自然语言处理

杨海钦

2025-2026-1学期

基于Yulia, Graham, Diyi, Jixing等课件更新

面试问题

- 假设你需要为公司设计一个 “垃圾邮件自动识别系统”(输入一封邮件文本，输出“垃圾邮件”或“正常邮件”)，请回答以下问题：
  - 对原始邮件文本(如含标题、正文、发件人签名的字符串)，你会做哪些核心预处理操作？请说明每个操作的目的。
  - 你会选择什么模型，分别说明它们适合/不适合该任务的核心原因。
  - 若数据集中“正常邮件占90%，垃圾邮件占10%”(类别不平衡)，仅用“准确率”评估模型是否合理？若不合理，应优先选择哪些评估指标？请解释理由。
  - 上线后发现“部分重要工作邮件被误判为垃圾邮件”(漏判正常邮件)，你会从哪些角度优化模型？

大纲

- 文本分类的应用

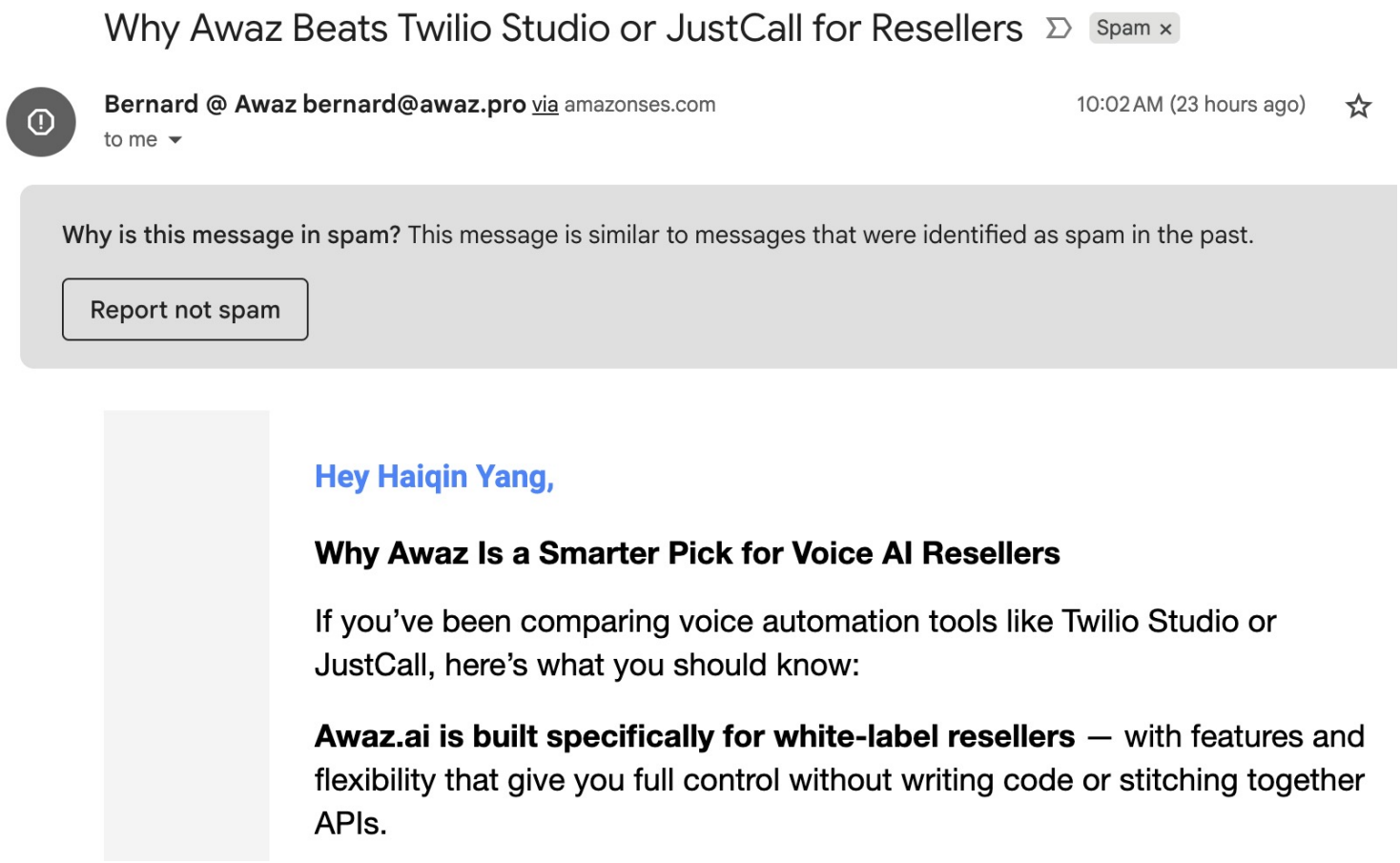
- 文本分类的流程
  - 如何提前特征
  - 如何构建判定函数
    - 基于规则的方法
    - 基于概率模型的学习方法
      - 朴素贝叶斯

自然语言处理系统通用算法框架

- 创建一个函数将输入X映射到输出Y，其中X和/或Y涉及语言
- 任务

输入X	输出Y	任务
文本	连续文本	语言模型 Language modeling
文本	其他语言的文本	机器翻译 Machine Translation
文本	概要	自动摘要 Automatic Summarization
文本	标签	文本分类 Text Classification
文本	语言结构	语言分析 Language Analysis
语音/ 文本	文本/ 语音	语音识别 ASR (Automatic Speech Recognition) /语音合成 TTS (Text-to-Speech)
图片/ 文本	文本/ 图片(视频)	图片描述 Image Captioning /文生图 T2I (文生视频 T2V)

垃圾邮件？Spam Email?



Why Awaz Beats Twilio Studio or JustCall for Resellers Spam ×

Bernard @ Awaz bernard@awaz.pro via amazonses.com 10:02 AM (23 hours ago)

to me

Why is this message in spam? This message is similar to messages that were identified as spam in the past.

Report not spam

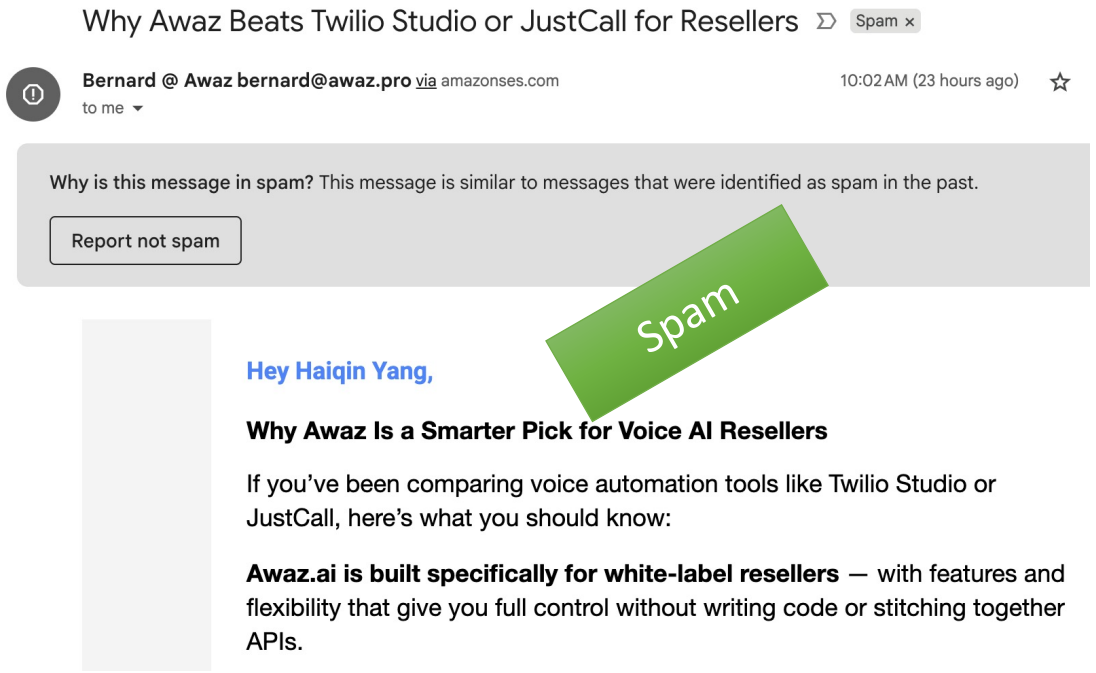
Hey Haiqin Yang,

Why Awaz Is a Smarter Pick for Voice AI Resellers

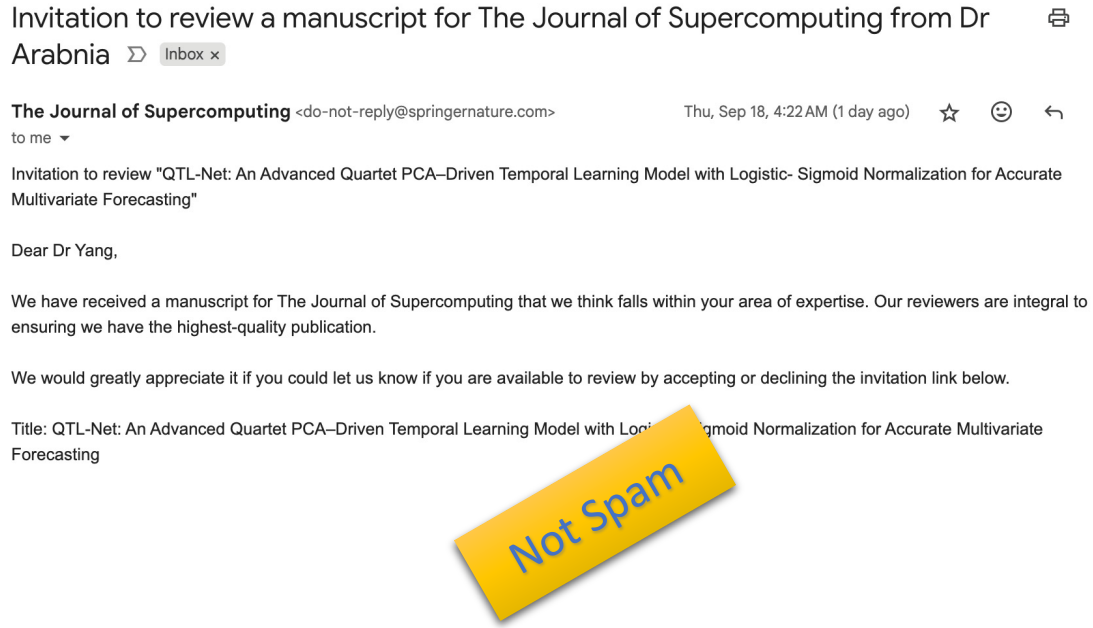
If you've been comparing voice automation tools like Twilio Studio or JustCall, here's what you should know:

Awaz.ai is built specifically for white-label resellers — with features and flexibility that give you full control without writing code or stitching together APIs.

垃圾邮件分类 Spam Classification



Spam

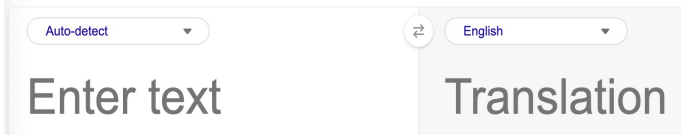


Not Spam

语种识别 Language ID

Аяны замд тур зогсон тэнгэрийн байдлыг ажиглаад хедлех зуур гутал дор шинэхэн орсон цас шаржигнан дуугарч байв. Цасны тухай бодол сонин юм. Хот хурээ тийш цас орвол орно л биз гэсэн хэнэггуй бодол маань хедее талд, говийн ээрэм хендийд, малын бэлчээрт, малчдын хотонд болохоор солигдож эргэцуулэн бодох нъ хачин. Цас хэр орсон бол?

Београд, 16. јуН 2013. године — Председник Владе Републике Србије Ивица Дачић честитао је кајакашици златне медаље иолимпијској дисциплини К-1, 500 метара, као и у двоструко дужој стази освојене на првенству Европе у Португалији.



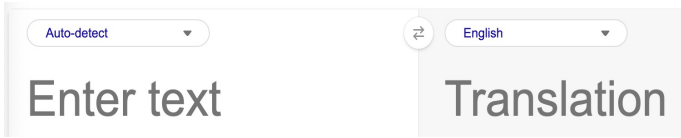
Beograd,16. jun 2013. godine - Predsednik Vlade Republike Srbije Ivica Dačic čestitao je kajakašici zlatne medalje u olimpijskoj disciplini K-1,500 metara,kao iu dvostruko dužoj stazi osvojene na prvenstvu Evrope u Portugaliji.

Nestrankarski Urad za vladno odgovornost ZDA je objavil eksplozivno mnenje,da je vlada predsednika Donalda Trumpa kršila zvezno zakonodajo, ko je zadrževala izplačilo kongresno potrjene vojaške pomoči Ukrajini zaradi političnih razlogov. Predstavniški dom kongresa je prav zaradi tega sprožil ustavno obtožbo proti Trumpu.

语种识别 Language ID

Аяны замд тур зогсон тэнгэрийн байдлыг ажиглаад хедлех зуур гутал дор шинэхэн орсон цас шаржигнан дуугарч байв. Цасны тухай бодол сонин юм. Хот хурээ тийш цас орвол орно л биз гэсэн хэнэггүй бодол маань хедее талд говийн аяам хөндийд, малын бэлчээрт, малчдын хотонд болохоор mongolian ээн бодох нъ хачин. Цас хэр орсон бол?

Београд, 16. јуН 2013. године — Председник Владе Републике Србије Ивица Дач serbian кашици златне медаље иолимпијској дисциплини К-1, 500 оструко дужој стази освојене на првенству Европе у Португалији.



Beograd,16. jun 2013. godine - Predsednik Vlade Republike Srbije Ivica Dačic čestitao je kajakašici zlatne m bosnia disciplini K-1,500 metara,kao iu dvostruko dužoj stazi osvoje ope u Portugaliji.

Nestrankarski Urad za vladno odgovornost ZDA je objavil eksplozivno mnenje,da je vlada predsednika Don slovenian vezno zakonodajo, ko je zadrževala izplačilo kongresno potrjene vojaške pomoči Ukr razlogov. Predstavniški dom kongresa je prav zaradi tega sprožil ustavno obtožbo proti Trumpu.

情感分析 Sentiment Analysis

- ...zany characters and richly applied satire, and some great plot twists

- It was pathetic. The worst part about it was the boxing scenes...

- ...awesome caramel sauce and sweet toasty almonds. I love this place!

- ...awful pizza and ridiculously overpriced...

情感分析 Sentiment Analysis

By **John Neal**

This review is from: **Accoutrements Horse Head Mask (Toy)**

When I turned State's Witness, they didn't have enough money to put me in the Witness Protection Program, so they bought me this mask and gave me a list of suggested places to move. Since then I've lived my life in peace and safety knowing that my old identity is forever obscured by this life-saving item.

By **Christine E. Torok**

Verified Purchase (What's this?)

First of all, for taste I would rate these a 5. So good. Soft, true-to-taste fruit flavors like the sugar variety...I was a happy camper.

BUT (or should I say BUTT), not long after eating about 20 of these all hell broke loose. I had a gastrointestinal experience like nothing I've ever imagined. Cramps, sweating, bloating beyond my worst nightmare. I've had food poisoning from some bad shellfish and that was almost like a skip in the park compared to what was going on inside

情感分析 Sentiment Analysis

By John Neal

This review is from: Accoutrements Horse Head Mask (Toy)

When I turned State's Witness, they didn't have enough money to put me in the Witness Protection Program, so they bought me this mask and some fake identification papers to move. Since then I've lived my life in peace and safety knowing that my old identity is forever obscured by this life-saving item.

By Christine E. Torok

Verified Purchase (What's this?)

First of all, for taste I would rate these a 5. So good. Soft, true-to-taste fruit flavors like the sugar variety...I was a happy camper.

BUT (or should I say BUTT), not long after eating about 20 of these all hell broke loose. I had a gastrointestinal experience like nothing I've ever imagined. Cramps, sweating, bloating beyond my worst nightmare. I've had food poisoning from some bad shellfish and that was almost like a skip in the park compared to what was going on inside

主题分类 Topic Classification

MEDLINE Article

Brain and Cognition

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

?

- MeSH Subject Category Hierarchy
  - Antagonists and Inhibitors
  - Blood Supply
  - Chemistry
  - Drug Therapy
  - Embryology
  - Epidemiology
  - …

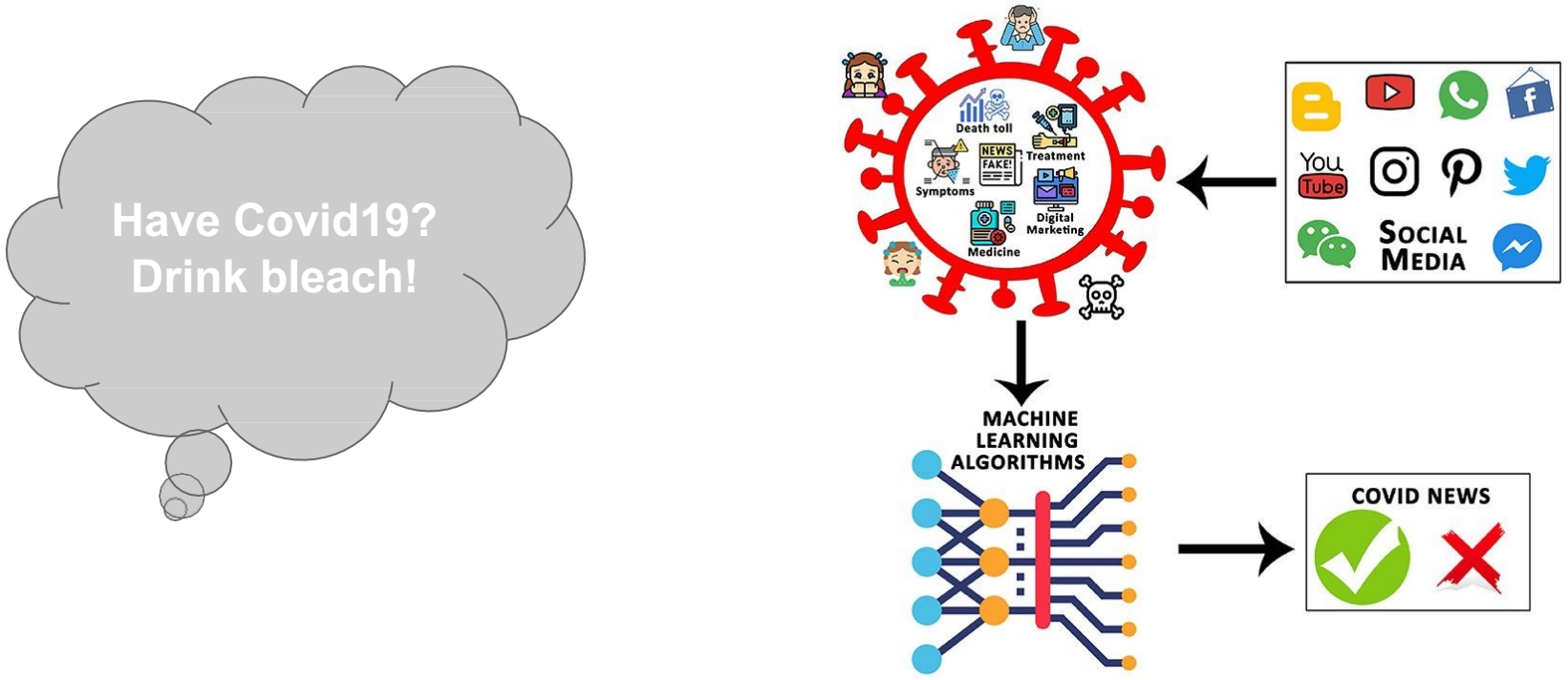
作者归属Author Attribution：作者是男性还是女性？

By 1925 Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam.

Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of the greatest assets...

S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. “Gender, Genre, and Writing Style in Formal Written Texts,” Text, volume 23, number 3, pp. 321–346

事实验证Fact Verification:
可信还是虚假 Trustworthy or Fake?



Detecting COVID-19-Related Fake News Using Feature Extraction

Suleman Khan, Saqib Hakak, N. Deepa, B. Prabadevi, Kapal Dev and Silvia Trelova

文本分类/Text Classification

- 对文本内容进行分类：
  - 垃圾邮件检测(二分类：垃圾邮件/非垃圾邮件)
  - 情感分析(二元或多元)
    - 电影、餐厅、产品评论(正面/负面或1-5星)
    - 政治论点(赞成/反对，或赞成/反对/中立)
    - 主题分类(多向：体育/金融/旅游等)
  - 语言识别(多种方式：语言、语系)
  - …
- 对文本的作者进行分类(作者归属)
  - 人工生成还是机器生成?
  - 母语识别(例如，量身定制语言辅导)
  - 疾病诊断(精神或认知障碍)
  - 识别性别，方言，教育背景，政治倾向(例如，在法医学[法律事务]，广告/营销，竞选活动，虚假信息)
  - …

文本分类/Text Classification

- **目标**：获得函数f，使得在给定输入x的情况下预测$\hat{y}$, 其预测尽可能准确；其中f由$\{(x_i, y_i)\}_{i=1}^{N}$训练获得

复习:文本分类步骤

- **特征提取**：从文本中提取出决策所需的显著特征
- **分数计算**：计算一个或多个可能性的分数
- **决策函数**：从几种可能性中选择一种
- **精度计算** Accuracy Calculation
- **误差分析** Error Analysis

- **特征提取**：$\mathbf{h} = F(\mathbf{x})$
- **分数计算**：
  - 二分类：$s = \mathbf{w} \cdot \mathbf{h} = f(\mathbf{h})$
  - 多分类：$\mathbf{s} = \mathbf{W}\mathbf{h} = f(\mathbf{h})$
- **决策函数**：
  - $\hat{y} = \text{decide}(s/\mathbf{s})$

后续课程，我们将探讨

1. 如何将文本“吸收/digest”成函数可用的形式F?
   (关键词：特征、特征提取、特征选择、表征)
   feature, extraction/selection, representation

2. 我们可以用什么样的策略来创建决策函数f?
   (关键词：建模 models)

3. 如何评估决策函数s?
   (关键词：评估 evaluation)

如何将文本“吸收/digest”成函数可用的形式F?

分类：特征(测量/Measurements)

- 测量并获得特征

4.2, 212, 3.4, 1332

diameter, weight, softness, color

5.2, 315, 5.7, 4567

diameter, weight, softness, color

文本分类-特征提取

- 如何衡量文本?

- 例子
  - I love this movie! It’s sweet, but with satirical humor. The dialogue is great, and the adventure scenes are fun… It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it just to about anyone. I’ve seen it several times, and I’m always happy to see it again whenever I have a friend who hasn’t seen it before.

文本分类-特征提取

- 如何衡量文本?

- 例子
  - I love this movie! It's sweet, but with satirical humor. The dialogue is great, and the adventure scenes are fun… It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it just to about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it before.

词袋模型 Bag-of-Words (BOW)

- 给定一个文档d(例如，一篇电影评论)—如何表示d?

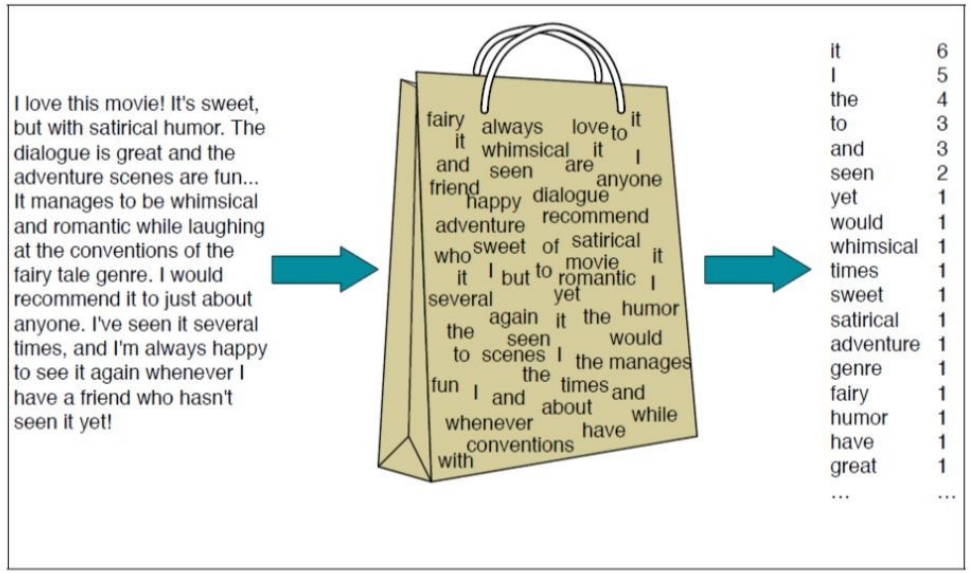


Figure 7.1 Intuition of the multinomial naive Bayes classifier applied to a movie review. The position of the words is ignored (the *bag of words* assumption) and we make use of the frequency of each word.

Figure from J&M 3rd ed. draft, sec 7.1

词袋模型：特征提取，独立性假设

I love this movie! It's sweet, but with satirical humor. The dialogue is great, and the adventure scenes are fun… It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it just to about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it before.

(几乎)整个词典(Lexicon)

Word	Count	Relative Frequency
love	10	0.0007
great	…	
recommend		
laugh		
happy		
…		
several		
…		

其他特征

- 词级别
  - 内容词、停用词
  - 标点符号？词元？词形还原？是否小写？
- 词序
  - 双元组bigrams，三元组trigrams，n元组(n-grams)
- 语法结构，句子解析树
- 单词的词性
- 词向量
- …

小结：不同文本表示的优劣势

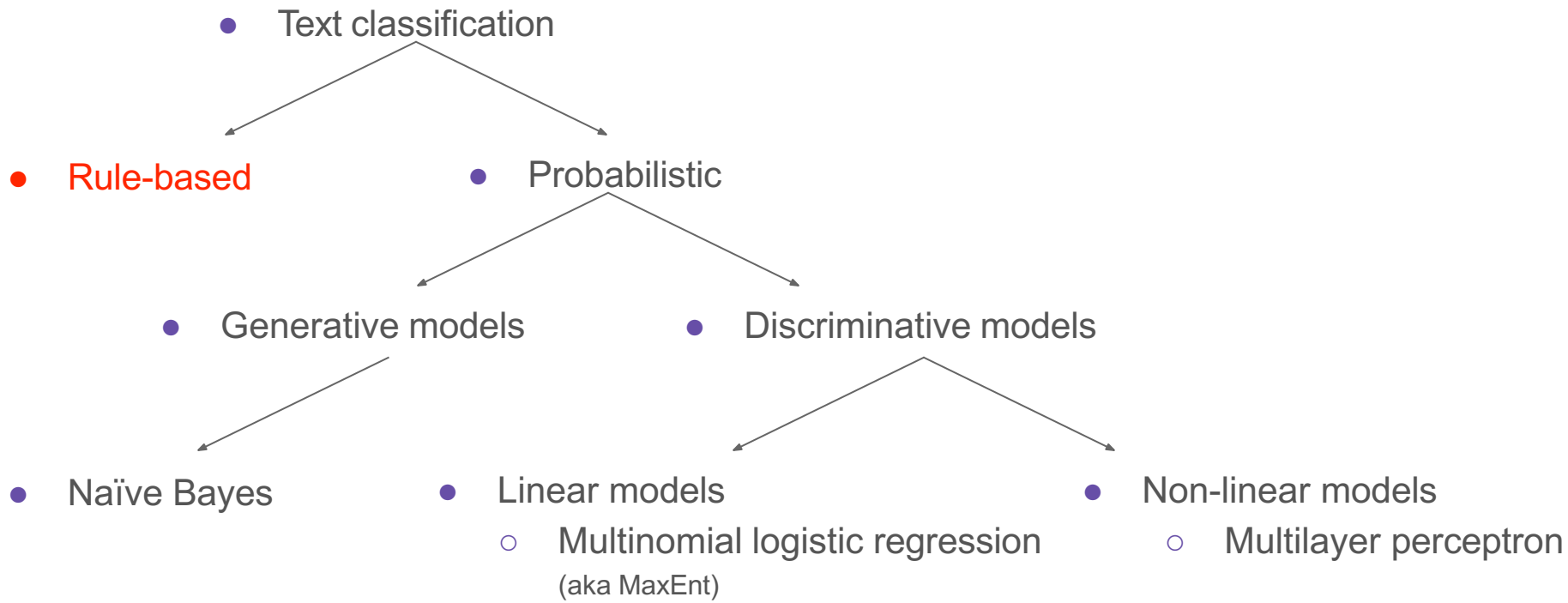
- 词袋模型 Bag-of-Words(BOW)
  - 容易获得，不需特别计算
  - 可变大小，忽略句子结构

- 手工制作/hand-crafted的特征
  - 完全控制，可以使用NLP流水线/管道(pipeline)，类特有的特征
  - 过度具体，不完整，需要使用NLP管道

- 通过学习的特征表示
  - 能学会所有相关信息
  - 需要学习

如何创建决策函数f?

文本分类模型


Text classification
Rule-based
Probabilistic
Generative models
Discriminative models
Naïve Bayes
Linear models
Multinomial logistic regression
(aka MaxEnt)
Non-linear models
Multilayer perceptron

复习：基于规则的文本分类

Ch2. pp. 22

```
def classify(x: str) -> str:
    sports_keywords = ["baseball", "soccer", "football", "tennis"]
    if any(keyword in x for keyword in sports_keywords):
        return "sports"
    else:
        return "other"
```

何时使用？ 无训练数据，仅凭直觉

基于规则的文本分类:挑战

Sentiment: Half submarine flick, half ghost story, all in one a criminally neglected film.

→ 很难提前确定哪些词是有信息的(以及它们携带了什么信息！)

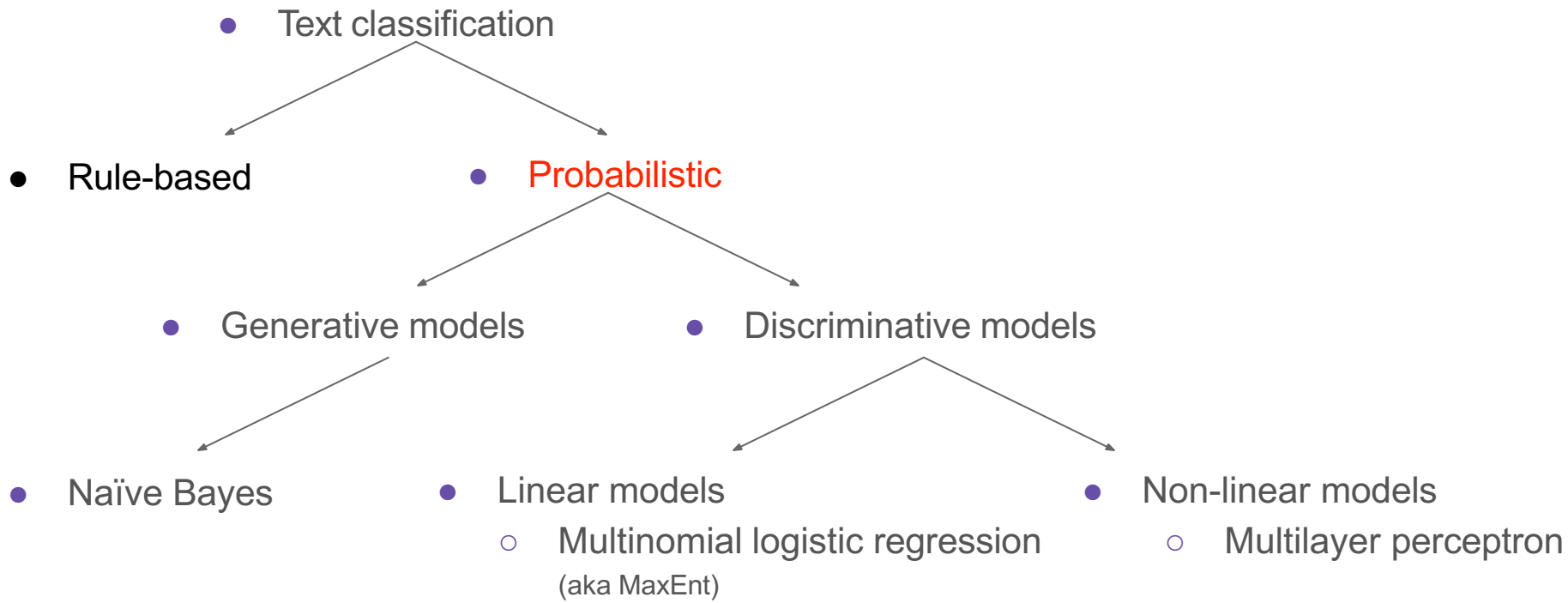
Sentiment: It's not life-affirming, it's vulgar, it's mean, but I liked it.

→ 语用学在词的层次上建模很复杂，语序(句法)很重要，但很难用规则来编码！

Language ID: All falter, stricken in kind.

→ 简单特征可能会误导结果！

文本分类模型


Text classification
Rule-based
Probabilistic
Generative models
Discriminative models
Naïve Bayes
Linear models
Multinomial logistic regression
(aka MaxEnt)
Non-linear models
Multilayer perceptron

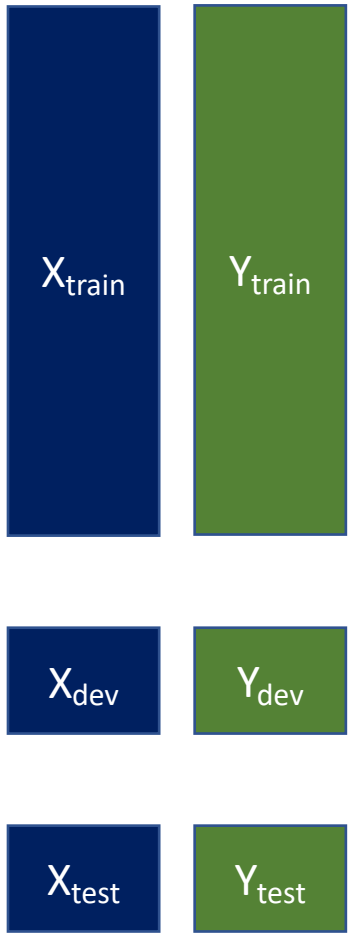
复习：机器学习定义

- 美国工程院院士Tom Mitchell的定义
  - 如果一个计算机程序在某类任务T(Task)上的性能P(Performance)，随着经验E(Experience)的增加而提升，那么我们称该程序从经验E中学习
  - <T, P, E>

机器学习典型范式

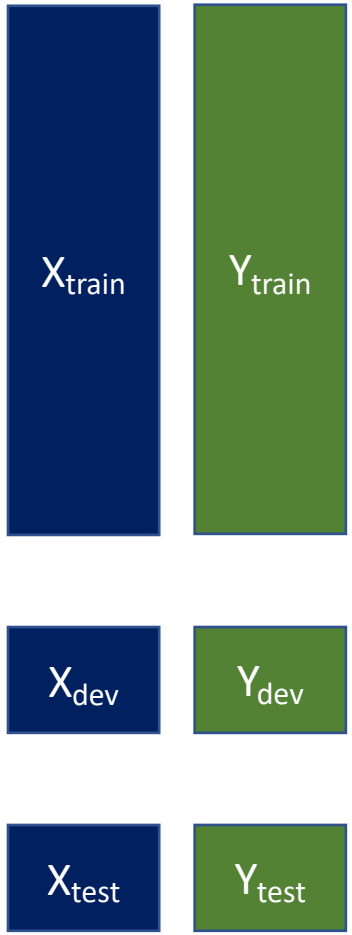
- 监督学习 Supervised Learning
  - 分类: 标注数据 labeled data
    - 二分类（喜欢，不喜欢）
    - 多类分类（政治、运动、新闻、…）
    - 多标签分类（#党派 #周五 #失败）
  - 回归:响应值是连续数值
    - 可能性: 喜欢的可能性[0-1]
    - 停留时长
    - …
- 非监督学习 Unsupervised Learning
  - 降维、聚类
- 半监督学习 Semi-supervised Learning
  - 标签数据+非标签数据
- 弱监督学习 Weakly-supervised Learning
  - 弱标签数据（非精确/低成本标签）
- 自监督学习 Self-supervised Learning
  - 非标签数据: 数据自生成监督信号
- 迁移学习 Transfer Learning
  - 源领域知识(已学模型/数据规律)+ 目标领域任务(小样本/新场景)
- …

复习:三组数据集

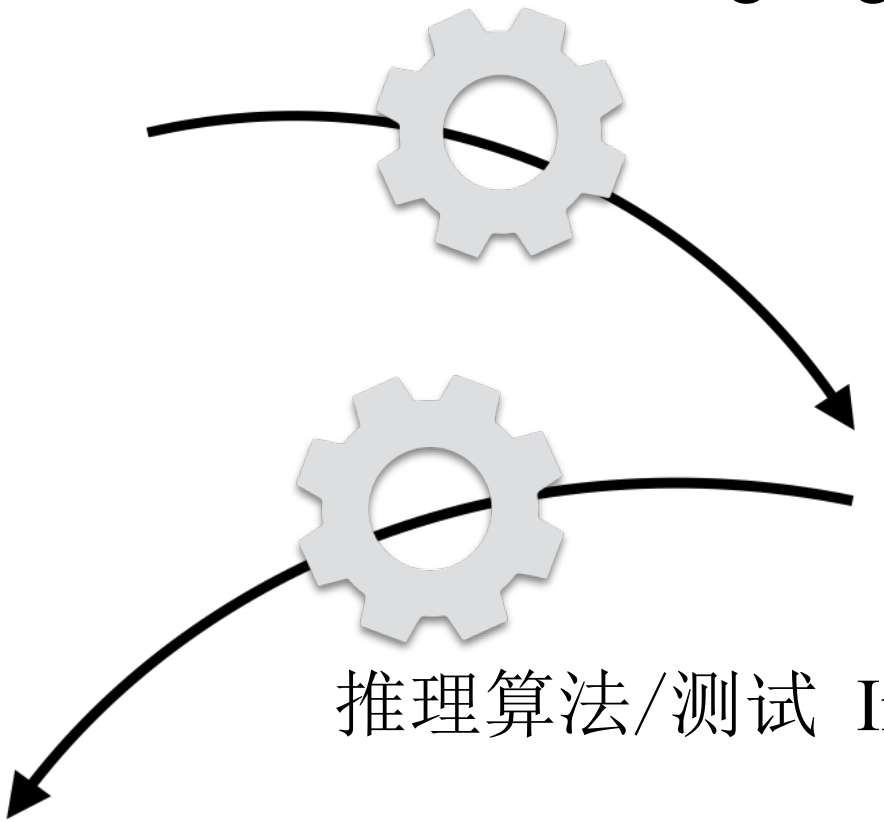


- **训练集(training set)**：通常数据集较大，在系统设计、创建和参数学习时使用
- **开发集/验证集(dev/validation set)**：用于测试不同设计决策(获得超参数)的较小数据集
- **测试集(test set)**：反映最终测试场景的数据集，用于评估模型性能

复习：文本分类/Text Classification



学习算法/训练 Learning Algorithm/Training



可学习的特征提取器 f
打分函数权值 $\mathbf{w}$

$$\mathbf{h} = F(\mathbf{x})$$
$$s = \text{category}(\mathbf{w} \cdot \mathbf{h}) = f(\mathbf{h})$$

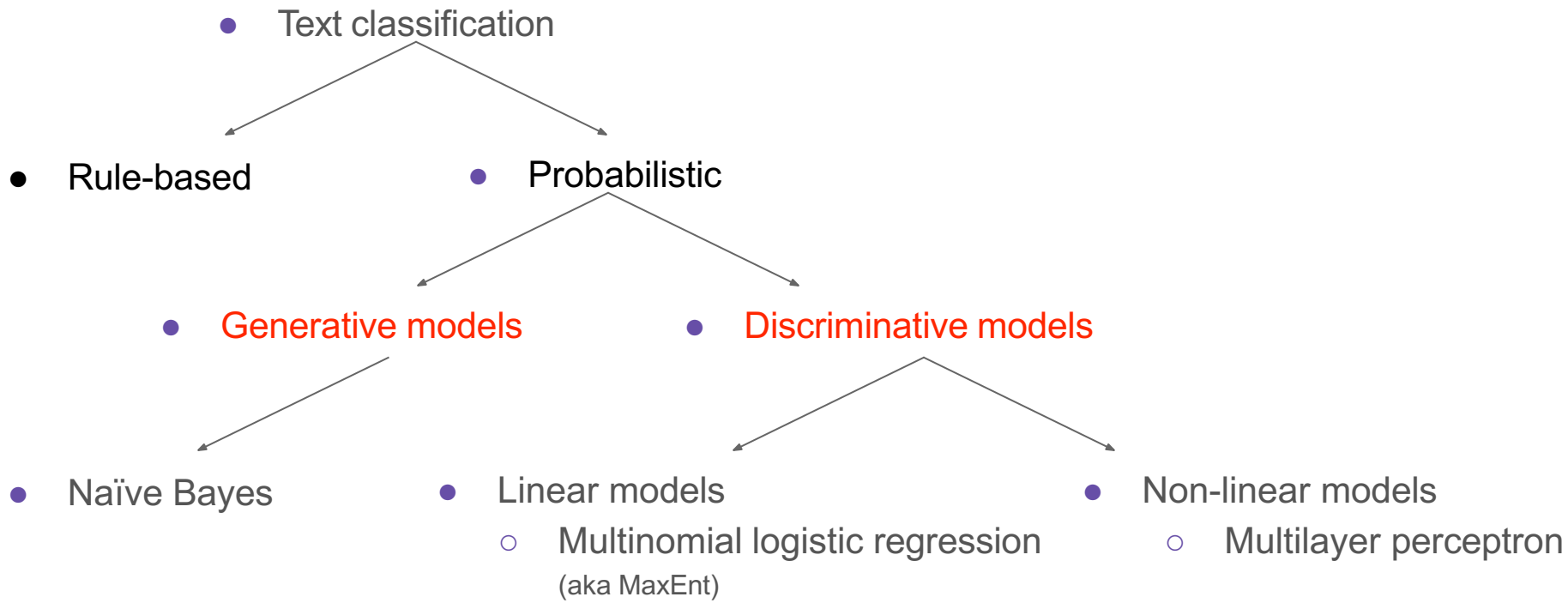
推理算法/测试 Inference Algorithm/Test

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

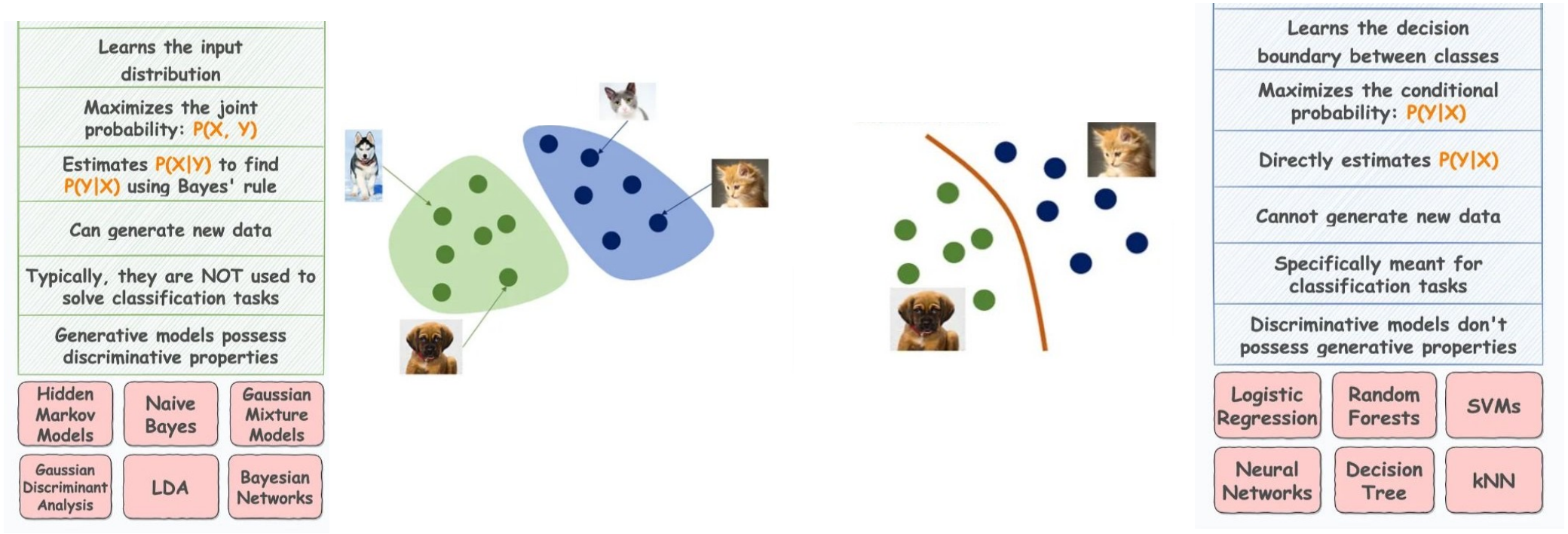
文本分类模型


Text classification
Rule-based
Probabilistic
Generative models
Discriminative models
Naïve Bayes
Linear models
Multinomial logistic regression
(aka MaxEnt)
Non-linear models
Multilayer perceptron

生成式模型和判别式模型
Generative vs. Discriminative Models

- 生成式模型：计算输入数据本身的概率的模型
  - $P(X)$:独立概率
  - $P(X,Y)$:联合概率
- 判别式模型：在给定输入数据的情况下计算输出概率的模型
  - $P(Y|X)$：条件概率

生成式模型和判别式模型



https://blog.dailydoseofds.com/p/an-intuitive-guide-to-generative

https://medium.com/@jordi299/about-generative-and-discriminative-models-d8958b67ad32

生成式模型和判别式模型

- 生成文本分类：学习模型获得联合概率函数$P(X,Y)$，并确定
  - $\hat{y} = \underset{\tilde{y}}{\operatorname{argmax}} P(X,\tilde{y})$
- 判别式文本分类：学习模型获得条件概率函数$P(Y|X)$，并确定
  - $\hat{y} = \underset{\tilde{y}}{\operatorname{argmax}} P(\tilde{y}|X)$

生成式文本分类和判别式文本分类

目标:从文档d中找到正确的类c

- 朴素贝叶斯 Naive Bayes
  - $\hat{c} = \underset{c \in Y}{\operatorname{argmax}} P(d, c)$
  - $= \underbrace{P(d|c)}_{\text{似然 Likelihood}} \cdot \underbrace{P(c)}_{\text{先验概率 Prior}}$
- 逻辑回归 Logistic Regression
  - $\hat{c} = \underset{c \in Y}{\operatorname{argmax}} \underbrace{P(c|d)}_{\text{后验概率 Posterior}}$

生成式文本分类：朴素贝叶斯

- 简单分类法
  - 基于贝叶斯规则
- 依赖于非常简单(朴素)的文档表示
  - 条件独立假设
  - 给定目标类，特性是条件独立的（因此得名“naïve”）
  - 词袋模型，没有相对顺序
- 对于更复杂的模型来说，这是一个很好的基线/baseline

Andrew Y. Ng and Michael I. Jordan, On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes, Advances in Neural Information Processing Systems 14 (NIPS), 2001.

朴素贝叶斯

情感分析：电影评论

- 给定一份文档d（例如，一篇电影评论）
- 确定它属于哪个类c：
  - 如：正向的，负向的，中性的,
  - 即 positive/negative /neutral

- 对每个c,计算$P(c|d)$
  - P(“正向的”|d)， P(“负向的”|d)， P(“中性的”|d)
  - 选择对应P值最大的类，作为样本的类别

朴素贝叶斯

- 给定文档d和类c，使用贝叶斯规则：

$$\underbrace{P(c|d)}_{\text{后验概率 Posterior}} = \frac{P(d|c)P(c)}{P(d)}$$

朴素贝叶斯

- 给定文档d和类c，使用贝叶斯规则：

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

$$\underbrace{P('\text{positive}'|d)}_{\text{后验概率 Posterior}} \propto \underbrace{P(d|'\text{positive}')}_{\text{似然 Likelihood}} \underbrace{P('\text{positive}')}_{\text{先验概率 Prior}}$$

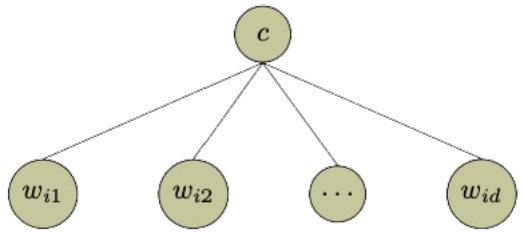
朴素贝叶斯

- 给定文档d和类c，使用贝叶斯规则：

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

$$\underbrace{P('\text{positive}'|d)}_{\text{后验概率 Posterior}} \propto \underbrace{P(d|'\text{positive}')}_{\text{似然 Likelihood}} \underbrace{P('\text{positive}')}_{\text{先验概率 Prior}}$$

朴素贝叶斯：独立假设



- 词袋假设：假设位置不影响
- 条件独立：给定类c，假设特征概率$P(w_i|c)$是独立的

$$P(d|c) = P(w_1, w_2, \cdots, w_n|c)$$

$$= P(w_1|c) \times P(w_2|c) \times \cdots \times P(w_n|c)$$

文档表示

I love this movie! It's sweet, but with satirical humor. The dialogue is great, and the adventure scenes are fun… It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it just to about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it before.

词袋模型

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

$$P(d|c) = P(w_1, w_2, \cdots, w_n|c) = \prod_i P(w_i|c)$$

生成式文本分类：朴素贝叶斯

- $C_{NB} = \underset{c}{\text{argmax}}\, P(c|d) = \underset{c}{\text{argmax}} \frac{P(d|c)P(c)}{P(d)}$

$$\propto \underset{c}{\text{argmax}}\, P(d|c)P(c)$$

$$= \underset{c}{\text{argmax}}\, P(w_1, w_2, \cdots, w_n|c)P(c)$$

$$= \underset{c}{\text{argmax}}\, P(c)\prod_i P(w_i|c)$$

- 贝叶斯规则
- 相同分母
- 词袋表征
- 条件独立假设

防止下溢：log变换

- 问题：将大量概率值相乘可能导致浮点下溢
- 因为$\log(xy) = \log(x) + \log(y)$
  - 最好将概率的对数求和，而不是将概率值相乘
- 具有未归一化的对数概率得分的类仍然是最有可能的分类

$$C_{NB} = \underset{c}{\operatorname{argmax}} P(c) \prod_i P(w_i|c)$$

$$C_{NB} = \underset{c}{\operatorname{argmax}} \log P(c) \sum_i \log P(w_i|c)$$

- 模型由权重和的最大值决定

多项式朴素贝叶斯模型的训练
Multinomial Naïve Bayes

$$C_{NB} = \underset{c}{\text{argmax}} \log P(c) \sum_i \log P(w_i|c)$$

- 通过(标注)训练数据, 学习$P(c)$和$P(w_i|c)$
- 将类别c的所有文档连接到一个超级文档中
- 使用超级文档中w_i的频率来估计单词的概率

$$\hat{P}(c_j) = \frac{count(c = c_j)}{N_{doc}} \qquad \hat{P}(w_i|c_j) = \frac{count(w_i, c_j)}{\sum_{w \in V} count(w, c_j)}$$

极大似然估计中遇到的问题

$$C_{NB} = \underset{c}{\text{argmax}}\, P(c) \prod_i P(w_i|c)$$

- 某个词没有出现在训练样本中
  - 如：正向样本中没有出现“fantastic”这个词
- 会出现什么问题？

$$\hat{P}(\text{“fantistic”}|c = \text{positive}) = \frac{count(\text{“fantistic”}, \text{positive})}{\sum_{w \in V} count(w, \text{positive})} = 0$$

- 该类概率为0

解决方案：拉普拉斯(加1)平滑

$$\hat{P}(w_i|c_j) = (1 - \lambda_{\text{unk}})P(w_i|c_j) + \lambda_{\text{unk}}P(w_i)$$

- 假设：
  - $P(w_i) = \frac{1}{|V|}$

$$\Longrightarrow \hat{P}(w_i|c_j) = \frac{count(w_i, c_j) + 1}{\sum_{w \in V} count(w, c_j) + |V|}$$

多项式朴素贝叶斯模型的训练

从训练语料库中提取词汇

- 计算$\hat{P}(c_j)$
- For each c_j do
  - $docs_j$←所有类别为c_j的文档
  - $\hat{P}(c_j) = \frac{count(c=c_j)}{N_{doc}}$

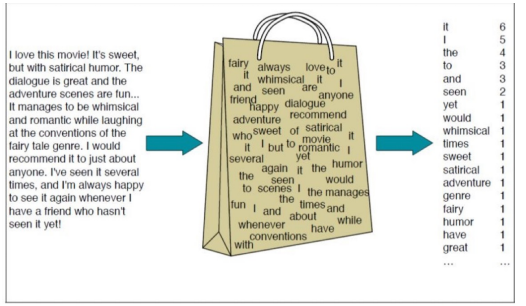
- 计算$\hat{P}(w_i|c_j)$
  - $Text_j$←包含所有$docs_j$的单个文档
  - 对于词汇表中的每个单词
    - n_i←w_i在文本中出现的次数
    - $\hat{P}(w_i|c_j) \leftarrow \frac{n_i+\alpha}{n+\alpha|V|}$

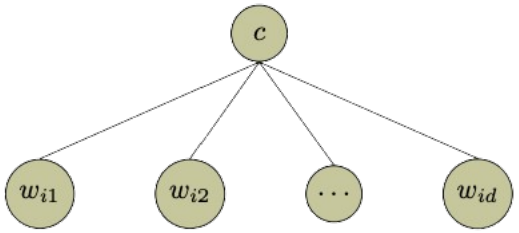
In-class Practice I

- 已知下面训练数据，求测试集中的文档类别

	Doc	Content	Cat.
Training	1	China Beijing China	C
	2	China China Shanghai	C
	3	China Hongkong	C
	4	Japan Tokyo China	J
Test	5	China China China Tokyo Japan	?

不同方案的效果





- 基于规则

```
Train accuracy: 0.4345739700374532
Dev/test accuracy: 0.4214350590372389
```

- 基于BOW & Structure Perceptron

```
Train accuracy: 0.7332631086142322
Dev/test accuracy: 0.5676657584014533
```

- 基于BOW & Naïve Bayes

```
Train accuracy: 0.858497191011236
Dev/test accuracy: 0.631244323342416
```

https://github.com/hqyang/nlp-codes/tree/main/01-simpleclassifier

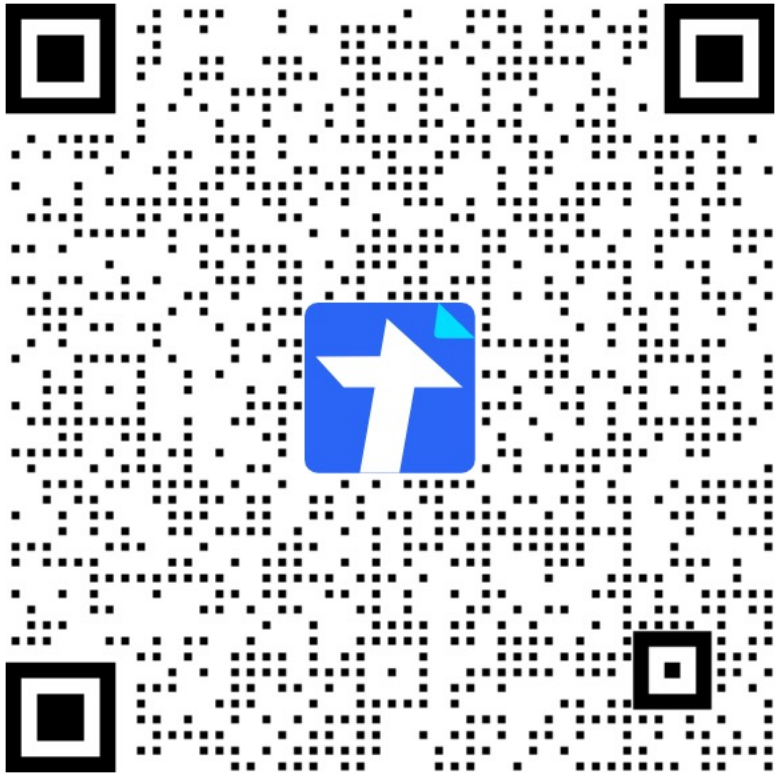
小结：朴素贝叶斯不是那么朴素

- 朴素贝叶斯是一个概率模型
  - 起名源于假设：一个类的特征是相互独立的
- 优点
  - 非常快，低存储要求
  - 对不相关特征的鲁棒性
    - 不相关的特征相互抵消亦不影响结果
  - 在具有许多同等重要特征的领域中表现很出色(如垃圾邮件过滤)
    - 其他模型(如决策树)在数据不足时，受到碎片化影响
  - 独立性假设成立时模型最优:如果独立性假设成立，那么它就是该问题的最优贝叶斯分类器
  - 文本分类一个很可靠的基线
    - 后续会看到其他分类器获得更好的性能

一句话总结

- 文本分类的应用场景
- 文本分类的定义及完整流程
  - 特征抽取的主要方式
  - 文本分类模型的两大分类
- 朴素贝叶斯
  - 核心假设
  - 贝叶斯公式
  - 对数转换
  - 拉普拉斯平滑

1. 作业1(ddl: Oct. 19, 23:59:00)
2. 项目分组 (ddl: Oct. 19, 23:59:00)
3. 课外阅读
   - [SLP3]Ch.4-5
   - [E]Ch.2



Ans. to In-class Practice I

	Doc	Content	Cat.
Training	1	China Beijing China	C
	2	China China Shanghai	C
	3	China Hongkong	C
	4	Japan Tokyo China	J
Test	5	China China China Tokyo Japan	?

- Prior
  - $P(C) = 3/4$
  - $P(J) = 1/4$
- Conditional Probabilities:
  - $P(\text{“China”}|C)=(5+1)/(8+6)=6/14=3/7$
  - $P(\text{“Tokyo”}|C)=(0+1)/(8+6)=1/14$
  - $P(\text{“Japan”}|C)=(0+1)/(8+6)=1/14$
  - $P(\text{“China”}|J)=(1+1)/(3+6)=2/9$
  - $P(\text{“Tokyo”}|J)=(1+1)/(3+6)=2/9$
  - $P(\text{“Japan”}|J)=(1+1)/(3+6)=2/9$
- Choose a class:
  - $P(C|d_5) \propto 3/4 * (3/7)^3 * 1/14 * 1/14 \approx 0.0003$
  - $P(J|d_5) \propto 1/4 * (2/9)^3 * 2/9 * 2/9 \approx 0.0001$
- C is chosen.