

《大数据的机器学习方法》

段方

北京大学 博士后

1 背景

1.1 数据驱动还是应用驱动？

1.2 应用价值的显现化

1.3 对内服务和对外服务

1.4 建设应用的生态圈

2 大数据应用开发方法

2.1 **CRISP-DM (Cross-Industry Standard Process for Data Mining)** 方法论

2.2 业务理解 (**Business Understanding**)

2.3 数据理解 (**Data Understanding**)

2.4 数据准备 (**Data Preparation**)

2.5 建模 (**Modeling**)

2.6 评估 (**Evaluation**)

2.7 部署 (**Deployment**)

2.8 数据挖掘之外的方法论

2.9 【案例】 电信行业外来工识别及营销

3 大数据的互联网思维

3.1 互联网思维概述

3.2 用户思维

3.2.1 用户是谁

3.2.2 用户需要什么？

3.3 极致思维

3.3.1 产品极致设计

3.3.2 服务的极致

3.4 简约思维

3.4.1 产品的简约

3.4.2 简约的流程

3.5 迭代思维

3.5.1 产品的快速迭代

3.5.2 开发过程的快速迭代

3.6 流量思维

3.6.1 现有客户再有钱

3.6.2 扩展客户使用流量

3.7 平台思维

3.7.1 APP store 的借鉴

3.7.2 生态圈的构筑

3.8 跨界思维

3.8.1 数据的跨

界 3.8.2 应用

的跨界 3.9 数据

思维

3.9.1 收集数

据 3.9.2 分

析数据 3.10

社会化思维

3.10.1 影响社

会

3.10.2 社会角度思考

3.11 互联网思维在大数据系统中的落地

4 机器学习基础

4.1 多维分析方法

4.1.1 OLAP 分析

4.1.2 上钻和下钻

4.1.3 用 OLAP 分析问题

4.2 分析算法

4.2.1 回归算法

线性回归

逻辑回归

4.2.2 决策树算法

C4.5 算

法 **CART**

算法

4.2.3 贝叶斯算法

朴素贝叶斯算法

BBN (Bayesian Belief Network) 算法

4.2.4 基于核的算法

支持向量机 **SVM** 算法

线性判别分析(**Linear Discriminate Analysis** , **LDA**)

4.2.5 聚类算法

K-MEANS 算法

期望最大化算法(**Expectation Maximization** , **ME**)

4.2.6 关联规则算法

Apriori 算法

4.2.7 降低维度算法

主成份分析(**Principle Component Analysis** , **PCA**)算法

偏最小二乘回归(**Partial Least Square Regression** , **PLS**)算法

4.2.8 集成算法

随机森林算法

梯度推进机

4.3 【案例】 机器学习方法在电信行业应用案例汇编

5 大数据机器学习过程

5.1 需求的收集、整理

5.1.1 需求的收集方

5.1.2 需求的表述方法

5.2 数据的收集

5.2.1 能收集哪些数据

5.2.2 数据的维度

5.2.3 数据的质量

5.2.4 数据的管理

5.3 分析方法的选择

5.3.1 统计分析方法

5.3.2 OLAP 分析

5.3.3 数据挖掘方法

5.4 应用开发原型

5.4.1 原型系统的开发

5.4.2 原型系统的评估

5.5 应用的上线

5.5.1 谁负责使用

5.5.2 谁负责验收

5.6 快速迭代开发

5.6.1 如何快速迭代

5.6.2 迭代后的上线

5.7 上线后评估

5.7.1 客户满意度

5.7.2 效率提升

5.7.3 效益评估等

5.8 效益的评估方法

5.8.1 经济效益

5.8.2 社会效益

5.8.3 管理效益等

6 数据分析工具

6.1 Cognos

6.2 BRIO

6.3 SAS

6.4 SPSS

6.5 R 语言

7 大数据机器学习案例

7.1 【案例 1】 客户离网分析案例

7.2 【案例 2】 客户群体细分及营销

7.3 【案例 3】 4G 业务营销案例

7.4 【案例 4】 电信网络智慧运维案例

7.5 【案例 5】 人工智能技术在电信领域中应用案例

8 总结