人工智能与数据挖掘导论

同学们，大家好！

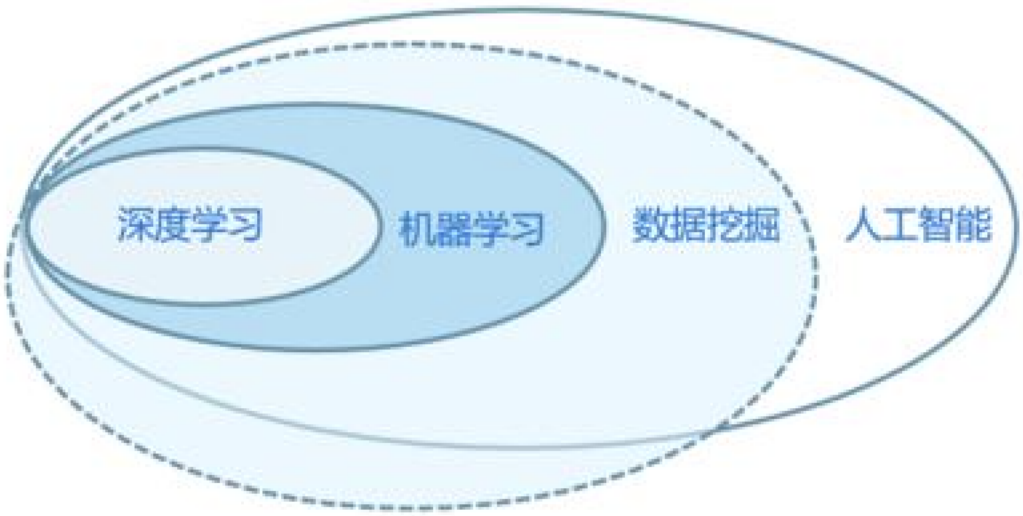
今天给大家分享的是人工智能与数据挖掘导论的相关内容。
从11个方面来介绍相关的内容：

- 人工智能的定义
- 人工智能和数据挖掘的关系
- 数据挖掘的定义
- 数据挖掘过程
- 相关理论算法
- 数据挖掘工具
- 相关代码实战
- 数据挖掘相关书籍
- 数据挖掘在各行业的应用
- 学习了数据挖掘可以从事的岗位
- 小结

人工智能的定义

人工智能是一门学科，指由人制造出来的机器所表现出来的智能。这种智能的最理想状态是像人一样拥有学习、推理等能力。简单来说，是指可模仿人类智能来执行任务，并基于收集的信息对自身进行迭代式改进的系统和机器。

人工智能和数据挖掘的关系



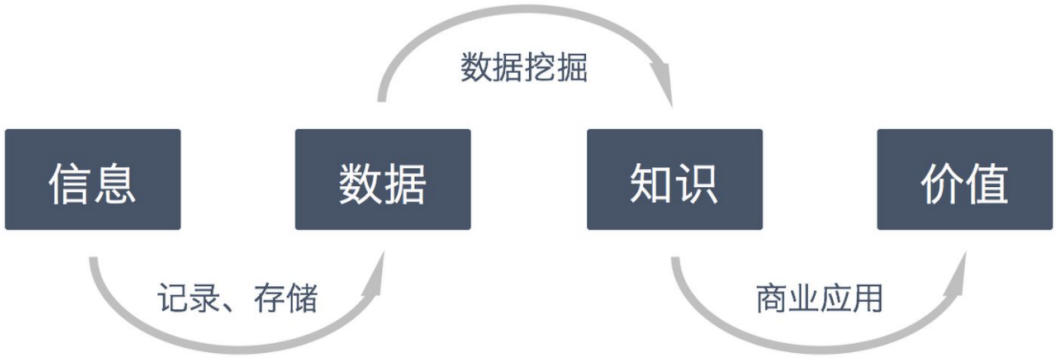
数据挖掘的定义

数据挖掘（Data Mining）是指通过大量数据集进行分类的自动化过程，以通过数据分析来识别趋势和模式，建立关系来解决业务问题。换句话说，数据挖掘是从大量的、不完全的、有噪

声的、模糊的、随机的数据中提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。

如今，大数据就像是一座潜力无穷的金矿，最核心的价值需要通过挖掘分析才能体现。各行各业的互联网化，让数据得到更广泛的应用。

而从事数据分析、数据挖掘、数据处理的岗位的人才其实相当稀缺， 如今掌握数据挖掘思维和技能，将会为你的升职加薪打牢基础。



通常我们把信息转化为价值，要经历信息、数据、知识、价值四个层面，而数据挖掘就是中间的重要环节，是从数据中发现知识的过程。

数据挖掘过程

从形式上来说，数据挖掘的开发流程是迭代式的。开发人员通过如下几个阶段对数据进行迭代式处理：

- 解读需求：我们需要完成什么样的目标?
- 搜集数据：我们需要从哪里获得数据?
- 预处理数据：如何处理出完整、干净的数据?
- 选择算法：该如何选择一个适合我需求的算法?
- 评估模型：如何确认我们的模型已经达标?
- 解释模型：我们的模型是否可以解决业务需求?

1.解读需求

绝大多数的数据挖掘工程都是针对具体领域的，因此数据挖掘工作人员不应该沉浸在自己的算法模型世界里，而应该多和具体领域的专家交流合作以正确的解读出项目需求，且这种合作应当贯穿整个项目生命周期。

2.搜集数据

在大型公司，数据搜集大都是从其他业务系统数据库提取。很多时候我们是对数据进行抽样，在这种情况下必须理解数据的抽样过程是如何影响取样分布，以确保评估模型环节中用于训练（train）和检验（test）模型的数据来自同一个分布。

3.预处理数据

预处理数据可主要分为数据准备和数据归约两部分。其中前者包含了缺失值处理、异常值处理、归一化、平整化、时间序列加权等；而后者主要包含维度归约、值归约、以及案例归约。

4.选择算法

算法分为很多类型，包括分类、聚类、回归算法等，要根据数据的格式和类型、数据的特点等，以及业务的场景和目标等选择合适的算法，以解决业务目标为项目目标，不断的调整算法的细节和过程，使算法能不断的适应现有的算法和场景。

5.评估模型

这一步就是在不同的模型之间做出选择，找到最优模型。很多人认为这一步是数据挖掘的全部，但显然这是以偏概全的，甚至绝大多数情况下这一步耗费的时间和精力在整个流程里是最少的。

6.解释模型

数据挖掘模型在大多数情况下是用来辅助决策的，人们显然不会根据“黑箱模型”来制定决策。如何针对具体环境对模型做出合理解释也是一项非常重要的任务。

相关理论算法

数据挖掘算法主要包括：分类、聚类、回归、关联分析等算法，以下将从宏观上描述这些算法。

1、分类问题

KNN 算法：对若干种水果进行分类
决策树：根据天气情况决定去哪里玩什么
朴素贝叶斯：根据人名来预测其性别
支持向量机（SVM）：用一条线分开红豆与绿豆
人工神经网络：当前最火热的深度学习模型的基础

2、聚类问题

k-means 聚类：擒贼先擒王，找到中心点，它附近的都是一类
DBScan 聚类：打破形状的限制，使用密度聚类

3、回归问题

线性回归与逻辑回归：找到一个函数去拟合数据

4、关联分析

Apriori 与 FP-Growth：最经典的就是啤酒与尿布的故事

数据挖掘工具

Python科学计算工具包：Numpy、Pandas
Python画图相关工具包：Matplotlib
机器学习相关算法库：Sklearn
Python代码编译器：Pycharm、Jupyter Notebook

相关代码实战

https://www.cnblogs.com/bigbigbird/p/12960460.html

数据挖掘相关书籍

Jiawei Han的《数据挖掘概念与技术》
Ian H. Witten / Eibe Frank的《数据挖掘实用机器学习技术》
Tom Mitchell的《机器学习》
TOBY SEGARAN的《集体智慧编程》
Anand Rajaraman的《大数据》
Pang-Ning Tan的《数据挖掘导论》
Matthew A. Russell的《社交网站的数据挖掘与分析》
《数据挖掘概念与技术》
《机器学习实战》

数据挖掘在各行业的应用

零售商可以部署数据挖掘，以更好地识别人们根据过去的购买习惯可能购买哪个产品，或者哪些商品在一年的某些时间可能热卖。这可以帮助商家规划库存和存储布局，同时也可以利用数据挖掘来做线下零售店铺的智能选址。

银行和其它金融服务提供商可以挖掘与其客户帐户、交易和渠道偏好相关的数据，以更好地满足他们的需求。它们还可以从他们的网站和社交媒体互动中分析数据，以增加现有客户的忠诚度并吸引新客户。

制造企业可以使用数据挖掘在生产过程中发现模式，从而可以精确地识别出瓶颈和有缺陷的方法，并设法提高效率。它们还可以将知识从数据挖掘应用于产品设计，并根据客户体验的反馈进行调整。

教育机构可以从数据挖掘中受益，例如分析数据集，以预测学生的未来学习行为和表现，然后利用这些知识来改进教学方法或课程。

医疗保健提供者可以挖掘和分析数据，以确定向患者提供护理和降低成本的更好的方法。在数据挖掘的帮助下，他们可以预测需要照顾的病人数量以及患者需要什么类型的服务。在生命科学领域，数据挖掘可用于从大量生物数据中获取洞察，帮助开发新药和其他治疗方法。

在包括医疗保健和零售在内的多个行业，你可以使用数据挖掘来检测诈骗和其它滥用行为——比传统的识别此类活动的方法要快得多。

学习了数据挖掘可以从事的岗位

数据挖掘算法工程师
机器学习算法工程师
推荐系统算法工程师
数据分析工程师

小结

在数据挖掘中，准备数据的初始行为（例如聚集然后使数据合理化）可以揭示可能危及数据机密性的信息或模式。因此，不经意地违反道德问题或法律要求是有可能的。因此数据挖掘的每一步还需要数据保护，以确保数据不被偷窃、改变或秘密访问。安全工具包括加密、访问控制和网络安全机制。

尽管存在这些挑战，但数据挖掘已成为很多组织IT战略的重要组成部分，这些组织力图通过收集或访问的所有信息获得价值。随着预测分析、人工智能、机器学习和其它相关技术的不断进步，这一驱动力无疑将加速。