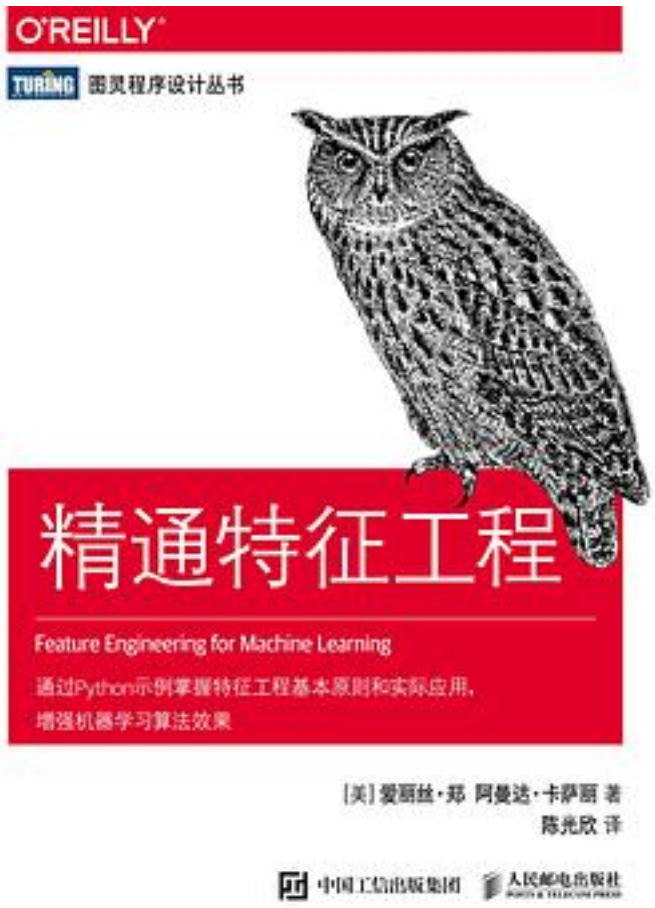
精通特征工程



精通特征工程_下载链接1_

著者:[美] 爱丽丝 · 郑

出版者:人民邮电出版社

出版时间:2019-4

装帧:平装

isbn:9787115509680

特征工程是机器学习流程中至关重要的一个环节，然而专门讨论这个话题的著作却寥寥无几。本书旨在填补这一空白，着重阐明特征工程的基本原则，介绍大量特征工程技术，教你从原始数据中提取出正确的特征并将其转换为适合机器学习模型的格式，从而轻

松构建模型，增强机器学习算法的效果。

然而，本书并非单纯地讲述特征工程的基本原则，而是通过大量示例和练习将重点放在了实际应用上。每一章都集中研究一个数据问题：如何表示文本数据或图像数据，如何为自动生成的特征降低维度，何时以及如何对特征进行标准化，等等。最后一章通过一个完整的例子演示了多种特征工程技术的实际应用。书中所有代码示例均是用Python编写的，涉及NumPy、Pandas、scikit-learn和Matplotlib等程序包。

- 数值型数据的特征工程：过滤、分箱、缩放、对数变换和指数变换

- 自然文本技术：词袋、n元词与短语检测

- 基于频率的过滤和特征缩放

- 分类变量编码技术：特征散列化与分箱计数

- 使用主成分分析的基于模型的特征工程

- 模型堆叠与k-均值特征化

- 图像特征提取：人工提取与深度学习

作者介绍:

爱丽丝·郑（Alice Zheng）

亚马逊广告平台建模和优化团队负责人，应用机器学习、生成算法和平台开发领域的技术领导者，前微软研究院机器学习研究员。

阿曼达·卡萨丽（Amanda Casari）

谷歌云开发者关系工程经理，曾是Concur Labs的产品经理和数据科学家，在数据科学、机器学习、复杂系统和机器人等多个领域都有丰富经验。

目录: 前言 ix
第1章 机器学习流程 1
1.1 数据 1
1.2 任务 1
1.3 模型 2
1.4 特征 3
1.5 模型评价 3
第2章 简单而又奇妙的数值 4
2.1 标量、向量和空间 5
2.2 处理计数 7
2.2.1 二值化 7
2.2.2 区间量化（分箱） 9
2.3 对数变换 13
2.3.1 对数变换实战 16
2.3.2 指数变换：对数变换的推广 19
2.4 特征缩放/ 归一化 24
2.4.1 min-max 缩放 24

2.4.2 特征标准化/ 方差缩放 24
2.4.3 ℓ2 归一化 25
2.5 交互特征 28
2.6 特征选择 30
2.7 小结 31
2.8 参考文献 32
第3章 文本数据：扁平化、过滤和分块 33
3.1 元素袋：将自然文本转换为扁平向量 34
3.1.1 词袋 34
3.1.2 n 元词袋 37
3.2 使用过滤获取清洁特征 39
3.2.1 停用词 39
3.2.2 基于频率的过滤 40
3.2.3 词干提取 42
3.3 意义的单位：从单词、n 元词到短语 43
3.3.1 解析与分词 43
3.3.2 通过搭配提取进行短语检测 44
3.4 小结 50
3.5 参考文献 51
第4章 特征缩放的效果：从词袋到tf-idf 52
4.1 tf-idf：词袋的一种简单扩展 52
4.2 tf-idf 方法测试 54
4.2.1 创建分类数据集 55
4.2.2 使用tf-idf 变换来缩放词袋 56
4.2.3 使用逻辑回归进行分类 57
4.2.4 使用正则化对逻辑回归进行调优 58
4.3 深入研究：发生了什么 62
4.4 小结 64
4.5 参考文献 64
第5章 分类变量：自动化时代的数据计数 65
5.1 分类变量的编码 66
5.1.1 one-hot 编码 66
5.1.2 虚拟编码 66
5.1.3 效果编码 69
5.1.4 各种分类变量编码的优缺点 70
5.2 处理大型分类变量 70
5.2.1 特征散列化 71
5.2.2 分箱计数 73
5.3 小结 79
5.4 参考文献 80
第6章 数据降维：使用PCA 挤压数据 82
6.1 直观理解 82
6.2 数学推导 84
6.2.1 线性投影 84
6.2.2 方差和经验方差 85
6.2.3 主成分：第一种表示形式 86
6.2.4 主成分：矩阵- 向量表示形式 86
6.2.5 主成分的通用解 86
6.2.6 特征转换 87
6.2.7 PCA 实现 87
6.3 PCA 实战 88
6.4 白化与ZCA 89
6.5 PCA 的局限性与注意事项 90
6.6 用例 91

6.7 小结 93
6.8 参考文献 93
第7章 非线性特征化与k-均值模型堆叠 94
7.1 k-均值聚类 95
7.2 使用聚类进行曲面拼接 97
7.3 用于分类问题的k-均值特征化 100
7.4 优点、缺点以及陷阱 105
7.5 小结 107
7.6 参考文献 107
第8章 自动特征生成：图像特征提取和深度学习 108
8.1 最简单的图像特征（以及它们因何失效） 109
8.2 人工特征提取：SIFT 和HOG 110
8.2.1 图像梯度 110
8.2.2 梯度方向直方图 113
8.2.3 SIFT 体系 116
8.3 通过深度神经网络学习图像特征 117
8.3.1 全连接层 117
8.3.2 卷积层 118
8.3.3 ReLU 变换 122
8.3.4 响应归一化层 123
8.3.5 池化层 124
8.3.6 AlexNet 的结构 124
8.4 小结 127
8.5 参考文献 128
第9章 回到特征：建立学术论文推荐器 129
9.1 基于项目的协同过滤 129
9.2 第一关：数据导入、清理和特征解析 130
9.3 第二关：更多特征工程和更智能的模型 136
9.4 第三关：更多特征= 更多信息 141
9.5 小结 144
9.6 参考文献 144
附录A 线性建模与线性代数基础 145
A.1 线性分类概述 145
A.2 矩阵的解析 147
A.2.1 从向量到子空间 148
A.2.2 奇异值分解（SVD） 150
A.2.3 数据矩阵的四个基本子空间 151
A.3 线性系统求解 153
A.4 参考文献 155
作者简介 156
封面简介 156
・・・・・・・(收起)

精通特征工程_下载链接1_

标签

机器学习

特征工程

Python

大数据

数据科学

计算机

数据挖掘

数据分析与机器学习

评论

还行

看一下开源版本

要吃透这本书的内容的前提是对线性代数的熟练掌握，因为这里面涉及到大量术语，虽然有讲解但还是很粗略。给出的代码很简洁实用，内容安排也比较合理。

---
https://github.com/apachecn/feature-engineering-for-ml-zh 这里粗看完了
就是还是肤浅地了解了个概念 大概是基础太差了 雁过不留痕

卧槽，才发现自己好久没看专业书籍了……

作为一个高数只学过数理统计的人，这本书看得太特么难受了，全是乱七八糟的名词，同一个概念，上下句间还要换种叫法，可以说很装逼了。给的代码集跟书上写的代码不是一路的，目前还没看出是干嘛用的。第二章欧式范数缩放的图非常有误导性且跟公式不搭配，当我们都已经会了吗？mix-max缩放的公式减号还丢了，差评。

---
feature engineering for
ml翻译成精通特征工程，真要从内容上看，翻成特征工程入门差不多，没多少新东西，也没多少实用的调参经验，看完还是像以前一样，一个个方法试错。另，像是写完没审直接出版了。

因为特征工程的书并不多，于是便入手了这本，我主要想看的是自然语言处理方面对于特征的处理。
看完之后很失望，讲的东西非常的少，而且很多都是老旧，很普遍的内容。
作为一本工具书，它对我的帮助实在是不大。

概括性的介绍了特征工程的一些方法，不够深入，而且专有名词很多，代码不错

作为单独介绍特征处理的书不是很多， 这本书还是不错的。
和模型构建相对紧密是最大的优点，给出了实例代码，不过没有提供直接数据下载，而且从数据网站上下载的数据往往和实例代码上的数据格式有冲突，无法直接边运行边学习。 扣一颗星

实战里更多是糙猛快，堆数据。书里不少方法和思路开阔眼界了，以后比赛里试试看

写得不是很浅显易懂，对实战提升较小

一般的特征工程都略知一二，这本书算是帮忙梳理了一边，完善了细节的感觉

感觉还可以,讲的听清楚的,如果看不懂,把线性代数复习一下,也就一半天

精通特征工程_下载链接1_

书评

我直言不讳，在我撰写本文的时候，本书在豆瓣评分偏低。不忍好书蒙尘，忍不住撰写此文。
工程领域的书籍不好写，实践性太强。工程中要处理的问题总是一个例子一个例子组成的，一个项目一个项目实操干出来的，具体例子和具体例子之间差异非常之大，方法论难以提炼。判断工程技术...

[https://github.com/apachecn/feature-engineering-for-ml-zh]
==========

特征工程是数据科学工程的核心，目前关于这个话题专门的书籍不多。本书通过概念（不是理论）和案例代码相结合的方式，还该了特征工程中的一些基础技术。包括分类型变量编码，数值型数据的分箱，变换。文本处理，PCA以及基于模型的特征工程。模型堆叠和k-均值特征化。最后简单介...

在图书馆看到的，感觉内容很棒，来豆瓣mark一下，上班有钱后买一本。吐个槽，书有点薄，59元略贵。虽然知识无价，不过对比国内出版物环境，嗯....相对有点点问题。
声明一下，这本书不是入门书籍，不适合机器学习入门/python入门的来看。
虽然英文名叫Feature Engineering for ...

精通特征工程_下载链接1_