供的练习题和实验题，使用书中的Python脚本将其运行起来。书中有一系列展示数据挖掘技术的实例，因此在阅读完本书后，你就能掌握这些技术了。这本书以Creative Commons协议发布，可以免费下载。你可以任意分发这本书的副本，或者重新组织它的内容。也许将来我会提供一本纸质的书籍，不过这里的在线版本永远是免费的。

目录

这本书以PDF格式免费发行，点击下面每一章的标题，您就会被定向到一个页面，其中包含了这一章的PDF文件和示例代码的下载链接。此外，你还可以在这个页面中发表评论，举出书中的问题和错误，哪部分难以理解等。我会根据这些评论来修订本书。

讲述什么是数据挖掘，它所能解决的问题的是什么，以及在阅读完本书后，你可以做些什么。

第二章：推荐系统入门

介绍协同过滤，基本的距离算法，包括曼哈顿距离、欧几里得距离、闵科夫斯基距离、皮尔森相关系数。使用Python实现一个基本的推荐算法。

第三章：隐式评价和基于物品的过滤算法

这章开始讨论可供选择的用户评价体系。用户能够显示地給于评价（好、差、五星评价等），或者隐式地給于评价——如果用户在亚马逊购买了一个MP3，我们则认为他是“喜欢”这件商品的。

第四章：分类

上一章中我们使用用户对商品的评价来进行推荐，这一章我们会使用商品本身的特性来进行推荐。这种算法在潘多拉等网站中采用。

第五章：进一步探索分类

本章会讨论如何评价分类器的效果，方法包括十折交叉验证、留一法、以及Kappa检验等，同时还会引入kNN算法。

第六章：朴素贝叶斯

我们会在这章探索朴素贝叶斯分类算法，使用概率密度函数来处理数值型数据。

第七章：朴素贝叶斯算法和非结构化文本

这一章我们会尝试使用朴素贝叶斯算法来对非结构化文本进行分类。我们是否能够判断出Twitter上的一片影评是正面评价还是负面的呢？

第八章：聚类

我们会讨论层次聚类和kmeans聚类。