主动学习中文材料

资料

- https://github.com/baifanxxx/awesome-active-learning

思路

- 通过机器学习的方法获取到那些比较**“难”**分类的样本数据，让人工再次确认和审核，然后将人工标注得到的数据再次使用有监督学习模型或者半监督学习模型进行训练，逐步提升模型的效果，将人工经验融入机器学习的模型中
    - 这里所谓的**难**，需要根据不同的情况去定义

流程

- 机器学习模型：包括机器学习模型的训练和预测两部分
- 待标注的数据候选集提取：依赖主动学习中的查询函数（Query Function）
    - 这里就是**难**的具体定义
    - 主动学习的核心
        - 不确定性采样的查询（**Uncertainty Sampling**）
            - 置信度最低
            - 边缘采样
            - 熵方法
        - 基于委员会的查询（**Query-By-Committee**）
            - 投票熵
            - 平均 KL 散度
        - 基于模型变化期望的查询（**Expected Model Change**）
            - 基于误差减少的查询（**Expected Error Reduction**）
            - 基于方差减少的查询（**Variance Reduction**）
        - 基于密度权重的查询（**Density-Weighted Methods**）
- 人工标注：专家经验或者业务经验的提炼
- 获得候选集的标注数据：获得更有价值的样本数据
- 机器学习模型的更新：通过增量学习或者重新学习的方式更新模型，从而将人工标注的数据融入机器学习模型中，提升模型效果

分类

- 流式主动学习【Stream-based Active Learning】
- 批式主动学习【Pool-based Active Learning】