自然语言处理

杨海钦

2025-2026-1学期

基于[SLP]课件更新

面试问题

1. 你会在文本分类中会使用手动特征吗？用了哪些？为什么？
2. 什么是避免分类器的伤害？通常有哪些伤害？有什么解决方案？

大纲

- 回顾
- 手动特征抽取
- 避免分类器的伤害

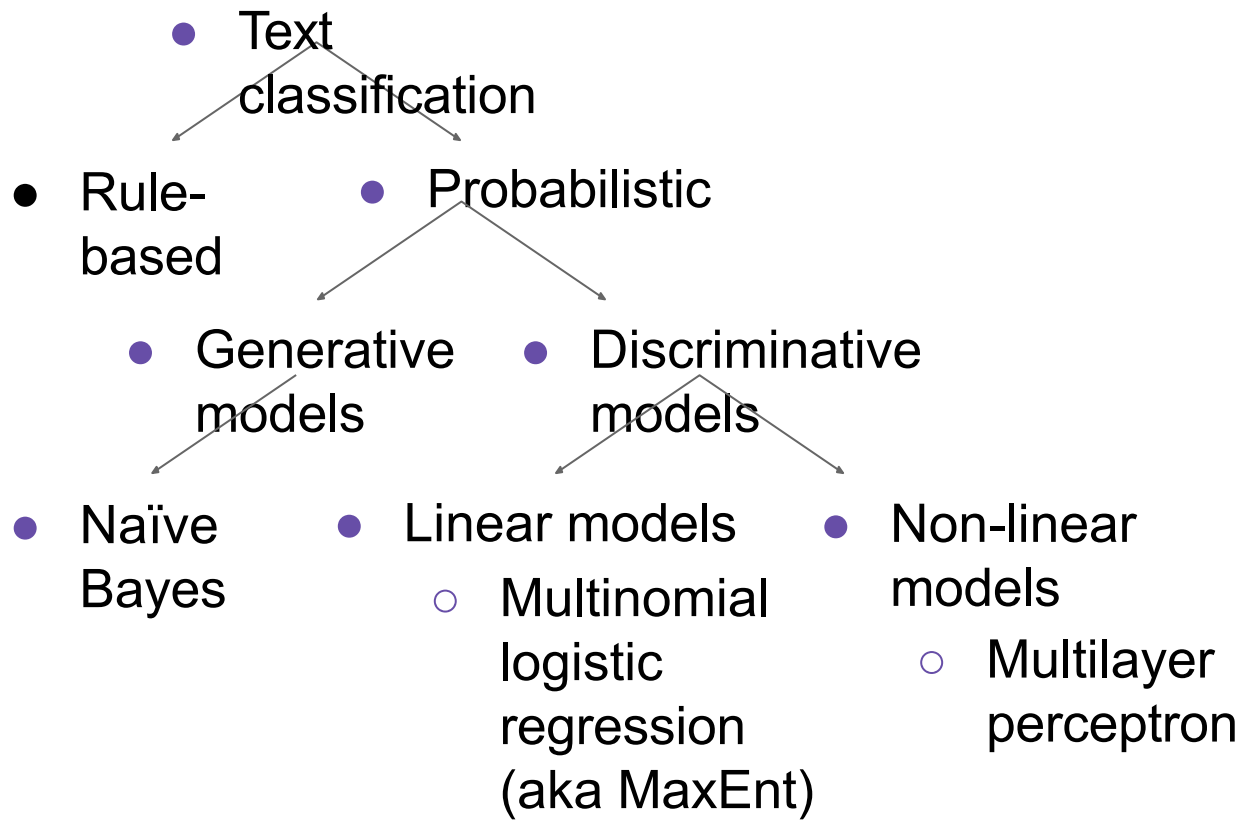
监督分类形式化定义
Supervised Classification

- 从标注数据中学习分类模型
  - 性质(“特征/feature”)和它们的重要性(“权重/weights”)
- X: 数据属性或特征的集合$\{x_1, x_2, \cdots, x_n\}$
  - 如：水果的测量值或从输入文档中提取的每个单词的计数
- Y: “类”标签，来自标签集$\{y_1, y_2, \cdots, y_k\}$
  - 如：水果类型，垃圾邮件/非垃圾邮件，积极/消极/中性

- **训练:** 给定标注的训练数据$\{(x_i, y_i)\}_{i=1}^{N}$，学习判定函数$f: x \in X \rightarrow y \in Y$
- **推理**：使得学到的f，对未来样本x的预测$\hat{y}$, 其预测尽可能准确

实现分类的监督机器学习方法

- 大量方法
  - 朴素贝叶斯
  - 逻辑回归
  - 神经网络、支持向量机
  - k-近邻
  - 大语言模型
    - 基于分类的微调
    - 基于分类的提示词



任务及传统手动特征抽取

核心任务	优先抽取特征	
文本分类(通用)	1. n-gram (1-3 gram) 2. 词性分布 (名词/动词/形容词占比)	3. 词汇丰富度 4. 领域词典匹配词频
信息检索	1. TF-IDF， BM25 2. 文档长度与平均长度比值	3. 核心术语词频 (标题/首段关键词) 4. n-gram短语
情感分析	1. 情感词典匹配(正面/负面/中性词频)	2. 标点符号占比 (感叹号/问号) 3. 程度副词词频
垃圾文本检测	1. 特殊字符/数字占比 2. 无关词汇(广告术语)词频	3. 文本长度异常值 4. 重复短语词频

TF-IDF

- 相较于词频，TF-IDF还综合考虑词语的稀有程度

$$\mathrm{TF-IDF}(t,d) = \frac{\mathrm{TF}(t,d)}{\mathrm{DF}(t)} = \mathrm{TF}(t,d) \cdot \mathrm{IDF}(t)$$

- 其中，
  - t: 单词/term
  - d: 文档/document
  - $\mathrm{TF}(t,d)$: t在d中的出现频次
  - $\mathrm{DF}(t)$: 有多少篇文档包含t
  - IDF: DF的倒数(inverse), $\mathrm{IDF}(w) = \log_{10}\left(\frac{\text{语料库总文档数}}{\text{包含目标词的文档数}}\right)$

BM25

- 如何衡量词语与文档的关联程度?
- BM25 (Best Match): TF-IDF的改进变种
  - 相似的句子将得到更高的投票

$$\mathrm{BM25}(D,Q)=\sum_{q_i\in Q}\mathrm{IDF}(q_i)\cdot\frac{\mathrm{TF}(q_i,D)\cdot(k_1+1)}{\mathrm{TF}(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{avg\ DL}\right)}$$

  - 其中
    - k_1和 b是两个常数， $avg\ DL$是所有文档的平均长度

- k_1越大， TF对文档得分的正面影响越大　　k_1默认1.2，范围0-3
- b越大，文档长度对得分的负面影响越大　　b默认0.75，范围0-1

In-class Practice

避免分类器的伤害/harms

- 像任何NLP算法一样，分类器可能会造成伤害
- 对所有分类器，无论是朴素贝叶斯还是其他算法，都存在伤害

详细见[SLP]Sec. 4.10

代表性伤害/Representational Harms

- 贬低一个社会群体的制度所造成的危害
  - 如，对该群体的负面刻板印象/perpetuating negative stereotypes
- Kiritchenko和Mohammad 2018年的研究
  - 通过句子对检查200个情感分析系统
  - 仅替代名字，其他保持相同：
    - 非洲裔美国人的常用名(Shaniqua)或欧洲裔美国人(Stephanie)
    - 如"I talked to Shaniqua yesterday"和"I talked to Stephanie yesterday"
- 结果: 系统给非裔美国人名字的句子分配了更低的情绪值和更多的负面情绪
- 下游任务的危害:
  - 使对非裔美国人的刻板印象永久化
  - 非裔美国人在情感等NLP工具（广泛用于市场研究、心理健康研究等）上受到不同对待

Saif M. Mohammad, Svetlana Kiritchenko: Understanding Emotions: A Dataset of Tweets to Study Interactions between Affect Categories. LREC 2018

审查的危害/Harms of Censorship

- 毒性检测是检测仇恨言论、辱骂、骚扰或其他类型的有毒语言的文本分类任务
  - 广泛用于在线内容审核
- 毒性分类器错误地把无毒句子标记了有毒，仅仅因为它们提到少数群体身份（如“盲人”或“同性恋”）
  - 女性（Park et al., 2018），残疾人（Hutchinson et al., 2020），同性恋者（Dixon et al., 2018；Oliva et al., 2021）
- 下游任务的危害:
  - 对残疾人和其他群体的言论进行审查
  - 这些团体的言论在网上变得不那么显眼了
  - 算法可能会推动作者避开这些词汇，以致人们不太可能写这些词或涉及这些群体

性能差异/Performance Disparities

1. 因为缺乏数据或标签，文本分类器在许多语言上表现较差
2. 文本分类器甚至在资源丰富的语言(如英文)中表现更差
   - 示例任务：语言识别/language identification，NLP pipeline的第一步 ("Is this post in English or not?")
   - 非洲裔美国作家(Blodgett and O'connor 2017)或印度作家(Jurgens et al., 2017)的英语语言检测表现更差

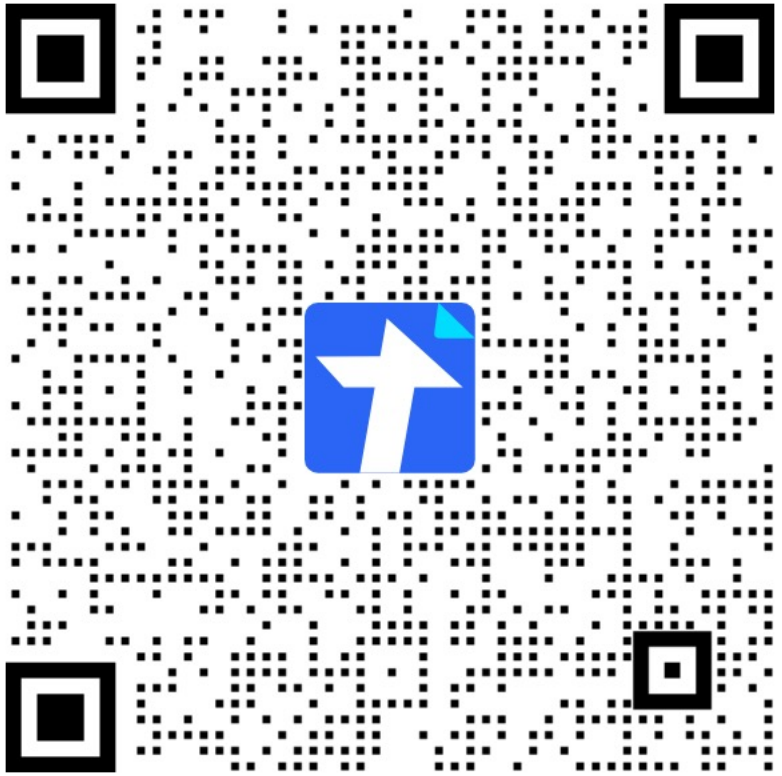
1. Su Lin Blodgett, Brendan O'Connor: Racial Disparity in Natural Language Processing: A Case Study of Social Media Africa-American English. CoRR abs/1707.00061 (2017)
2. David Jurgens, Yulia Tsvetkov, Dan Jurafsky: Incorporating Dialectal Variability for Socially Equitable Language Identification. ACL (2) 2017: 51-57

文本分类中的危害/Harms in text classification

- **原因**:
  - 数据中的问题；NLP系统放大了训练数据中的偏见
  - 标签中的问题
  - 算法中的问题（如模型训练的优化的方向）
- **普遍性**: 同样的问题出现在整个NLP（包括大型语言模型）中
- **解决方案**: 没有一般的缓解或解决方案
  - 但是减轻危害是一个活跃的研究领域
  - 我们可以用一些标准的基准和工具来衡量一些危害
  - 发布模型卡片/model card
    - 训练算法和参数
    - 训练数据源、动机和预处理
    - 评测数据源，动机和预处理
    - 预期用途和用户
    - 不同人群或其他群体和环境情况下的模型表现

一句话总结

- 传统手工特征抽取方法
  - TF-IDF, BM25
- 分类器的伤害
  - 代表性伤害/representational harms
  - 审查的危害/Harms of Censorship
  - 性能差异/Performance Disparities
  - Solution: model cards



- 课外阅读
  - [SLP3] Ch. 4.10; Ch.14.1

附录: BM25

BM25的变种

Robertson et al.	$\sum_{t \in q} \log\left(\frac{N - df_t + 0.5}{df_t + 0.5}\right) \cdot \frac{tf_{td}}{k_1 \cdot \left(1 - b + b \cdot \left(\frac{L_d}{L_{avg}}\right)\right) + tf_{td}}$
Lucene (default)	$\sum_{t \in q} \log\left(1 + \frac{N - df_t + 0.5}{df_t + 0.5}\right) \cdot \frac{tf_{td}}{k_1 \cdot \left(1 - b + b \cdot \left(\frac{L_{dlossy}}{L_{avg}}\right)\right) + tf_{td}}$
Lucene (accurate)	$\sum_{t \in q} \log\left(1 + \frac{N - df_t + 0.5}{df_t + 0.5}\right) \cdot \frac{tf_{td}}{k_1 \cdot \left(1 - b + b \cdot \left(\frac{L_d}{L_{avg}}\right)\right) + tf_{td}}$
ATIRE	$\sum_{t \in q} \log\left(\frac{N}{df_t}\right) \cdot \frac{(k_1 + 1) \cdot tf_{td}}{k_1 \cdot \left(1 - b + b \cdot \left(\frac{L_d}{L_{avg}}\right)\right) + tf_{td}}$
BM25L	$\sum_{t \in q} \log\left(\frac{N + 1}{df_t + 0.5}\right) \cdot \frac{(k_1 + 1) \cdot (c_{td} + \delta)}{k_1 + c_{td} + \delta}$
BM25+	$\sum_{t \in q} \log\left(\frac{N + 1}{df_t}\right) \cdot \left(\frac{(k_1 + 1) \cdot tf_{td}}{k_1 \cdot \left(1 - b + b \cdot \left(\frac{L_d}{L_{avg}}\right)\right) + tf_{td}} + \delta\right)$
BM25-adpt	$\sum_{t \in q} G_q^1 \cdot \frac{\left(k_1' + 1\right) \cdot tf_{td}}{k_1' \cdot \left(1 - b + b \cdot \left(\frac{L_d}{L_{avg}}\right)\right) + tf_{td}}$
$\mathrm{TF}_{lo\delta op} \times \mathrm{IDF}$	$\sum_{t \in q} \log\left(\frac{N + 1}{df_t}\right) \cdot \left(1 + \log\left(1 + \log\left(\frac{tf_{td}}{1 - b + b \cdot \left(\frac{L_d}{L_{avg}}\right)} + \delta\right)\right)\right)$

Chris Kamphuis, Arjen P. de Vries, Leonid Boytsov, Jimmy Lin: Which BM25 Do You Mean? A Large-Scale Reproducibility Study of Scoring Variants. ECIR (2) 2020: 28-34

In-class Practice: TF-IDF

Q: 根据下面三篇英文文档，计算其关于“machine”的TF-IDF值：

- D1: "I love machine learning. Machine is amazing!"
- D2: "Data science and machine learning are popular."
- D3: "Python is a useful programming language."

Ans. to In-class Practice: TF-IDF

Q: 根据下面三篇英文文档，计算其关于“machine”的TF-IDF值：

- D1: "I love machine learning. Machine is amazing!"
- D2: "Data science and machine learning are popular."
- D3: "Python is a useful programming language.”

解:

1. 根据TF的定义，$\mathrm{TF}(w,d)=\frac{\text{目标词在文档中出现的次数}}{\text{文档总词数}}$，因此

- $\mathrm{TF}(\text{”machine”},d_1)=2/7\approx 0.29$
- $\mathrm{TF}(\text{”machine”},d_2)=1/7\approx 0.14$
- $\mathrm{TF}(\text{”machine”},d_3)=\frac{0}{7}=0$

2. 计算IDF(“machine”的逆文档频率，$\mathrm{IDF}(w)=\log_{10}\left(\frac{\text{语料库总文档数}}{\text{包含目标词的文档数}}\right)$

- $\mathrm{IDF}(\text{”machine”})=\log_{10}(3/2)=\log_{10}(1.5)\approx 0.18$

3. 计算“machine”在各文档中的TF-IDF值:

- 在d_1中，$\text{TF-IDF}(\text{”machine”},d_1)\approx 0.29^{*}0.18\approx 0.05$
- 在d_2中，$\text{TF-IDF}(\text{”machine”},d_2)\approx 0.14^{*}0.18\approx 0.03$
- 在d_3中，$\text{TF-IDF}(\text{”machine”},d_3)=0^{*}0.18=0$