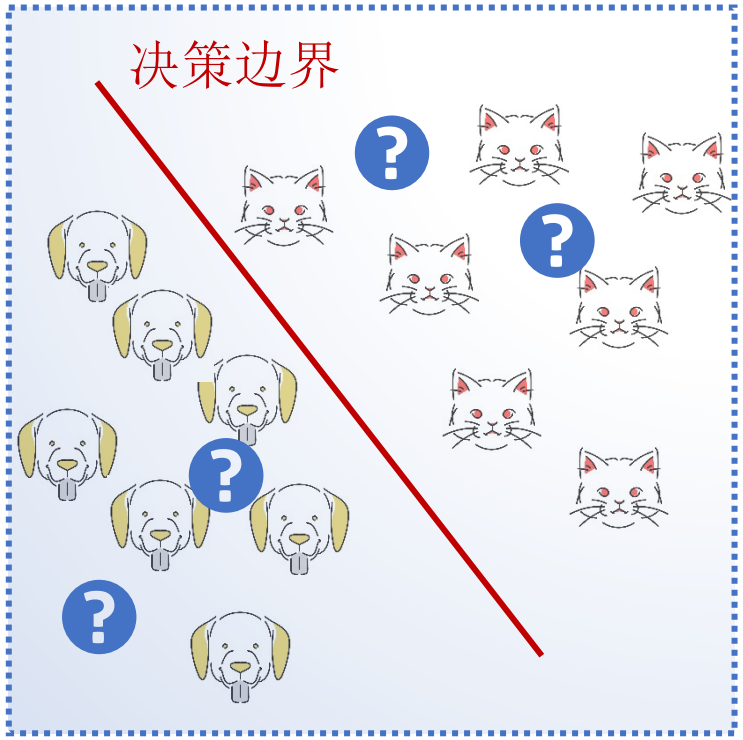
弱监督文本分类

陆俞因

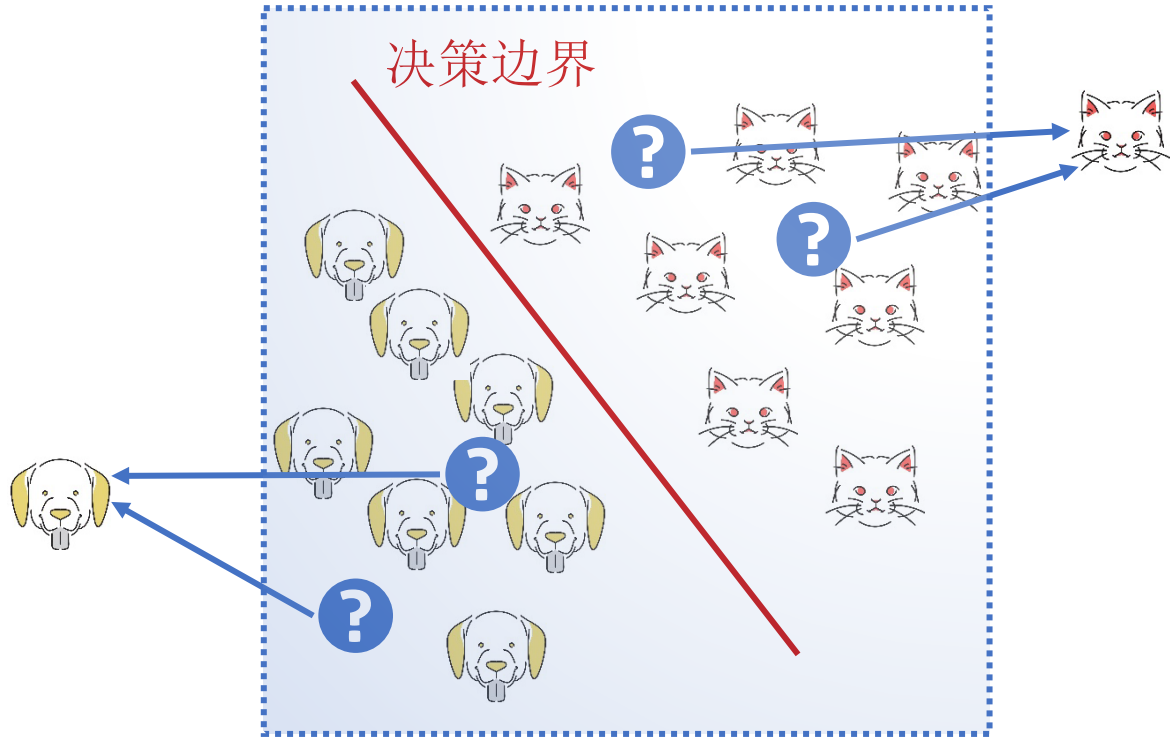
2023.05.16

Supervised Classification
有监督分类

➢训练



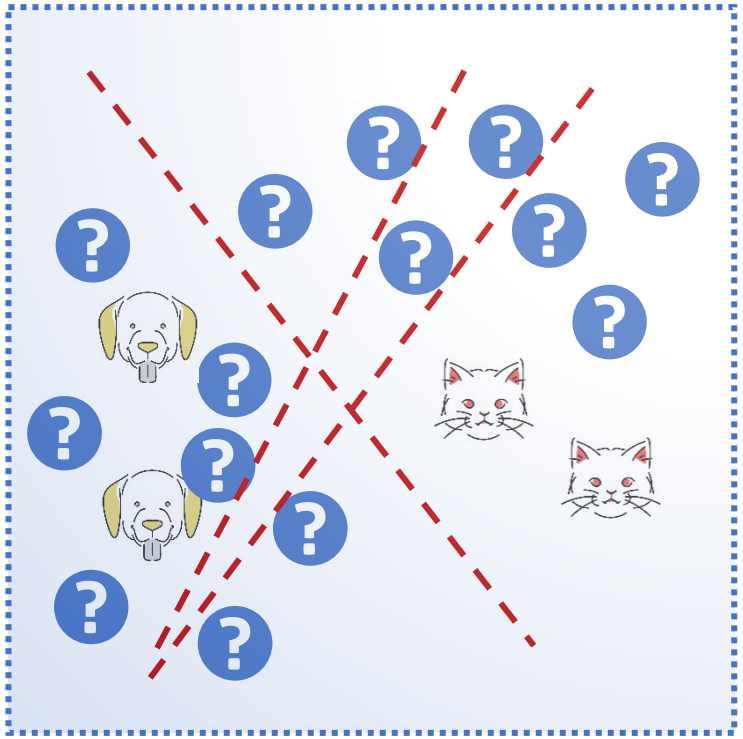
➢预测



实际的情况……

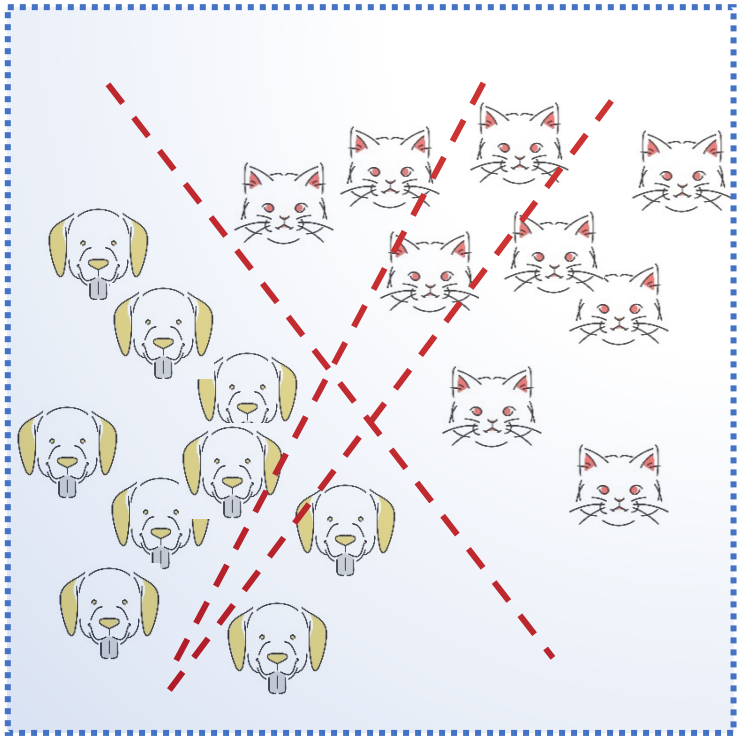
- 缺少足够的标记数据

难以学习到合理的决策边界



- 预测

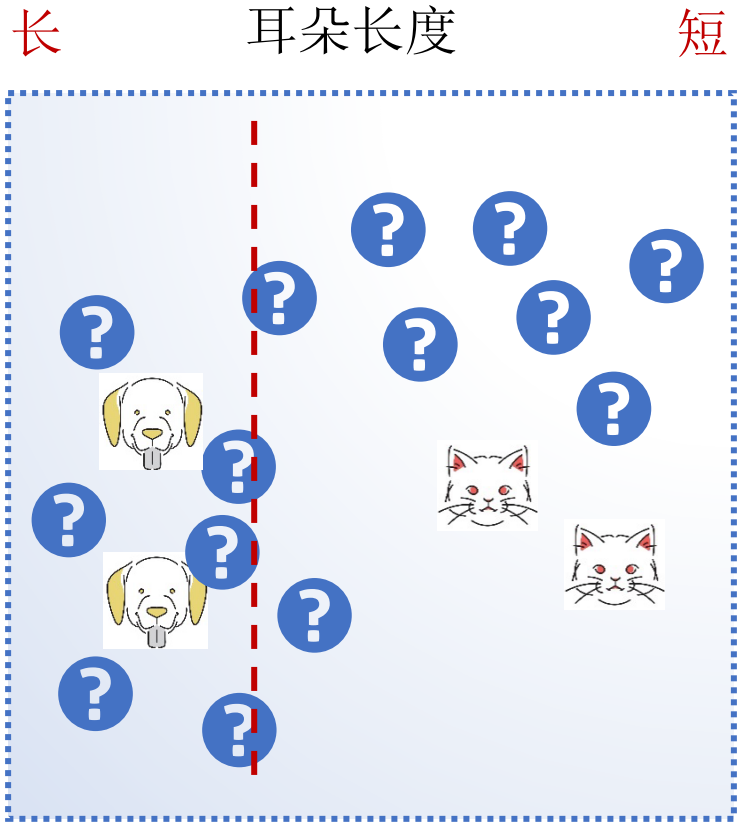
准确率低

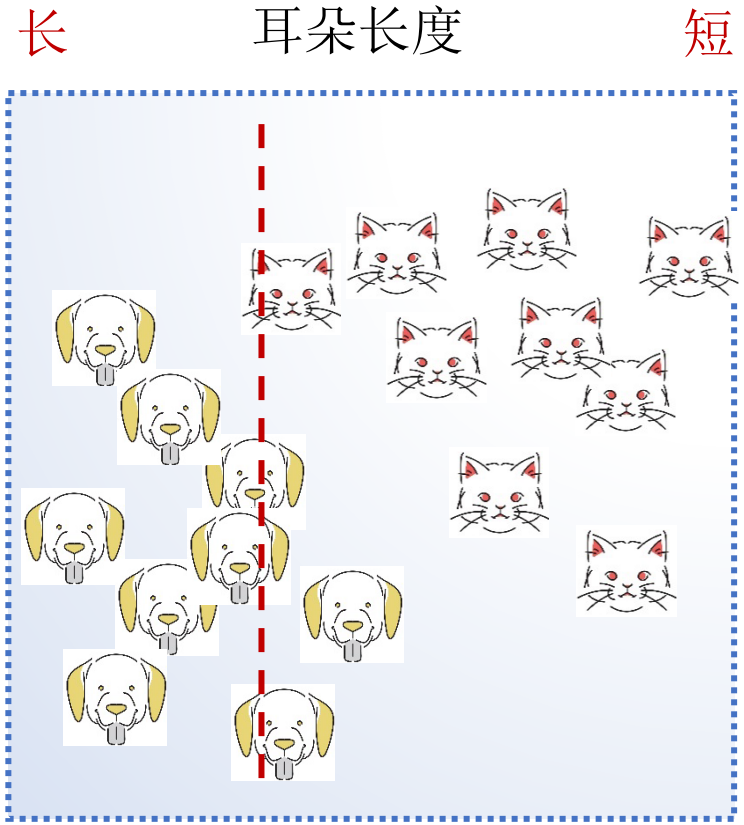


Weakly-Supervised Classification
弱监督分类

➢利用弱监督信息进行分类，尽可能提升分类准确率

- 少量有标记样本
- 关于各个类的先验知识
- ……





Self-Training

自训练算法

➢一种经典且有效的弱监督分类方法

① 根据弱监督信息（少量标记样本或类先验知识）初始化分类器 C

② 通过分类器求得各个无标记样本 i 属于每个类 k 的概率

$$[p(i \in 1), \ldots, p(i \in k), \ldots p(i \in K)] = C(i)$$

③ 取 M 个分类信心最强的样本作为训练集

④ 在训练集上更新分类器 C

⑤ 重复步骤②~④，直至终止条件

Self-Training

自训练算法

➢终止条件

① 达到最大迭代次数

② 所有样本的分类信心都超过一个最低阈值

③ 更新分类器后，分类标签改变的样本少于一个阈值

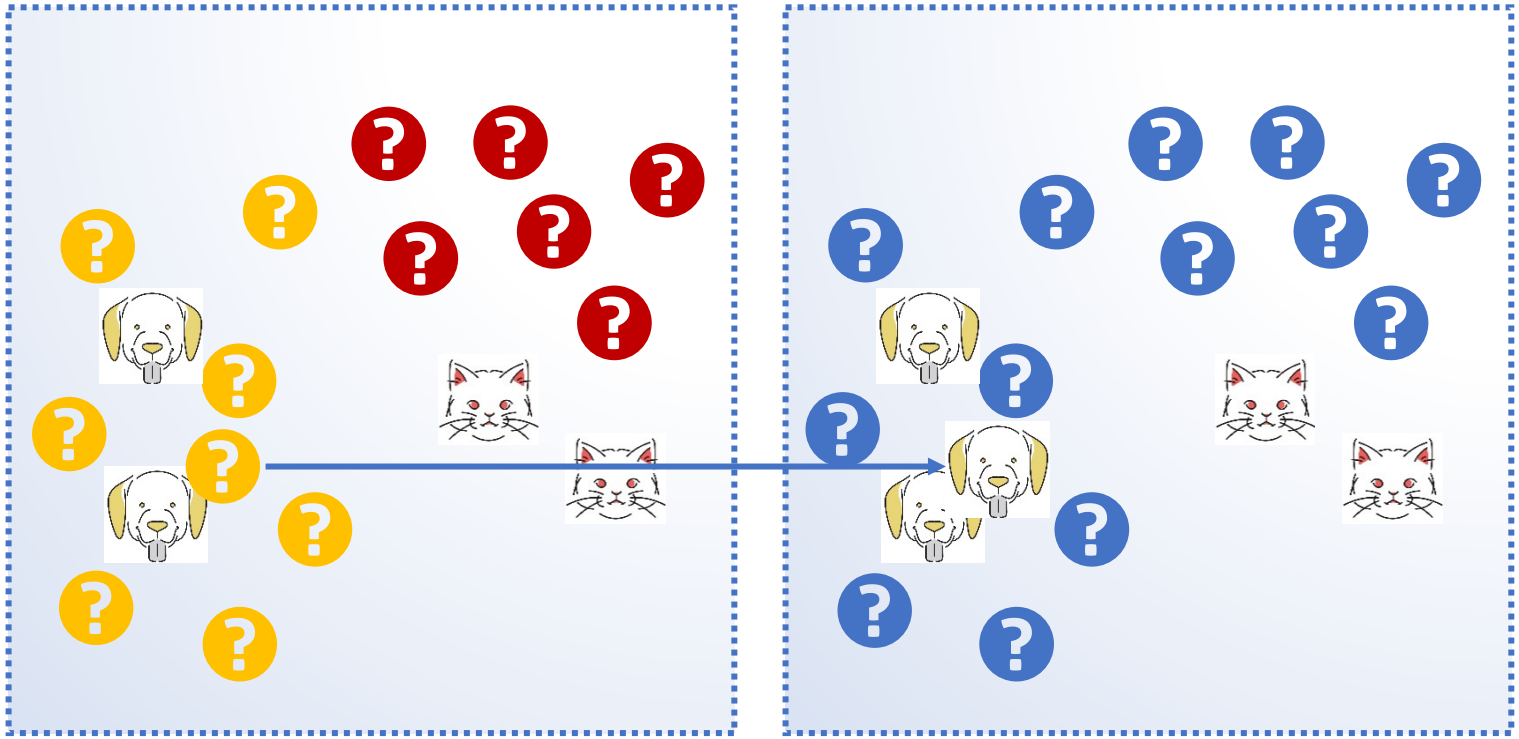
自训练算法示例

➢假设：

- 弱监督信息 = 少量有标记样本
- 分类器 = 2-Nearest Neighbor (2NN)
- 分类信心 =与最近邻居的距离
- $M = 1$

分类得到各个无标记样本的“伪标签”

扩展训练集



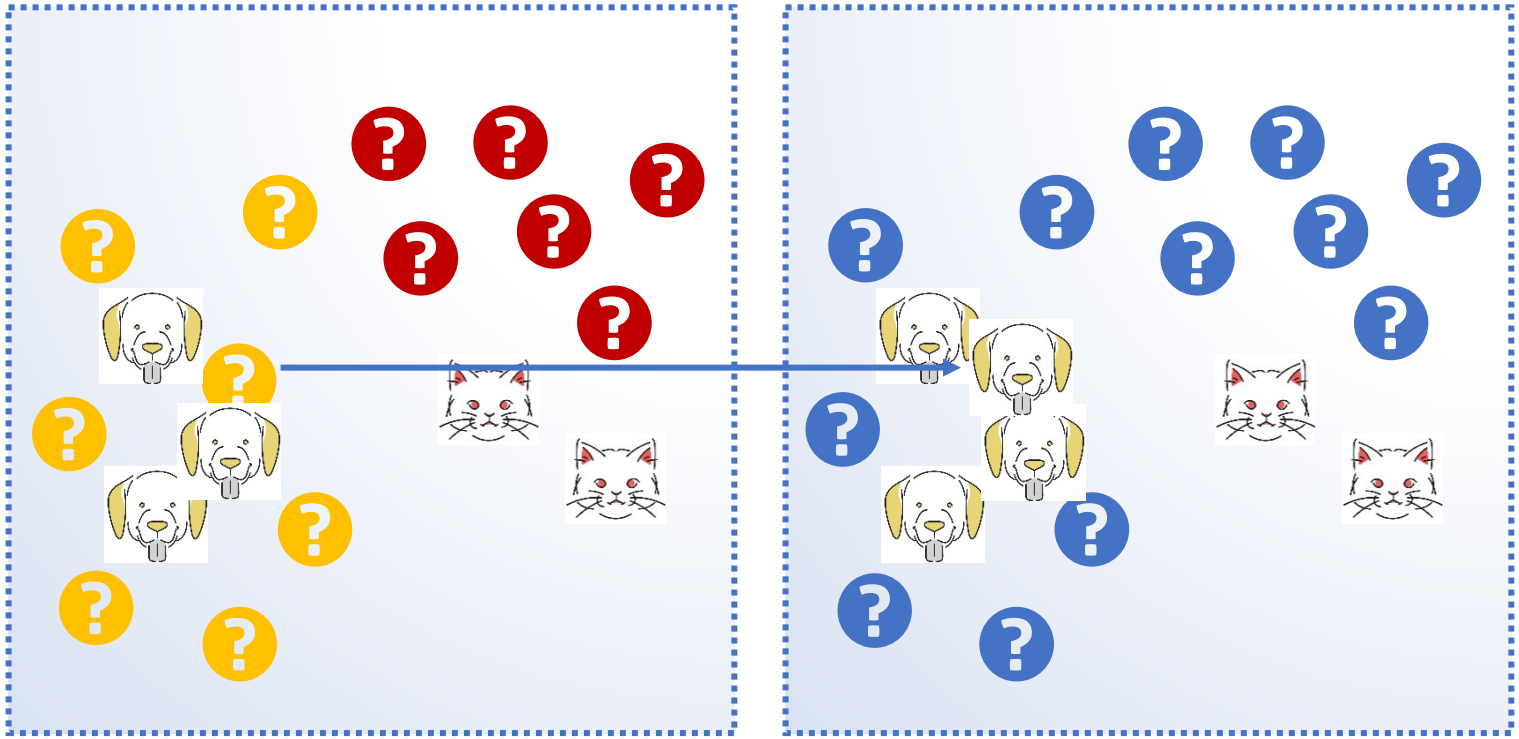
自训练算法示例

➢假设：

- 弱监督信息 = 少量有标记样本
- 分类器 = 2-Nearest Neighbor (2NN)
- 分类信心 = 与最近邻居的距离
- $M = 1$

分类得到各个无标记样本的“伪标签”

继续扩展训练集



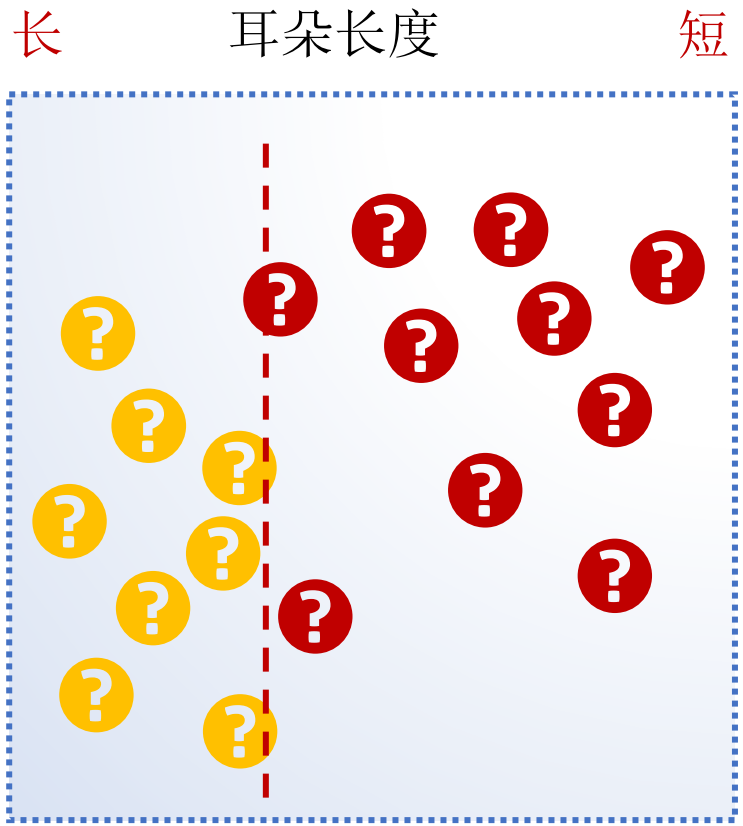
自训练算法示例

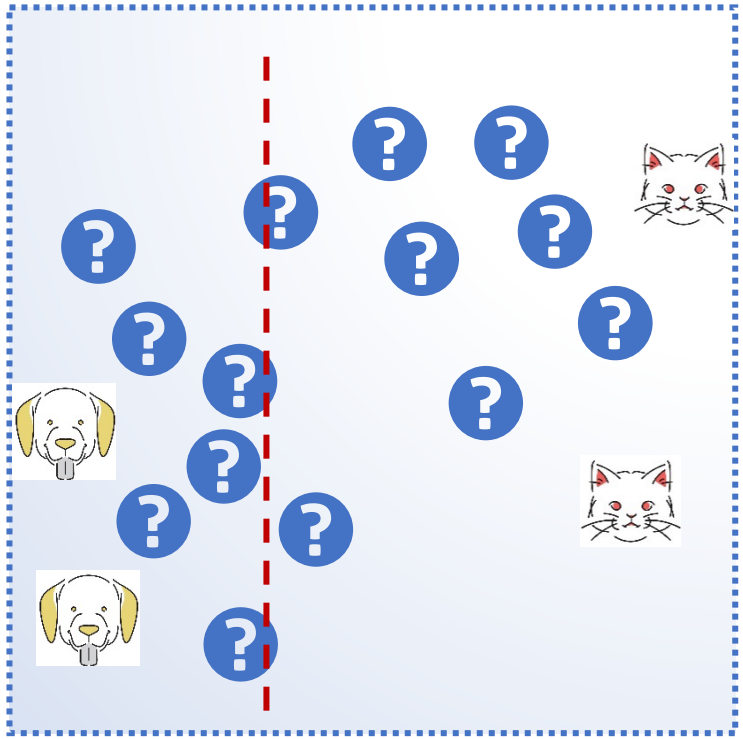
➢假设：

- 弱监督信息 = 关于类的先验知识
- 分类器 = 线性分类器
- 分类信心 = 与决策边界的距离
- $M = 2$

分类得到各个无标记样本的“伪标签”

扩展训练集



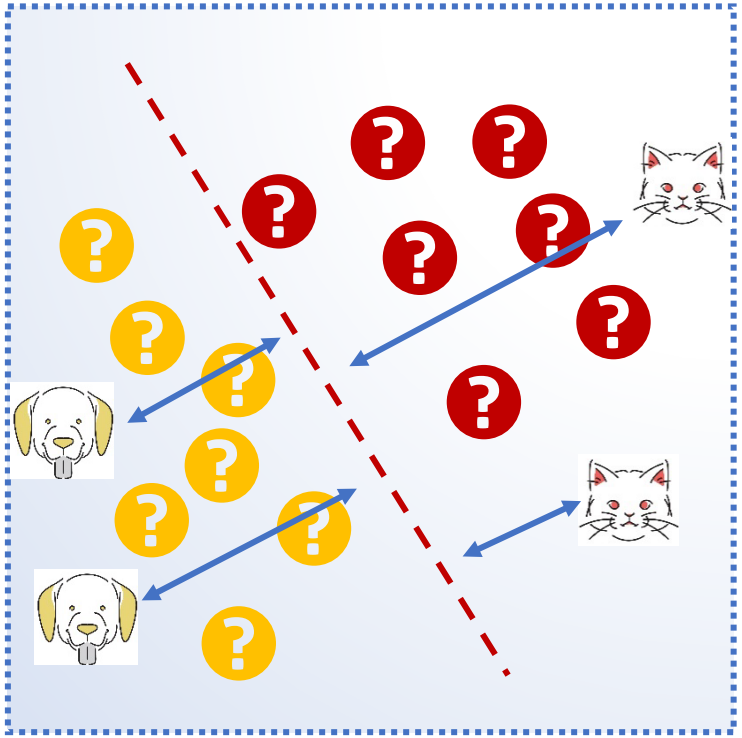


自训练算法示例

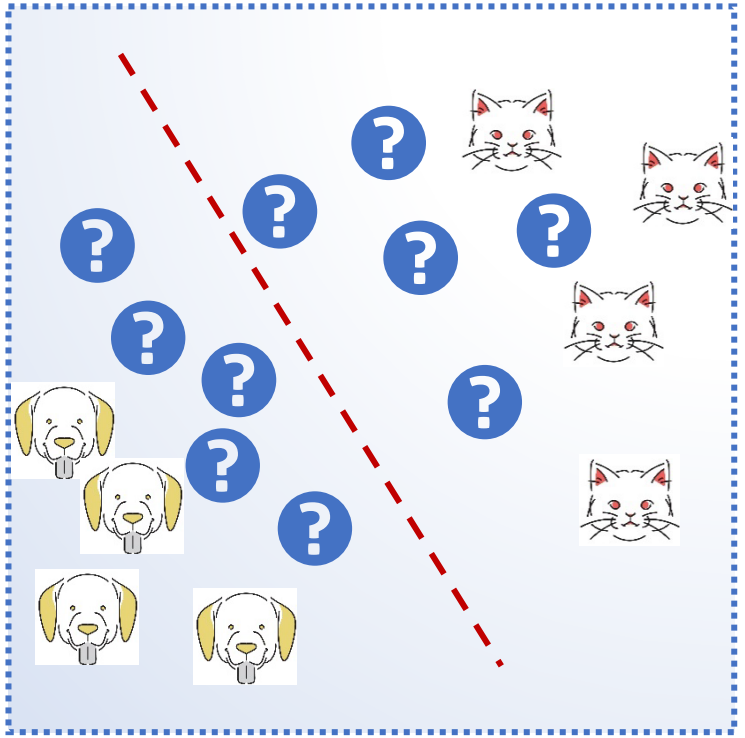
➢假设：

- 弱监督信息 = 关于类的先验知识
- 分类器 = 线性分类器
- 分类信心 = 与决策边界的距离
- $M = 2$

训练分类器



继续扩展训练集



Weakly-Supervised Text Classification
弱监督文本分类

➢无标记文本

ID	Documents
D_1	I cheered for Lakers winning NBA.
D_2	I am sad that Heat lost.
D_3	Great news! Scientists discovered …
D_4	The new film is not satisfactory.
	……

➢弱监督信息

- 少量有标记文本

ID	Documents
D_1	I cheered for Lakers winning NBA.

Sports

- 类的名称

sports

arts

science

- 类的关键词

NBA, basketball, …

film, music, …

technology, computer, …

Weakly-Supervised Text Classification
弱监督文本分类

➢采用自训练算法，关键问题：

1. 如何根据弱监督信息初始化分类器？

   或如何根据弱监督信息初始化文本的“伪标签”，进而初始化分类器？

2. 采用什么分类器？

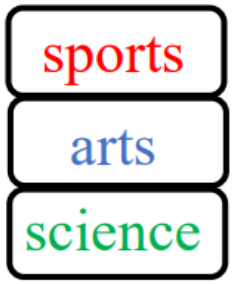
3. 如何衡量分类信心？

4. 如何扩展训练集？

➢假设：弱监督信息 = 类的名称

无标记文本

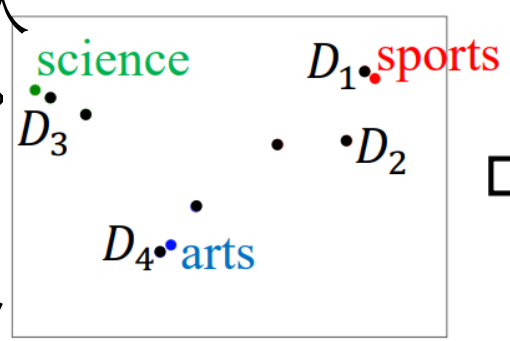
ID	Documents
D_1	I cheered for Lakers winning NBA.
D_2	I am sad that Heat lost.
D_3	Great news! Scientists discovered …
D_4	The new film is not satisfactory.
	……



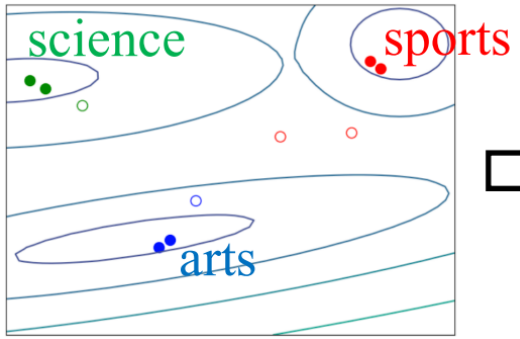
类的名称

文本嵌入

词嵌入

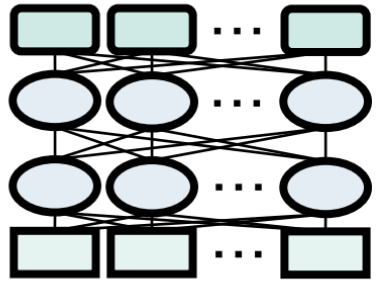


估计分类信心



扩展训练集

训练分类器



Final Project

期末项目：弱监督文本分类

- 基本要求：设计并实现一个弱监督文本分类模型
  - 报告模型在两个数据集(暂定)上的性能
  - 使用至少两个评价方法
  - 推荐采用自训练算法
- 进阶要求:(下节课介绍)
  - 未知类型识别
- 验收：第17、18周(暂定)自愿课堂展示，有加分
- 提交文件：21******_wangxiaoming.zip，内含
  ① 实验报告：21******_wangxiaoming.pdf
  ② 代码：/code　　如果代码分成多个文件，需要写readme

Text Classification
文本分类

① 构建词表

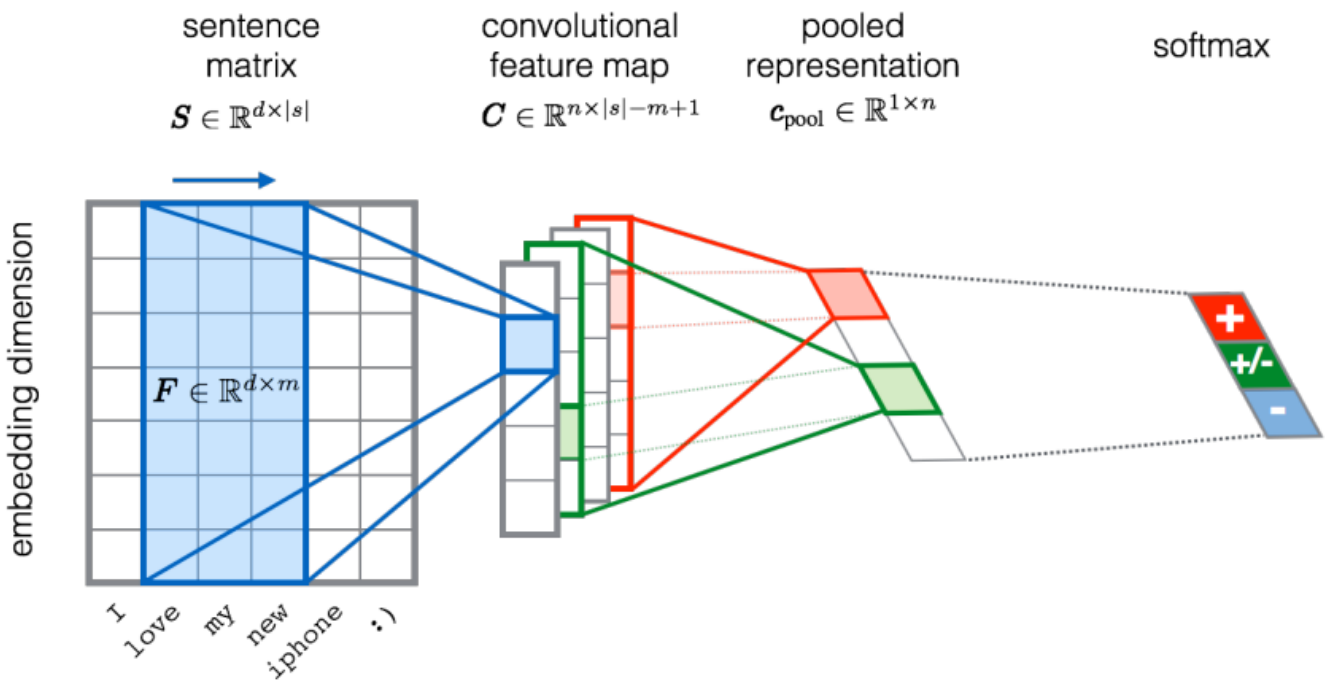
词清洗：低频词，标点，数字等

② 构建输入数据

统一序列长度为 L

① 对于长度大于 L 的句子：截取前 L 个单词，或截取前 $(L/2)$ 个单词+最后 $(L/2)$ 个单词

② 对于长度小于 L 的句子：用一个特殊字符 [PAD] 填充



$L = 10$

I	love	my	new	iphone	:)	[PAD]	[PAD]	[PAD]	[PAD]