第二章 数据获取

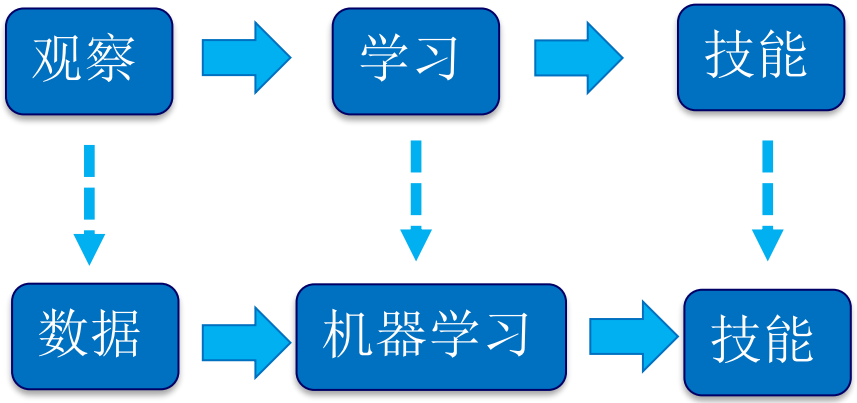
内容提要

- 背景
- 数据采集
- 数据标注
- 改进现有数据和模型

收集数据的理由

- 新的应用程序(数据挖掘、机器学习)不一定有足够的标签数据。
  --传统的机器翻译和对象检测：需要数十年积累的海量训练数据。
  --新的应用：缺少可用训练数据，需要人工标注(昂贵和领域专家)

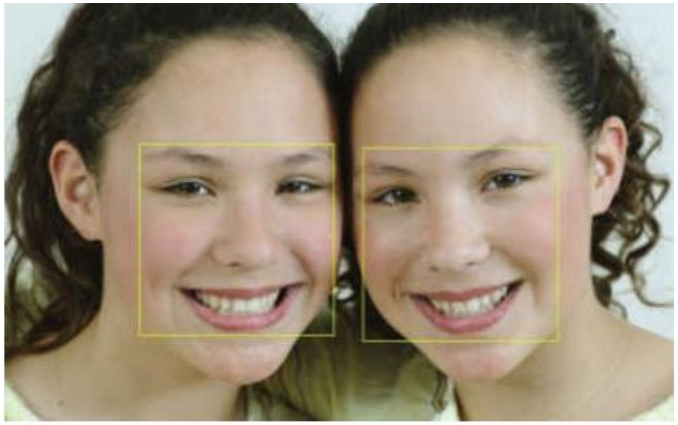
- 与传统的机器学习不同，深度学习技术会自动生成特征，这节省了特征工程的成本，但反过来可能需要大量的标注数据。



相关的应用

- 机器学习(ML)、自然语言处理(NLP)和计算机视觉(CV)

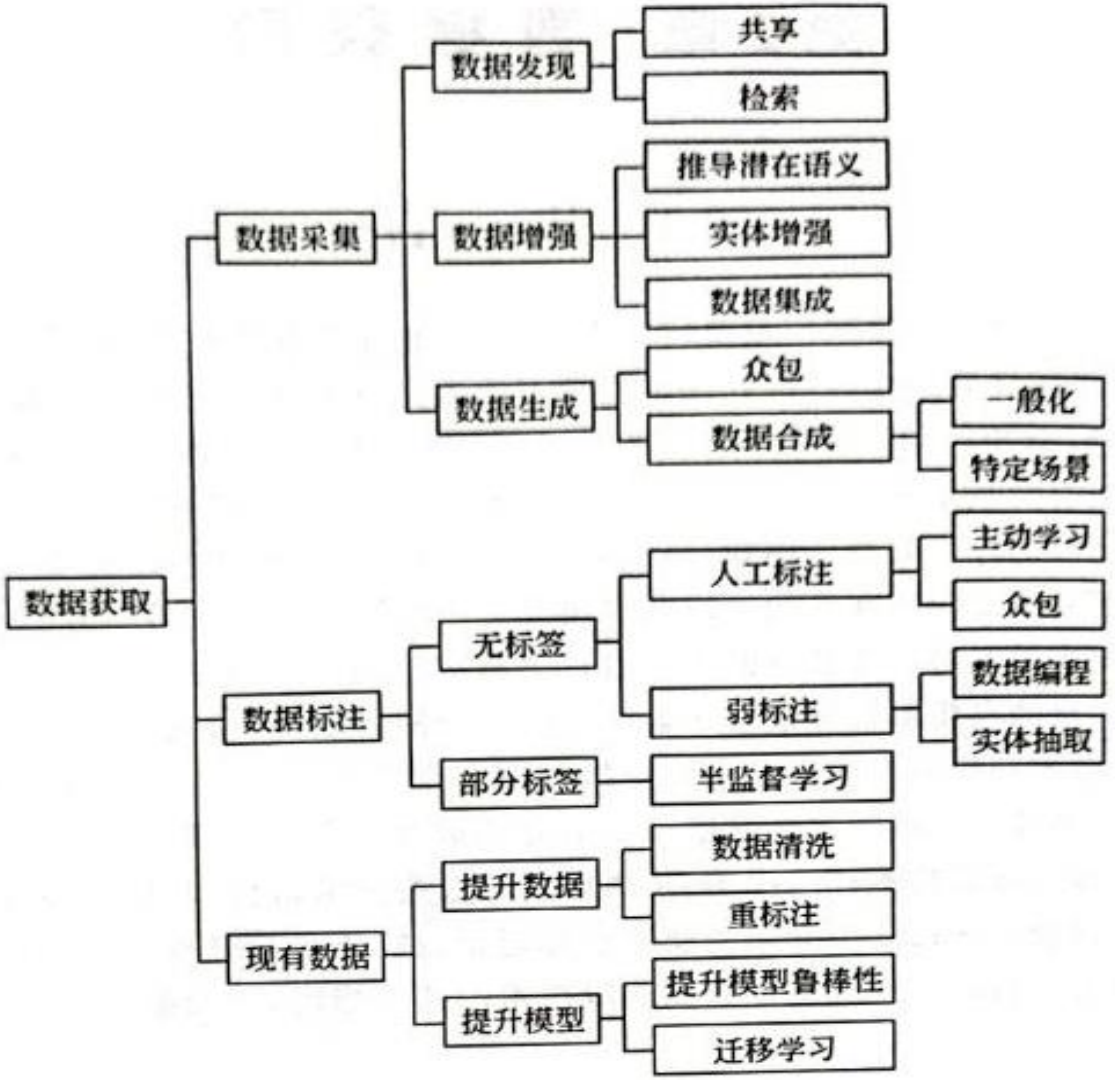
端到端(端到端)机器学习应用程序：收集、清洗、分析、可视化和特性工程



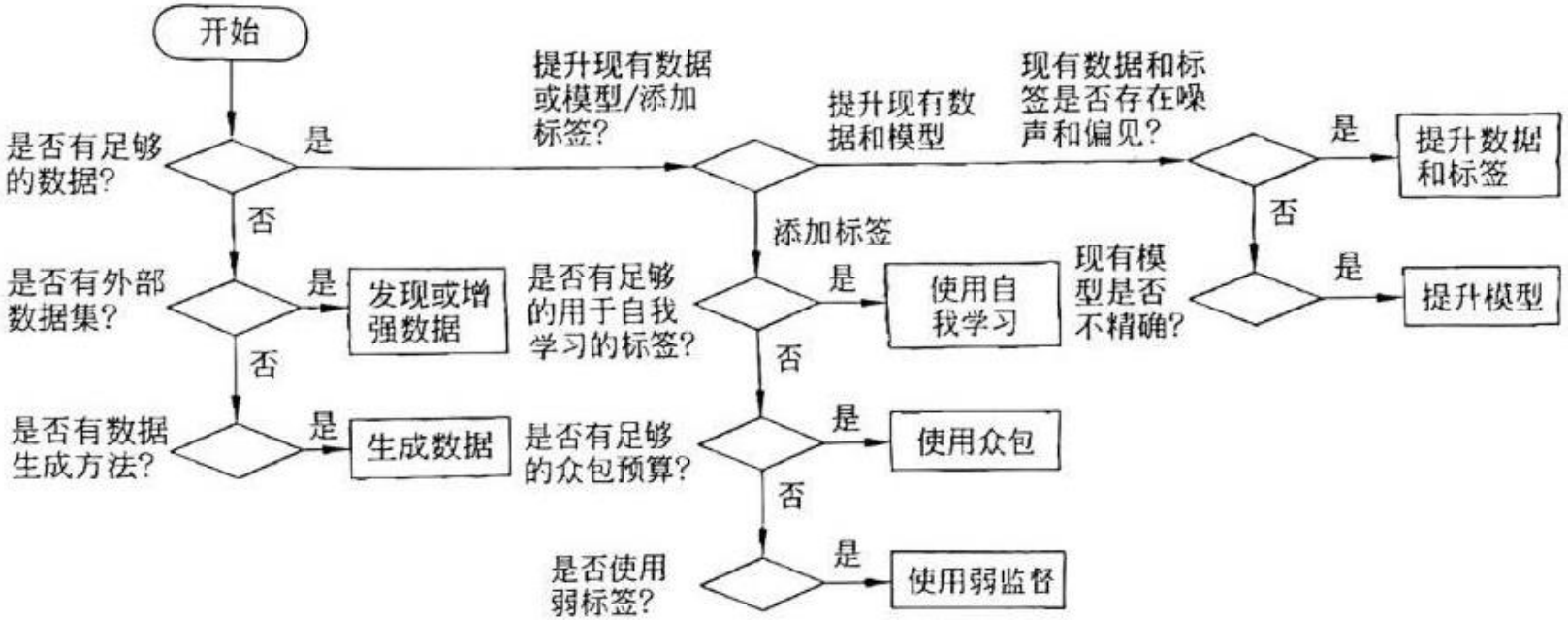
背景

- **大数据时代迫切需要精确和可伸缩的数据收集技术**
  - 从数据管理角度，数据获取有三种方式：
  - 共享和搜索新数据集：数据采集技术可用于发现、扩充或生成数据集
  - 一旦数据集可用，就可以使用各种数据标记技术来标记单个示例
  - 与其标记新的数据集，不如改进现有的数据集或在经过培训的模型之上进行培训。

- 数据获取的三种方式


数据获取
数据采集
数据发现
共享
检索
数据增强
推导潜在语义
实体增强
数据集成
数据生成
众包
数据合成
一般化
特定场景
数据标注
无标签
人工标注
主动学习
众包
弱标注
数据编程
实体抽取
部分标签
半监督学习
现有数据
提升数据
数据清洗
重标注
提升模型
提升模型鲁棒性
迁移学习

- 数据获取流程


开始
是否有足够的数据?
是
否
是否有外部数据集?
是
发现或增强数据
否
是否有数据生成方法?
是
生成数据
提升现有数据或模型/添加标签?
提升现有数据和模型
现有数据和标签是否存在噪声和偏见?
是
提升数据和标签
否
添加标签
现有模型是否不精确?
是
提升模型
是否有足够的用于自我学习的标签?
是
使用自我学习
否
是否有足够的众包预算?
是
使用众包
否
是否使用弱标签?
是
使用弱监督

数据采集

- 数据采集的目标：找到可用于挖掘数据的数据集
- 有三种方法：数据发现、数据增强和数据生成。
  - 当一个人想要共享或搜索新的数据集时，数据发现是必要的。随着更多的数据集在Web和公司数据湖上可用，数据发现变得非常重要。
  - 数据增强是对现有数据集通过添加更多外部数据来增强的数据发现的补充。
  - 当没有可用的外部数据集时，可以使用数据生成，可以采用众包或合成生成数据集。

数据采集-数据发现

- 数据发现的两个步骤：

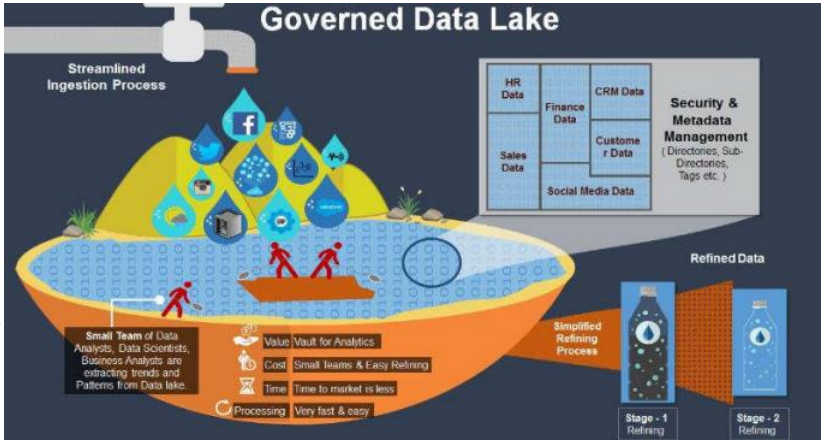
  - 数据共享：必须对生成的数据进行索引并发布以供共享。（自组织方法）

    技术：协同分析、基于Web的技术、协同分析和Web结合

  - 数据搜索：根据给定的数据挖掘任务搜索对应的数据集，关键在于如何扩展搜索以及判断数据集是否适合给定的挖掘任务。

    技术：数据湖、基于Web的技术

- **数据湖：**
  - 数据仓库：瓶装水的商店——经过清洁、包装和结构化以便于饮用
  - 数据湖：自然状态的一大片水体。数据湖的内容从源头流入，填满湖，湖的各种用户可以来检查、潜入或取样。
  - 数据湖（Data Lake）是一个以原始格式存储数据的存储库或系统，它按原样存储数据，而无需事先对数据进行结构化处理。一个数据湖可以存储结构化数据（如关系型数据库中的表），半结构化数据（如CSV、日志、XML、JSON），非结构化数据（如电子邮件、文档、PDF）和二进制数据（如图形、音频、视频）。



- **数据增强：获取数据的另一种方法是利用外部数据扩大现有数据集。**
  - 让有限的数据产生更多的数据，增加训练样本的数量以及多样性（噪声数据）
  - 以对图像进行不同方式的裁剪，让物体以不同的实例出现在图像的不同位置
  - 从数据中导出潜在语义：

    词向量模型+隐含主题模型

    NLP中将字和词连接起来就形成了一个新样本，也属于数据增强。

数据采集-数据生成

- **数据生成：当没有可用于训练的现有数据集时，需要从头构建新数据集。**
  - 手工构建：众包

    指的是一个公司或机构把过去由员工执行的工作任务，以自由自愿的形式外包给非特定的（而且通常是大型的）大众志愿者的做法。
  - 自动化技术：数据合成

数据采集-数据生成

- **众包技术可以分为两步：收集数据和预处理数据**
  - 收集数据：

    在众包平台发布具体任务，招募大众志愿者完成任务，收集到足够的数据
  - 预处理数据：

    对数据进行预处理，实体解析、连接数据集等，使其能用于对应的机器学习任务
  - 需要关注的问题：

    数据质量

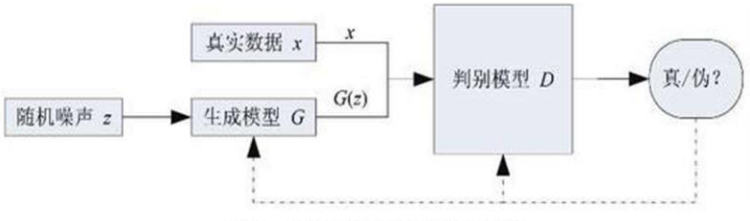
Amazon Mechanical Turk

- **自动数据生成：低成本和灵活性**
  - 生成对抗网络（GANs）

    基于真实数据生成合成数据

    --如MEDGAN基于真实患者记录

    生成合成患者记录



  - 面向特定应用程序的自动化生成技术

    --如合成图像的自动生成，合成文本的自动生成

- **数据标注**
  - 对未经处理的初级数包括语音、图片、文本、视频等进行加工处理并转换为机器可识别信息的过程。
- **数据标注的目的在于标记单个示例**
- **数据标注主要有三类方法：**
  - 使用现有标签
  - —利用已经存在的标签
  - 众包技术
  - —使用众包技术标记
  - 弱监督学习
  - —在较低成本的条件下，生成不太完美的标签

- 利用现有标签
  - 分类算法

—利用已有的标记数据训练分类模型，该分类模型应用于所有未曾标记的数据

  - 回归算法

—使用标记数据训练分类模型，为每个示例返回对应的实数（属于某个类型的概率）。

  - 基于图标签传播的算法

—从有限的有标记的例子集合开始的，但是利用基于它们的相似点的例子的图结构来推断剩下的例子的标签。

- **基于众包技术**
  - 手动标注每一条实例
  - 大量工人统共同完成，如何确保质量？
  - 开销巨大，如何降低成本？
  - 主动学习：主动学习的重点是选择最“有趣”的未标记示例给工人进行标记。如何选择？

Fig. 1. Three crowd workers are hired to categorize the breed of a dog as Scotch, Yorkshire, or Australian. They express their single-option and Cumulative crowd labels using checked boxes and confidence bars, respectively. While the average score of Cumulative labels correctly indicates higher chance for Australian, the majority of single-option labels incorrectly suggests Yorkshire as the truth.

- 主动学习技术：

1）不确定采样—最简单的一种方法，它选择了模型预测中最不确定的做为下一个样本。

2）决策理论—使用目标函数来判断需要标注的示例。

3）回归主动学习—回归问题，比如选择方差最大的示例

4）自我学习和主动学习集合--自我学习找到具有最高置信度的预测，并将它们添加到Labeled example中，同时主动学习找到具有最低置信度的预测(使用不确定抽样、逐委员会查询、顺序加权方法)，并将它们发送给人工标记。

- **弱监督学习**
  - 有大量数据并且手动标记成本不可接受
  - 半自动生成大量标签
  - 数据编程，被认为是一种用多个标注函数代替单个标注函数来生成大量弱标签的解决方案
  - 事实提取，另一种生成弱标签的方法，知识库包含从各种来源(包括Web)中提取的事实。事实可以描述一个实体的属性(例如，德国，首都，柏林)。

FREEBASE

YAGO

- **提升已有数据和模型**
  - 提升已有数据

    数据清理提升数据质量

    重复再标记

  - 提升已有模型

    模型对噪声和偏差的鲁棒（强壮）性：对噪声建模、分辨对抗样本

    迁移学习：通过现有模型，训练新模型

Thank you !!!