《机器学习实战》阅读计划

——大数据与机器学习群阅读计划（第1期）

领读人：Allen Moore

本书特色

- **简洁**：更多地讨论如何编码实现机器学习算法，而尽量减少讨论数学理论
- **实用**：更多地讨论如何转化数学矩阵描述的机器学习算法可以实际工作的应用程序
- **趣味**：更多地讨论如何使用机器学习应用程序解决生活出现的有趣问题

适合读者：需要进行数据处理，想获取并理解数据

总阅读时间长度（预估）：5周

每天阅读用时：2小时

答疑时间安排：每周1次，每周二晚大数据与机器学习群 20:00—22:00

图灵社区本书网址：http://www.ituring.com.cn/book/1021

图灵阅读计划网址：https://github.com/BetterTuring/turingWeChatGroups

阅读建议

为了更好地阅读本书，读者需要：

- **一些编程概念**：流程控制（比如递归）、数据结构（比如树结构）
- **一些数学知识**：线性代数、概率论
- **一些Python经验**：Python编程知识

阅读规划

第一部分（第1~7章）分类

阅读时长：2周

重点内容 & 难点内容

1. 机器学习基础
    - 机器学习的主要任务
    - 如何选择合适的算法
    - 开发机器学习应用程序的步骤
2. k-近邻算法
    - 实施 kNN 算法
    - 准备数据：归一化数值
    - 准备数据：将图像转换为测试向量
3. 决策树
    - 决策树的构造
4. 基于概率论的分类方法：朴素贝叶斯
    - 基于贝叶斯决策理论的分类方法
    - 准备数据：从文本中构建词向量
5. Logistic 回归
    - 基于 Logistic 回归和 Sigmoid 函数的分类
6. 支持向量机
7. 利用 AdaBoost 元算法提高分类性能
    - 基于数据集多重抽样的分类器
    - bagging：基于数据随机重抽样的分类器构建方法

补充

- 这一部分主要探讨监督学习（supervised learning）：给定输入样本集，机器就可以从中推演出指定目标变量的可能结果。
- 监督学习一般使用两种类型的目标变量：标称型和数值型。

  标称型目标变量的结果只在有限目标集中取值，如真与假、车辆用途分类集合 { bus、van、suv、jeep } ；

  数值型目标变量则可以从无限的数值集合中取值，如 0.100、42.001、1000.743 等。

- 分类算法针对标称型目标变量，主要有最简单的k_近邻算法，比较直观、容易理解但是相对难于实现的决策树，
- 基于概率论的朴素贝叶斯，优化算法的 Logistic 回归，非常流行的支持向量机，元算法——AdaBoost。
- 读者需要注意分类算法之间的区别与联系，务必掌握各个分类算法的优势以及思考如何在实际项目之间进行取舍和搭配。
- 此外，读者也需要注数据预处理和数据之间的转化技巧，以备不时之需。

第二部分（第8~9章）利用回归预测数值型数据

阅读时长：1周

重点内容 & 难点内容

1. 预测数值型数据：回归
    - 用线性回归找到最佳拟合直线
    - 局部加权线性回归
    - 岭回归
    - lasso
    - 前向逐步回归
    - 权衡偏差与方差
2. 树回归
    - 复杂数据的局部性建模
    - 连续和离散型特征的树的构建
    - 将 CART 算法用于回归
    - 树剪枝

补充

- 这一部分仍然主要探讨监督学习（见第一部分 补充 内容）。
- 回归针对的其目标变量是连续数值型，主要有线性回归、局部加权线性回归和收缩方法以及树回归。
- 读者需要注意分类算法与回归算法的不同，需要了解不同回归算法的特点以及线性回归和树回归之间的本质差异。

第三部分（第10~12章）无监督学习

阅读时长：1周

重点内容 & 难点内容

1. 利用 K-均值聚类算法对未标注数据分组
    - K-均值聚类算法
    - 使用后处理来提高聚类性能
    - 二分 K-均值算法
2. 使用 Apriori 算法进行关联分析
    - 关联分析
    - Apriori 原理
    - 使用 Apriori 算法来发现频繁集
    - 从频繁项集中挖掘关联规则
3. 使用 FP-growth 算法来高效发现频繁项集
    - 构建 FP 树

补充

- 这一部分介绍的是无监督机器学习方法。

  不同于有监督学习，无监督学习中并不存在类似分类和回归中的目标变量。

  无监督学习只需要从算法程序中得到这些数据的共同特征，无需用户知道搜寻的目标对象。

  主要有K-均值聚类算法、关联分析的Apriori算法以及改进关联分析的FP_Growth算法。
- 读者需要注意无监督机器学习算法与有监督学习算法的差异，需要留意K-均值聚类算法和K-近邻分类算法之间的联系，同时关注FP_Growth算法对于原始Apriori算法的改进点。

第四部分（第13~15章）其他工具

阅读时长：1周

重点内容 & 难点内容

1. 利用 PCA 来简化数据
    - 降维技术
    - PCA
2. 利用 SVD 简化数据
    - SVD 的应用
3. 大数据与 MapReduce
    - MapReduce：分布式计算的框架
    - Hadoop 流
    - 在 Amazon 网络服务上运行 Hadoop程序
    - MapReduce 上的机器学习

补充

- 降维的目标就是对输入的数目进行削减，由此剔除数据中的噪声并提高机器学习方法的性能，主要有主成分分析降维方法（PCA）以及矩阵分解技术（SVD）。
- 大数据（big data）指的就是为了避免数据量过大时，内存不够而使用硬盘虚拟内存导致作业变慢的问题。
- 将整个作业进行分片用于在多机下进行并行处理的技术，这种技术的实现形式之一就是MapReduce。
- 读者需要注意两种降维技术PCA与SVD之间的差异，以及了解MapReduce在AWS上的部署流程。