慕课网学习笔记

机器学习——入门理论（2017-08-26）

第一节 课程介绍

什么是机器学习

生活中的机器学习

数据分析与机器学习

常用的机器学习算法

机器学习的框架

Demo联系机器学习

第二节 认识机器学习

1. 什么是机器学习

利用计算机从历史数据中找到规律，并将这些规律应用到未来不确定的场景决策中.

机器学习的主体是计算机而不是人.

数据分析方向:是人员驱动的,依赖数据分析师

机器学习方向:是计算机驱动的,依赖于大量的历史数据,数据越大,寻找的规律越可靠.

1. 从数据中寻找规律

概率论和统计学是机器学习的基石.

统计学:对大批量的数据进行抽样方法,使用部分的数据来代替整体的数据.而现在计算机的能力上升,不在需要进行抽样进行计算.

对数据进行函数拟合:仅仅限于二维.三维数据,对于更多维度的数据,则不适合可视化的拟合函数实现.

1. 机器学习的原动力

使用数据取代专家:使用数据来告诉我们逻辑,而不是专家

大数据驱动:经济驱动,数据变现.

1. 业务逻辑系统的历史

基于专家的经验

基于统计:分维度分析 (OLAP)

基于机器学习:

在线机器学习(搜索/电商) 考虑实时的数据流

离线机器学习

电商节日不适合离线机器学习,使用消费者当天的数据实时计算模型,不断的调整模型.

第三节

1. 机器学习的典型应用

啤酒和纸尿裤的例子:购物篮分析,捆绑销售,调整货架的摆放位置.利用关联规则对购物篮进行分析.

1. 用户细分和精准营销

移动分类: 全球通客户(机场、车站的VIP客户) 动感地带(在校学生) 神州行(白领、务工人员)

使用机器学习的聚类，将用户进行分类，按照要求把数据进行分类。

1. 机器学习的典型应用

朴素贝叶斯的应用模型：垃圾邮件

信用卡欺诈：决策树，风险识别

互联网广告：CTR预估，搜索引擎的推荐广告排序。点击率的预估概率。线性逻辑回归

推荐系统：协同过滤，电商推荐系统，一起购买。

自然语言处理：情感分析、实体分析。根据文字判断用户对于一件事件，一件商品的态度。

将文章中的人名、地名进行识别

深度学习：图像识别

语音识别

个性化医疗

情感分析

人脸识别

自动驾驶

智慧机器人

私人虚拟助理

手势识别

视频内容自动识别

机器实时翻译

Artificial Intelligence

第四节 机器学习与数据分析的区别

1. 数据特点

数据分析处理的是交易数据

机器学习处理的是行为数据

交易数据：电商的订单、电信运营商的计费数据

行为数据：用户搜索、点击，浏览历史，搜集用户的行为数据。

两者数据量不是一个级别的。

对于交易数据：一致性原则，对于数据的一致性要求非常高

对于行为数据：不要求特别的一致性。

NOSQL只能够处理行为数据，对数据一致性打了折扣，强调的数据的分布式。

交易数据：采样分析。

行为数据：全量分析。

1. 解决的业务问题不同

数据分析：报告过去的事情

机器学习：预测未来的事情

1. 技术手段不同

数据分析：用户驱动、交互分析，维度低，使用OLAP进行分析，是比较成熟的技术，落伍了

机器学习：数据驱动、自动进行知识发现，数据规模大，必须依赖计算机进行处理，维度大。

1. 参与者不同

数据分析：分析师决定数据的结果，数据不是那么的重要

机器学习：算法不是特别的重要，但是数据的质量决定结果。

1. 目标用户

数据分析：公司的高层

机器学习：针对的是个体用户

第五节 机器学习算法分类

1. 根据有无监督进行分类，有Y值得就是监督学习，无Y值得就是无监督学习。

有监督学习

分类算法

半监督学习

强化学习，可能会有一些Y值

无监督学习

聚类，不知道如何分类，分成几类。

1. 针对解决的问题

分类与回归

聚类

标注

1. 算法分类（回归算法的本质）

生成模型：判断数据属于某一类的概率，多种情况的判别。（陪审团）

判别模型：判断数据的类型、分类。（法官）

1. 常见机器学习的算法

C4.5算法：分类

K-Mean算法：聚类

SVM算法：分类问题

Apriori：关联分析，啤酒纸尿裤， 代价非常高，过时了

FP-growth：对数据库两次扫描

EM：统计学习，比较抽象

PageRank：google 搜索引擎

AdaBoost：集装与推进，人脸识别，决策树的改进版

KNN：分类算法

朴素贝叶斯：分类，垃圾邮件识别

CART算法：分类

逻辑回归：百度推荐

RF：随机森林

GBDT：决策树算法改进

推荐算法：

LDA：文本分析

World2Vector： 文本挖掘

HMM CRF：文本挖掘

深度学习：图像识别技术

第六节 机器学习解决问题

1. 机器学习解决问题的框架

确定目标

业务需求：明确自己的业务需求

数据：搜集数据

特征工程：数据预处理，数据清理（最重要），数据提取，数据的转换

1. 训练模型

定义模型 f(x)

定义损失函数 预测结果和计算结果之间的差异

优化算法 比较数学的思维

1. 模型评估

交叉验证

效果评估

Raw Data With Lable

Training Data Validation Data

Feature Engineering

Scoring model on traing data set

Measure Model Performance

Precision Recall ROC AUC

1. Demo Picture Classification

把图片分类：

首先将图片数据格式化，转换成为一个向量数据，然后将所有的向量数据扔进模型K-mean模型中进行计算。

机器学习实战 阅读笔记

1. 前言

数据挖掘十大算法

C4.5决策树

K-mean

SVM支持向量机

Apriori

EM 最大期望算法

PageRank算法

AdaBoost算法

KNN K近邻算法

NB（朴素贝叶斯算法）

CART（分类回归树算法）

第一部分 分类

主要探讨监督学习（supervised learning）。在监督学习中，我们只需要给定输入样本，机器学习就可以从中推断出制定目标变量可能的结果。监督学习相对比较简单，机器只需要从输入数据中预测合适的模型，从中计算出目标变量的结果。

监督学习一般使用的两种类型的目标变量：

标称型：类似枚举,一般用于分类

数值型：有无数个可能的数值，一般用于回归

Chapter 1 机器学习基础

1. 人脸识别、手写数字识别、垃圾邮件过滤、电商推荐系统等等都是机器学习的应用领域。
2. 什么是机器学习

机器学习横跨计算机科学、数学、工程技术、统计学、概率学等等多门学科。

未来物联网的时代，有海量的传感器对数据进行采集，并且通过网络进行汇总。未来如何在海量数据中获取有价值的信息成为关键。

1. 机器学习的主要任务

把数据进行分类

回归：用于预测数值型数据。最常见的例子就是：数据拟合曲线——给定数据点的最优拟合曲线。

分类和回归属于监督学习，因为我们知道自己预测的是什么，也就是目标变量的分类信息。

同时还有无监督学习，此时数据没有类别信息，也不会给定目标值。在无监督学习的情况下，将数据集合分成由类似的对象组成的多个类的过程称之为聚类。将寻找描述数据统计值的过程称为密度估计。此外无监督学习还可以减少数据的维度，使用二维、三维图形展示数据。

1. 如何选择机器学习的算法

根据目的选择：

想要预测目标变量的值——选择监督学习算法，否则选择无监督学习算法

监督学习中，数据类型是离散型数据，则使用分类算法；如果是连续型数据，则使用回归算法。

如果不想预测变量的值，则可以选择无监督学习，进一步分析是否将数据划分为离散的组。如果是，则使用聚类算法；如果需要估计数据与每一个分组的相似程度，则需要使用密度估计算法。

对于数据进行分析：特征值是离散还是连续的变量，特征值是否存在缺失的值，何种原因造成的数据缺失，数据中是否存在异常值，存在小概率的特征值？

我们可以在一定程度上缩小算法的选择范围，一般不存在最好的算法或者是给出最好的结果算法。每一种算法都可以使用其他的机器学习的算法技术进行优化。

1. 开发机器学习应用程序的步骤

搜集数据：爬虫 使用公开数据源

准备输入数据：将数据格式化为Python方便操作的数据，如list

分析输入数据：人工分析数据，是否数据存在一些异常，空值，可以实现数据的降维进行展示。系统产生的数据一般不需要尽心这样的处理。

训练算法：将数据输入进算法，从中抽取知识和信息，保存为计算机可以处理的形式。

测试算法：测试算法的有效性，对于不满意的算法需要重新进行上面的步骤。

使用算法：应用在实际中，转化为应用。

1. Python语言优势

Python语法清晰

Python容易操作文本文件

Python使用广泛，存在大量的开源工具和文档。

Python中的第三方库Scipy Numpy等许多科学函数库都实现了向量和矩阵的操作，这些函数增加了代码的可读性，并且底层使用的是C 和Fortran编写的，学习机器学习中大量使用Scipy和Numpy库。

Matplotlib绘图库可以绘制2D 3D 图形

Pylab模块将上面Scipy Numpy Matplotlib集成在一起。

Python唯一的不足就是效率低下，但是可以将其中的一些使用C来实现，来解决性能问题。

1. Numpy函数库基础

在Numpy中，数组和矩阵是不一样的，他们都可以处理行列表示的数据，看起来十分的相似，但实际是不同的。可以通过mat() 函数将数组转换成为矩阵。

randMat = mat(random.rand(4,4))

invRandMat = randMat.I # 求矩阵的逆操作

randMat\* invRandMat # 输出的结果不是单位矩阵，存在一定的计算误差。除了对角线，其他的数据都是很小的偏差。

myEye = randMat \* invRandMat

myEye - eye(4)

1. 本章小结

学习机器学习必须了解数据实例，每个数据实例由多个特征值构成。分类是基本的机器学习任务，它分析未分类数据，以确定如何将其放入到已知群组中。为了构建和训练分类器，必须首先输入大量已知分类的数据，我们称为训练样本。

机器学习在很多的领域应用，前景几乎是无限的。

Chapter 2 KNN K邻近算法

Chapter 3 决策树

Chapter 4 基于概率论的分类方法：朴素贝叶斯算法

Chapter 5 Logistics回归

Chapter 6 支持向量机 SVM

Chapter 7 利用AdaBoost元算法提高分类性能

第二部分 利用回归预测数值型数据

Chapter 8 预测数值型数据：回归

Chapter 9 树回归

第三部分 无监督学习

Chapter 10 利用K-mean均值聚类算法对未标注的数据分组

Chapter 11 使用Apriori算法进行关联分析

Chapter 12 使用FP-Growth算法高效发现频繁项集

第四部分 其他工具

Chapter 13 利用PCA来简化数据

Chapter 14 利用SVD简化数据

Chapter 15 大数据与MapReduce

Python入门知识

线性代数

概率论复习

资源