自然语言处理

3. 文本分类

虞剑飞

南京大学智能科学与技术学院

2025.3.5

本章内容

- 3.1 引言
- 3.2 基于传统机器学习的文本分类方法
  - 3.2.1 文本表示
  - 3.2.2 特征选择
  - 3.2.3 分类算法
- 3.3 文本分类性能评估
  - 3.3.1 准确率、宏平均/微平均的召回率、精确率和F_1值
  - 3.3.2 P-R曲线
  - 3.3.3 ROC曲线、AUC

引言

- 真实生活中的文本分类问题

新闻分类

情感分类

引言

- 文本分类定义
  - 将给定的文本文档或句子分类为预定义的类别
    - 单标签多类别文本分类
    - 多标签多类别文本分类

单标签多类别文本分类

7月5日，周杰伦新专辑《最伟大的作品》开启预约，目前，QQ音乐上已经有超过270万人参与了预约。据悉，最伟大的作品》7月6日迎来MV首播，7月8日会开启专辑预售，7月15日专辑将正式上线。

多标签多类别文本分类

7月5日，日本相机暨影像产品协会最新的统计报告出炉，今年5月份全球数码相机出货量相较去年同期下跌63.7万台，约10.4%受全球芯片缺货的影响，数码相机已经连续10个月出现市场萎缩迹象，月总出货量连续18个月不足百万台。

引言

- 文本分类定义
  - 将给定的文本文档或句子分类为预定义的类别
    - 单标签多类别文本分类

引言

- 文本分类
  - 基准公开数据集

Dataset	Type	Number of labels	Size (train/test)	Avg. length (tokens)
SST	sentiment	5 or 2	8.5k / 1.1k	19
IMDb Review	sentiment	2	25k / 25k	271
Yelp Review	sentiment	5 or 2	650k / 50k	179
Amazon Review	sentiment	5 or 2	3m / 650k	79
TREC	question	6	5.5k / 0.5k	10
Yahoo! Answers	question	10	1.4m / 60k	131
AG's News	topic	4	120k / 7.6k	44
Sogou News	topic	6	54k / 6k	737
DBPedia	topic	14	560k / 70k	67

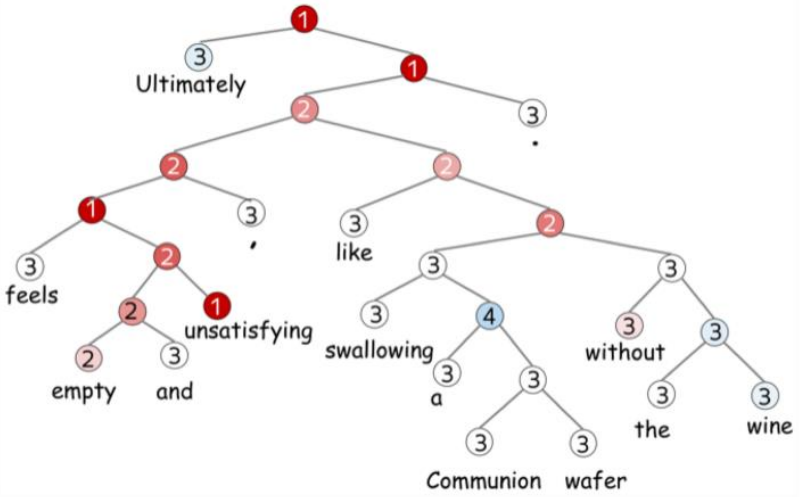
引言

- 文本分类
  - 基准公开数据集

Pick a dataset

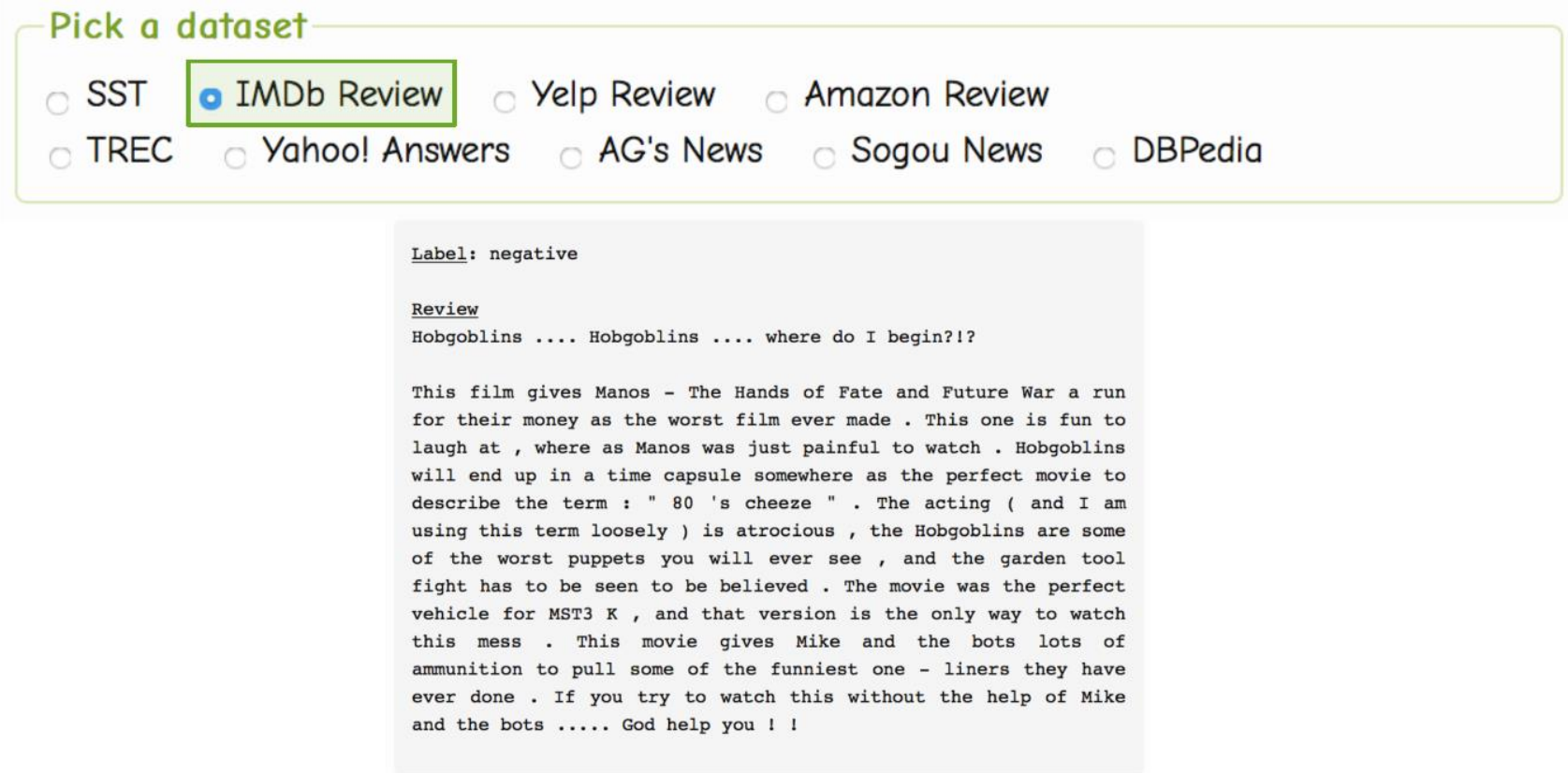
SST　IMDb Review　Yelp Review　Amazon Review
TREC　Yahoo! Answers　AG's News　Sogou News　DBPedia

Label: 1
Review:
Ultimately feels empty and unsatisfying , like swallowing a Communion wafer without the wine .



引言

- 文本分类
  - 基准公开数据集

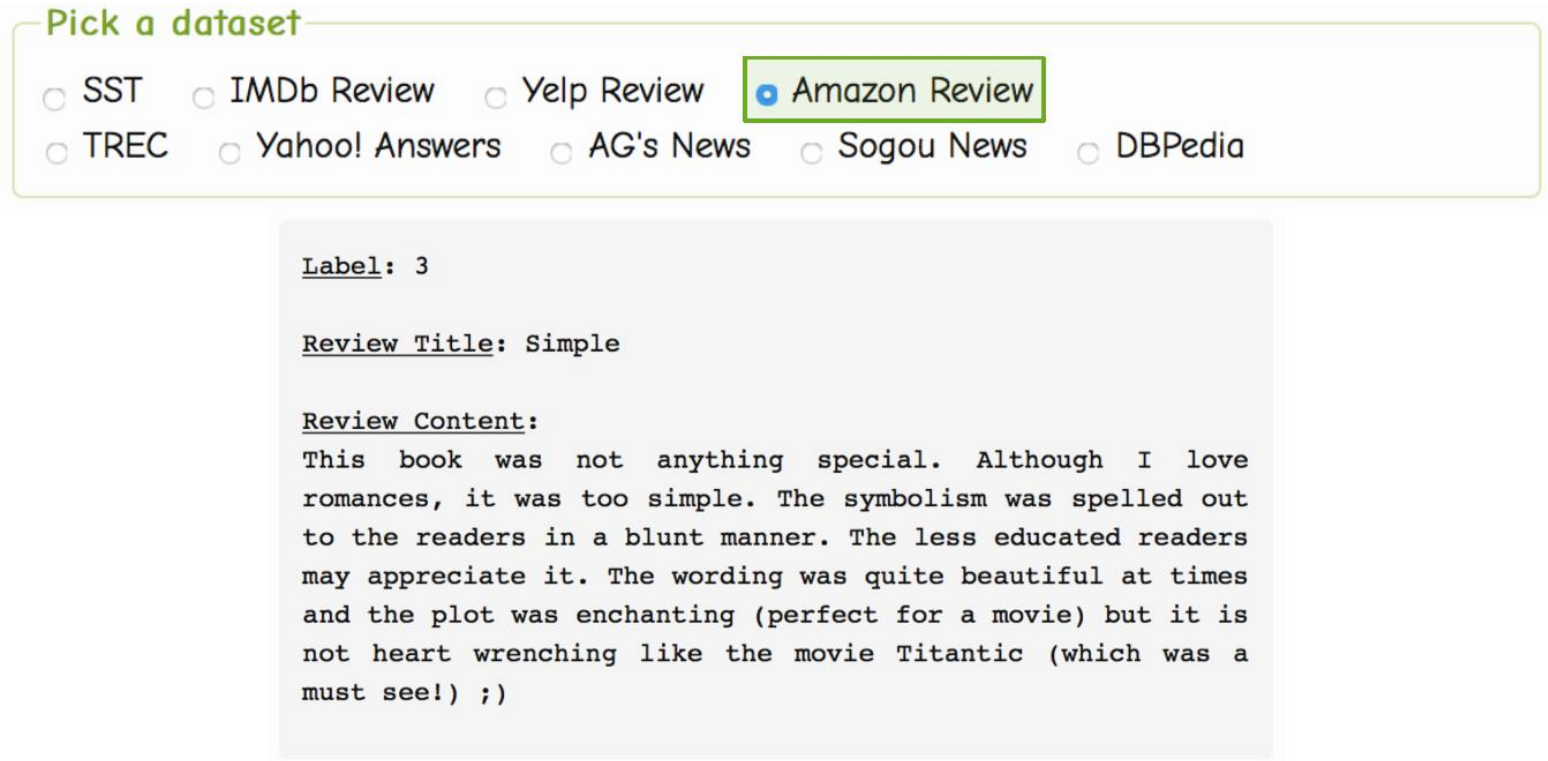


引言

- 文本分类
  - 基准公开数据集

引言

- 文本分类
  - 基准公开数据集



引言

- 文本分类
  - 基准公开数据集

Label: Society & Culture

Question Title: Why do people have the bird, turkey for thanksgiving?

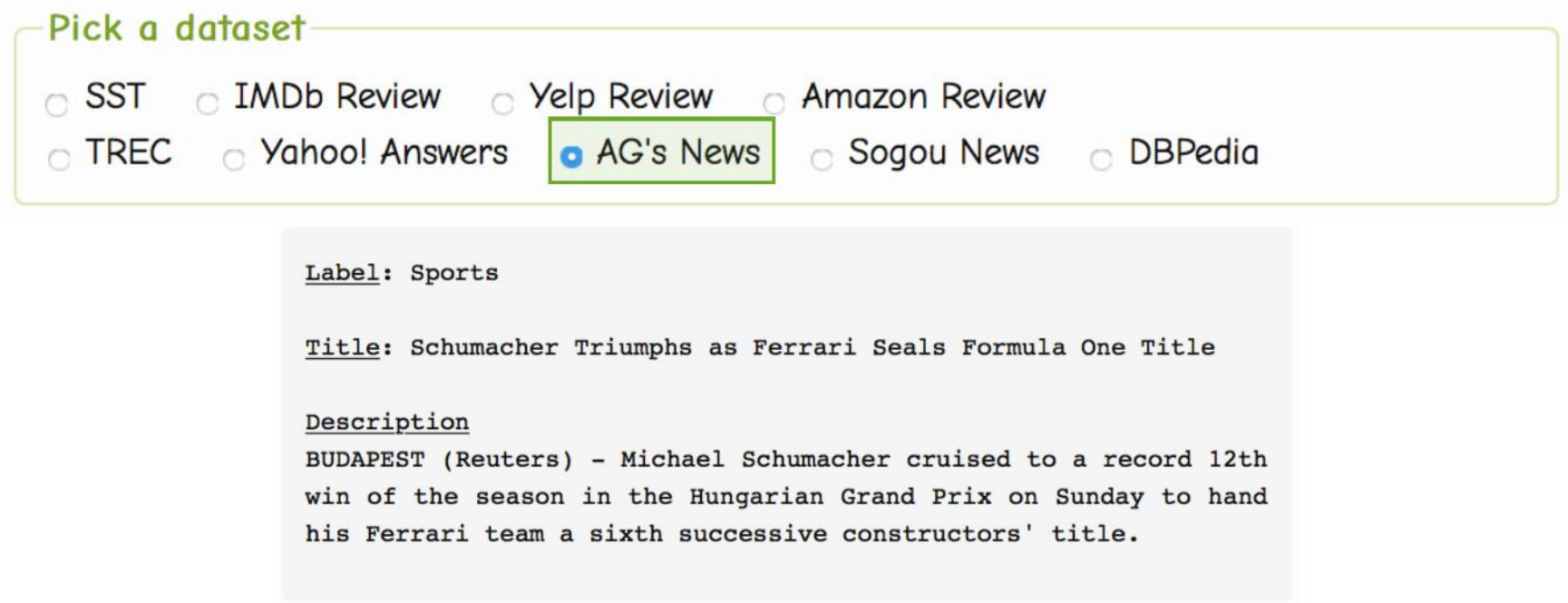
Question Content: Why this bird? Any Significance?

Best Answer
It is believed that the pilgrims and indians shared wild turkey and venison on the original Thanksgiving.

Turkey's "Americanness" was established by Benjamin Franklin, who had advocated for the turkey, not the bald eagle, becoming the national bird.

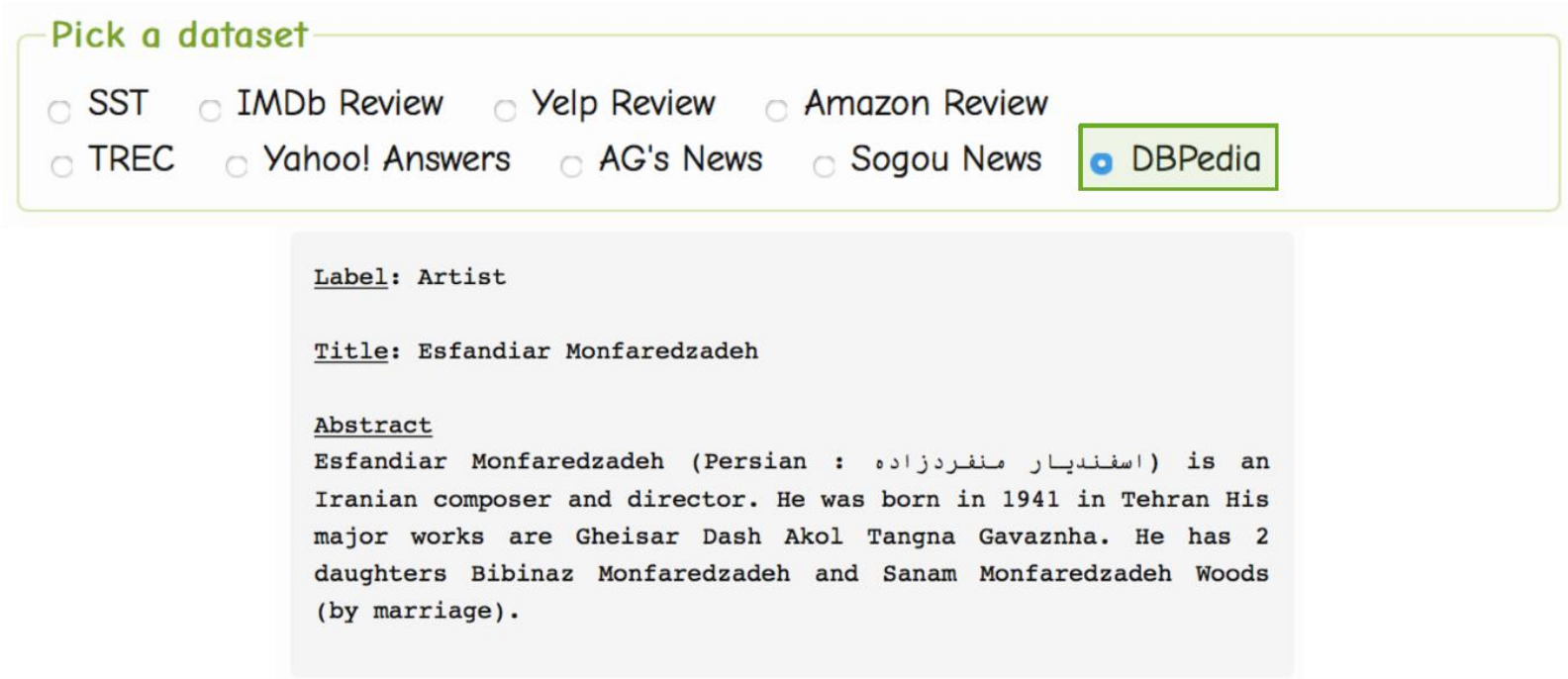
引言

- 文本分类
  - 基准公开数据集



引言

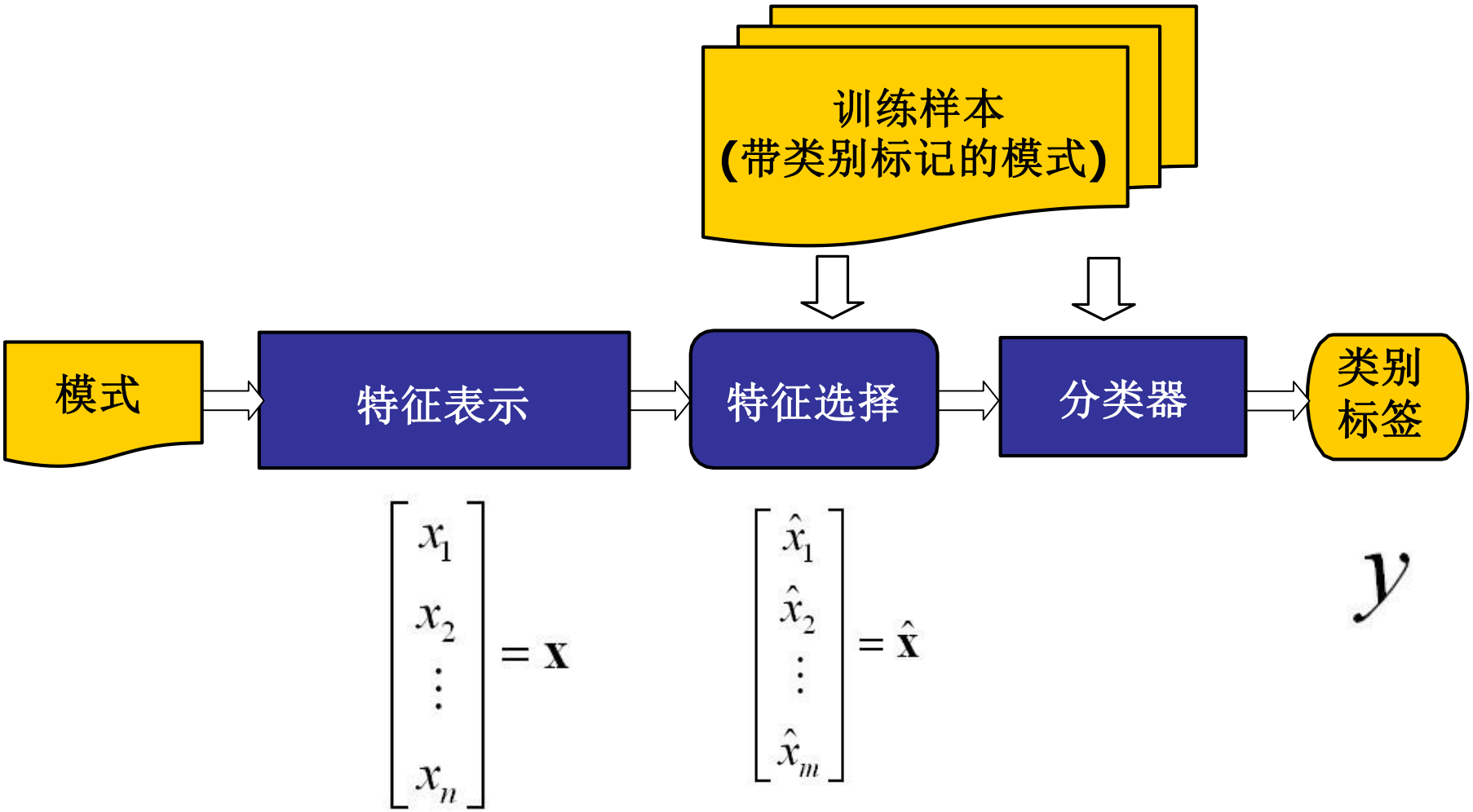
- 文本分类
  - 基准公开数据集



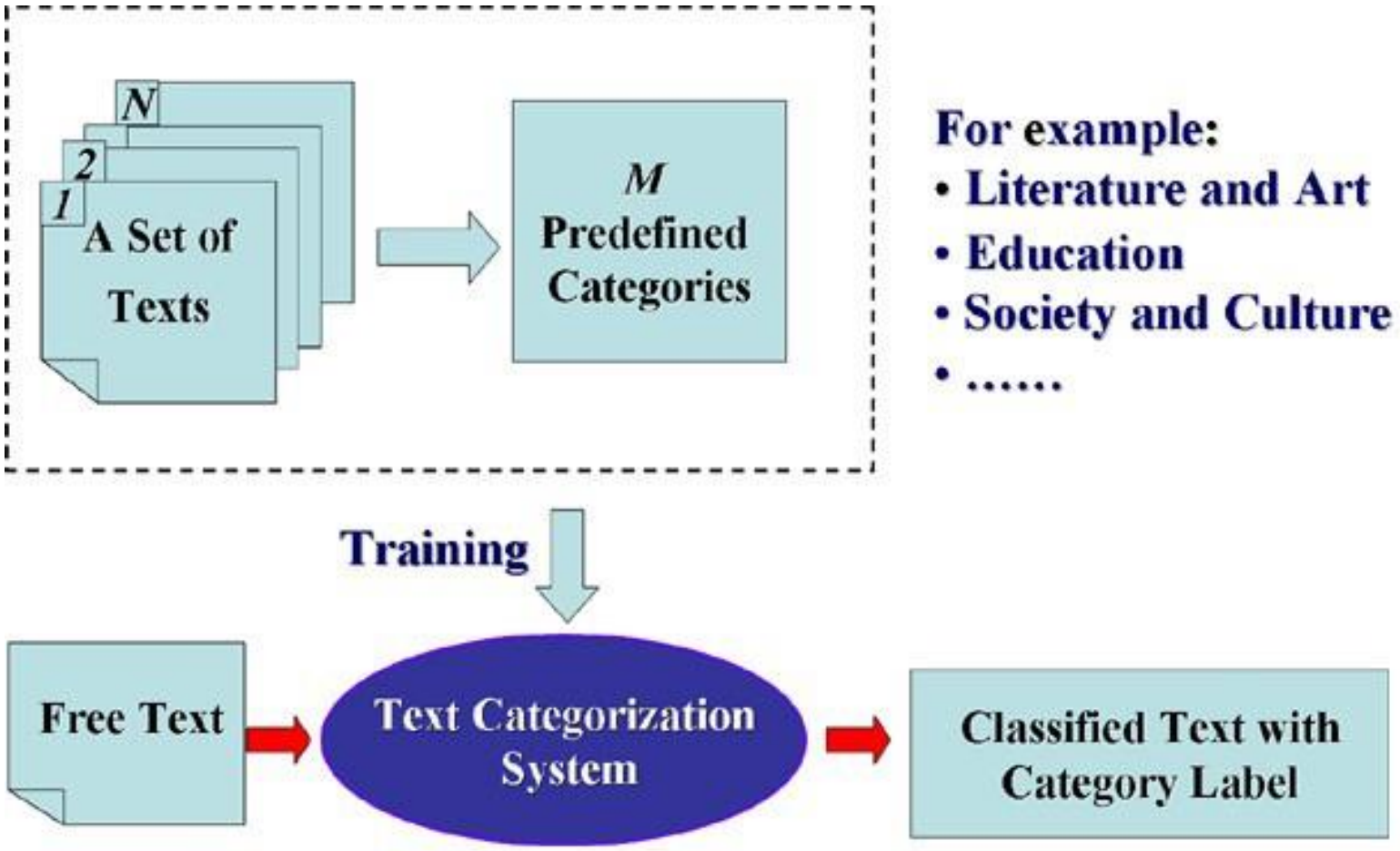
本章内容

- 3.1 引言
- 3.2 基于传统机器学习的文本分类方法
  - 3.2.1 文本表示
  - 3.2.2 特征选择
  - 3.2.3 分类算法
- 3.3 文本分类性能评估
  - 3.3.1 准确率、宏平均/微平均的召回率、精确率和F_1值
  - 3.3.2 P-R曲线
  - 3.3.3 ROC曲线、AUC

传统机器学习方法


训练样本
(带类别标记的模式)
模式
特征表示
特征选择
分类器
类别
标签
$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \mathbf{x}$
$\begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \vdots \\ \hat{x}_m \end{bmatrix} = \hat{\mathbf{x}}$
y

传统机器学习方法


N
2
1
A Set of Texts
M
Predefined Categories
For example:
• Literature and Art
• Education
• Society and Culture
• ……
Training
Free Text
Text Categorization System
Classified Text with Category Label

本章内容

- 3.1 引言
- 3.2 基于传统机器学习的文本分类方法
  - 3.2.1 文本表示
  - 3.2.2 特征选择
  - 3.2.3 分类算法
- 3.3 文本分类性能评估
  - 3.3.1 准确率、宏平均/微平均的召回率、精确率和F_1值
  - 3.3.2 P-R曲线
  - 3.3.3 ROC曲线、AUC

背景

- 文本
  - 文本是由文字和标点组成的字符串，短语、句子、段落和篇章都是不同粒度的文本
  - 字或字符组成词、短语，进而形成句子、段落和篇章

- 文本表示的必要性
  - 计算机进行文本理解，必须知道文本长什么样
  - 文本的形式化表示是反映文本内容和区分不同文本的有效途径

向量空间模型

- 基本概念
  - 向量空间模型（vector space model, VSM）由G. Salton 等人于1960s末期在信息检索领域提出，核心是将文本视为特征项的集合
  - 特征项：VSM中最小的语言单元，可以是字、词、短语等。文本表示为特征项集合 $(t_1, t_2, \ldots, t_n)$

Document	the	cat	sat	in	hat	with
the cat sat						
the cat sat in the hat						
the cat with the hat						

向量空间模型

- 基本概念
  - 向量空间模型（vector space model, VSM）由G. Salton 等人于1960s末期在信息检索领域提出，核心是将文本视为特征项的集合
  - 特征项：VSM中最小的语言单元，可以是字、词、短语等。文本表示为特征项集合 $(t_1, t_2, \ldots, t_n)$

Document	the_cat	cat_sat	sat_in	in_the	the_hat	cat_with	with_the
the cat sat							
the cat sat in the hat							
the cat with the hat							

向量空间模型

- 特征项
  - 词语（词组或短语）：若词语作为特征项，那么特征项的集合可视为一个词表。词表可从语料中统计获得， 可看作一个词袋， 向量空间模型被称为词袋模型（bag-of-words, BOW）

一个文本表示的例子

- 训练数据（带类别标签的文档）

教育	体育
北京 理工 大学 计算机 专业 创建 于 1958 年 是 中国 最早 设立 计算机 专业 的 高校 之一	北京 理工 大学 体育馆 是 2008 年 中国 北京 奥林匹克 运动会 的 排球 预赛 场地
北京 理工 大学 学子 在 第四 届 中国 计算机 博弈 锦标赛 中 夺冠	第五 届 东亚 运动会 中国 军团 奖牌 总数 创 新高 男女 排球 双双 夺冠

一个文本表示的例子

- 词袋表示（含40个词，即词表大小为40）
  - 假设只考虑将单个词作为特征项

1958 2008 奥林匹克 北京 博弈 场地 创 创建 大学 的 第四 第五 东亚 夺冠 高校 计算机 奖牌 届 锦标赛 军团 理工 男女 年 排球 设立 是 双双 体育馆 新高 学子 于 预赛 运动会 在 之一 中 中国 专业 总数 最早

向量空间模型

- 基本概念
  - 特征项权重：每个特征项在文本中的重要性不尽相同，用w表示特征项t的权重，相应地，文本可以表示为 $(t_1:w_1, t_2:w_2, \dots, t_n:w_n)$ 或 $(w_1, w_2, \dots, w_n)$

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1

向量空间模型

- 基本概念
  - 特征项权重：每个特征项在文本中的重要性不尽相同，用w表示特征项t的权重，相应地，文本可以表示为 $(t_1:w_1, t_2:w_2, \dots, t_n:w_n)$ 或 $(w_1, w_2, \dots, w_n)$

Document	the_cat	cat_sat	sat_in	in_the	the_hat	cat_with	with_the
the cat sat	1	1	0	0	0	0	0
the cat sat in the hat	1	1	1	1	1	0	0
the cat with the hat	1	0	0	0	1	1	1

向量空间模型

- 特征项的权重
  - 如何计算每个特征项对应的权重？
    - 布尔变量（是否出现）

$$w_i = \begin{cases} 1, & \text{如果}t_i\text{在文本}d\text{中} \\ 0, & \text{否则} \end{cases}$$

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	1	1	1	1	1	0
the cat with the hat	1	1	0	0	1	1

向量空间模型

- 特征项的权重
  - 如何计算每个特征项对应的权重？
    - 布尔变量（是否出现）

$$w_i = \begin{cases} 1, & \text{如果} t_i \text{在文本} d \text{中} \\ 0, & \text{否则} \end{cases}$$

    - 词频（Term Frequency, TF）

$$w_i = \log(tf_i + 1)$$

Document	the	cat	sat	in	hat	with
the cat sat	1	1	1	0	0	0
the cat sat in the hat	2	1	1	1	1	0
the cat with the hat	2	1	0	0	1	1

向量空间模型

- 特征项的权重
  - 如何计算每个特征项对应的权重？
    - 布尔变量（是否出现）

$$w_i = \begin{cases} 1, & \text{如果}t_i\text{在文本}d\text{中} \\ 0, & \text{否则} \end{cases}$$

    - 词频（Term Frequency, TF）

$$w_i = \log(tf_i + 1)$$

Document	the	cat	sat	in	hat	with
the cat sat	log2	log2	log2	0	0	0
the cat sat in the hat	log3	log2	log2	log2	log2	0
the cat with the hat	log3	log2	0	0	log2	log2

向量空间模型

- 特征项的权重
  - 如何计算每个特征项对应的权重？
    - 布尔变量（是否出现）

$$w_i = \begin{cases} 1, & \text{如果} t_i \text{在文本} d \text{中} \\ 0, & \text{否则} \end{cases}$$

    - 词频（Term Frequency, TF）

$$w_i = \log(tf_i + 1)$$

    - 逆文档频率（Inverse Document Frequency, IDF）

$$w_i = idf_i = \log\frac{N}{df_i}$$

    - TF-IDF

$$tf\_idf_i = tf_i \cdot idf_i$$

向量空间模型

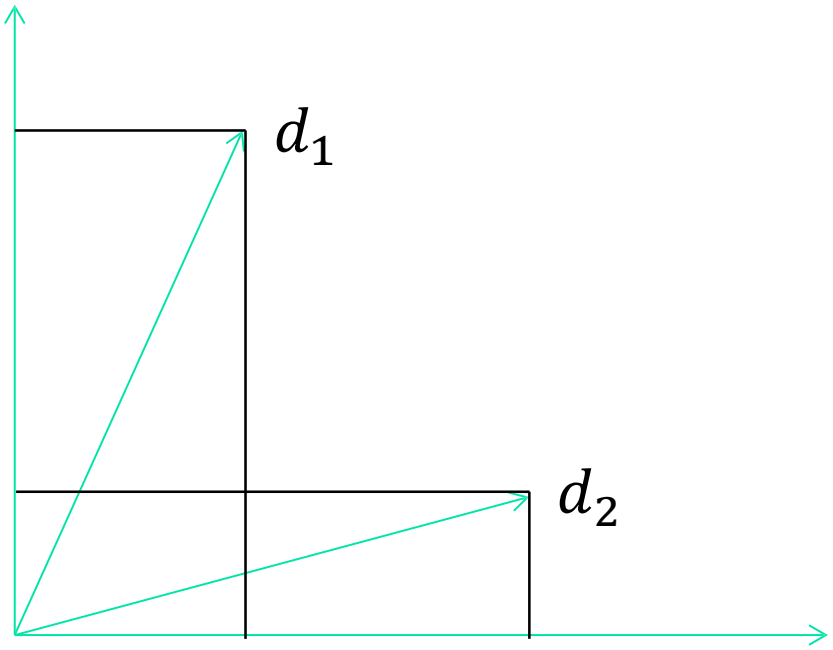
- 举例

文本1

$d_1 = (w_1, w_2, \cdots, w_n)$

文本2

$d_2 = (w'_1, w'_2, \cdots, w'_n)$



本章内容

- 3.1 引言
- 3.2 基于传统机器学习的文本分类方法
  - 3.2.1 文本表示
  - 3.2.2 特征选择
  - 3.2.3 分类算法
- 3.3 文本分类性能评估
  - 3.3.1 准确率、宏平均/微平均的召回率、精确率和F_1值
  - 3.3.2 P-R曲线
  - 3.3.3 ROC曲线、AUC

本章内容

- 3.1 引言
- 3.2 基于传统机器学习的文本分类方法
  - 3.2.1 文本表示
  - 3.2.2 特征选择
    - 文档频率（Document Frequency, DF）
    - 互信息（Mutual Information, MI）
    - 信息增益（Information Gain, IG）
  - 3.2.3 分类算法
- 3.3 文本分类性能评估
  - 3.3.1 准确率、宏平均/微平均的召回率、精确率和F_1值
  - 3.3.2 P-R曲线
  - 3.3.3 ROC曲线、AUC

特征选择-文档频率

- 文档频率（Document Frequency, DF）
  - 一个特征的文档频率是指在文档集中含有该特征的文档数目
  - 基本假设:
    - DF值低于某个域值的词条是低频词，它们不含或含有较少的类别信息
    - 将这样的词条从原始特征空间中除去，不但能够降低特征空间的维数,而且还有可能提高分类的精度
  - 出现文档数多的特征词被保留的可能性大

特征选择-文档频率

- 文档频率（Document Frequency, DF）
  - 用特征词在一个类别中出现的文档数表示这个特征词与该类别的相关度
  - 举例：文本集合A关于特征 t_i 与类别 c_j 的统计表

类别 / 特征	c_j	$\overline{c_j}$
t_i	A_{ij}	B_{ij}
$\overline{t_i}$	C_{ij}	D_{ij}

A_{ij}: 类别 c_j 的文档中出现特征t_i的文档数;

B_{ij}: 类别 $\overline{c_j}$ 的文档中出现特征t_i的文档数;

C_{ij}: 类别 c_j 的文档中未出现特征t_i的文档数;

D_{ij}: 类别 $\overline{c_j}$ 的文档中未出现特征t_i的文档数;

$$P(c_j) \approx (A_{ij} + C_{ij})/N_{all}$$

$$P(t_i) \approx (A_{ij} + B_{ij})/N_{all}$$

$$P(\overline{t_i}) \approx (C_{ij} + D_{ij})/N_{all}$$

$$P(c_j|t_i) \approx \frac{A_{ij} + 1}{A_{ij} + B_{ij} + C}$$

$$P(c_j|\overline{t_i}) \approx \frac{C_{ij} + 1}{C_{ij} + D_{ij} + C}$$

特征选择-文档频率

- 文档频率（Document Frequency, DF）
  - 举例：一共有20篇文档(N_{all}=20)，分为两类：教育类(c_1)和非教育类($\overline{c_1}$)。其中，教育类7篇(c_1=7)，非教育类13篇($\overline{c_1}$=13)。特征词t_2=“计算机”是教育类文档的特征之一。教育类的7篇文档中出现特征词“计算机”的个数，假设有5篇，那么教育类文档中没出现特征词“计算机”的文档数为2篇。同时，在非教育类的13篇文档中，假设有5篇出现了特征词“计算机”。

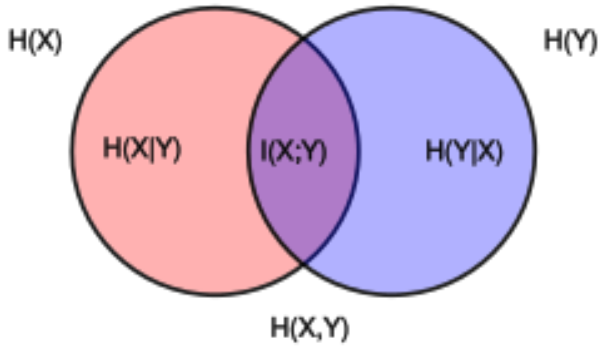
类别 / 特征	c_j	$\overline{c_j}$
t_i	$A_{12}=5$	$B_{12}=5$
$\overline{t_i}$	$C_{12}=2$	$D_{12}=8$

本章内容

- 3.1 引言
- 3.2 基于传统机器学习的文本分类方法
  - 3.2.1 文本表示
  - 3.2.2 特征选择
    - 文档频率（Document Frequency, DF）
    - 互信息（Mutual Information, MI）
    - 信息增益（Information Gain, IG）
  - 3.2.3 分类算法
- 3.3 文本分类性能评估
  - 3.3.1 准确率、宏平均/微平均的召回率、精确率和F_1值
  - 3.3.2 P-R曲线
  - 3.3.3 ROC曲线、AUC

特征选择-互信息

- 相关信息论概念



  - 熵（Entropy）

$$H(X) = -\sum_{x} p(x) \log p(x)$$

  - 联合熵（Joint Entropy）

$$H(X,Y) = -\sum_{x}\sum_{y} p(x,y) \log p(x,y)$$

  - 条件熵（Conditional Entropy）

$$H(Y \mid X) = \sum_x p(x) H(Y|X = x) = -\sum_x \sum_y p(x,y) \log p(y|x)$$
$$H(Y \mid X) = H(X,Y) - H(X)$$

特征选择-互信息

- 互信息（Mutual Information，MI）
  - 互信息是关于两个随机变量互相依赖程度的一种度量

$$I(X,Y) = H(X) - H(X \mid Y) = \sum_y \sum_x P(x,y) \log \frac{P(x,y)}{P(x)P(y)}$$

$$MI(t_i, c_j) = \log \frac{P(t_i, c_j)}{P(t_i)P(c_j)} \approx \log \frac{A_{ij} N_{all}}{(A_{ij} + C_{ij})(A_{ij} + B_{ij})}$$

点式互信息

$$MI_{\text{avg}}(t_i) = \sum_{j=1}^{C} P(c_j) MI(t_i, c_j)$$

类别 / 特征	c_j	$\overline{c_j}$
t_i	A_{ij}	B_{ij}
$\overline{t_i}$	C_{ij}	D_{ij}

本章内容

- 3.1 引言
- 3.2 基于传统机器学习的文本分类方法
  - 3.2.1 文本表示
  - 3.2.2 特征选择
    - 文档频率（Document Frequency, DF）
    - 互信息（Mutual Information, MI）
    - 信息增益（Information Gain, IG）
  - 3.2.3 分类算法
- 3.3 文本分类性能评估
  - 3.3.1 准确率、宏平均/微平均的召回率、精确率和F_1值
  - 3.3.2 P-R曲线
  - 3.3.3 ROC曲线、AUC

特征选择-信息增益

- 信息增益 (IG)
  - IG衡量特征能够为分类系统带来多少信息
  - 特征T_i对训练数据集C的信息增益定义为集合C的经验熵$H(C)$与特征T_i给定条件下C的经验条件熵 $H(C|T_i)$之差，即: $IG(C,T_i)=H(C)-H(C|T_i)$

参阅:李航《统计机器学习》

$$IG(t_i)=\{-\sum_{j=1}^{C}P(c_j)\log P(c_j)\}+\{P(t_i)[\sum_{j=1}^{C}P(c_j\mid t_i)\log P(c_j\mid t_i)]+P(\bar{t}_i)[\sum_{j=1}^{C}P(c_j\mid \bar{t}_i)\log P(c_j\mid \bar{t}_i)]\}$$

- 信息增益与互信息的关系

$$IG(t_i)=\sum_{j=1}^{C}P(t_i,c_j)MI(t_i,c_j)+\sum_{j=1}^{C}P(\bar{t}_i,c_j)MI(\bar{t}_i,c_j)$$

一个文本表示的例子

- 训练数据（带类别标签的文档）

教育	体育
北京 理工 大学 计算机 专业 创建 于 1958 年 是 中国 最早 设立 计算机 专业 的 高校 之一	北京 理工 大学 体育馆 是 2008 年 中国 北京 奥林匹克 运动会 的 排球 预赛 场地
北京 理工 大学 学子 在 第四 届 中国 计算机 博弈 锦标赛 中 夺冠	第五 届 东亚 运动会 中国 军团 奖牌 总数 创 新高 男女 排球 双双 夺冠

“计算机”的信息增益

特征 \ 类别	教育	体育
计算机	2	0
$\overline{\text{计算机}}$	0	2

$P(\text{计算机})=1/2$　　$P(\overline{\text{计算机}})=1/2$

$P(\text{教育}|\text{计算机})=(2+1)/(2+2)=3/4$

$P(\text{体育}|\text{计算机})=1/(2+2)=1/4$

$P(\text{教育}|\overline{\text{计算机}})=1/(2+2)=1/4$

$P(\text{体育}|\overline{\text{计算机}})=(2+1)/(2+2)=3/4$

$$IG(t_i)=\{-\sum_{j=1}^{C}P(c_j)\log P(c_j)\}+\{P(t_i)[\sum_{j=1}^{C}P(c_j\mid t_i)\log P(c_j\mid t_i)]+P(\bar{t}_i)[\sum_{j=1}^{C}P(c_j\mid \bar{t}_i)\log P(c_j\mid \bar{t}_i)]\}$$

$$IG(\text{计算机})=-0.5log0.5-0.5log0.5+0.5(0.75log0.75+0.25log0.25)+0.5(0.75log0.75+0.25log0.25)=-log0.5+0.75log0.75+0.25log0.25=0.1308$$

“北京”的信息增益

- 课堂练习

信息增益的例子

- 根据信息增益的特征排序

Features	IG
计算机 排球 运动会	0.1308
1958 2008 奥林匹克 博弈 场地 创 创建 第四 第五 东亚 高校 奖牌 锦标赛 军团 男女 设立 双双 体育馆 新高 学子 于 预赛 在 之一 中 专业 总数 最早 北京 大学 理工	0.0293
的 夺冠 届 年 是 中国	0.0000

信息增益的例子

- 选择的特征

计算机 排球 运动会 高校 大学 1958 2008 奥林匹克 博弈 场地 创 创建 第四 第五 东亚 奖牌 锦标赛 军团 男女 设立 双双 体育馆 新高 学子 于 预赛 在 之一 中 专业 总数 最早 北京 理工

- 精简后的训练数据

教育	体育
大学 计算机 计算机 高校	大学 运动会 排球
大学 计算机	运动会 排球

互信息的例子

- 根据互信息的特征排序

Features	MI
计算机 排球 运动会	0.2877
1958 2008 奥林匹克 博弈 场地 创 创建 第四 第五 东亚 高校 奖牌 锦标赛 军团 男女 设立 双双 体育馆 新高 学子 于 预赛 在 之一 中 专业 总数 最早 北京 大学 理工	0.1178
的 夺冠 届 年 是 中国	0.0000

本章内容

- 3.1 引言
- 3.2 基于传统机器学习的文本分类方法
  - 3.2.1 文本表示
  - 3.2.2 特征选择
  - 3.2.3 分类算法
- 3.3 文本分类性能评估
  - 3.3.1 准确率、宏平均/微平均的召回率、精确率和F_1值
  - 3.3.2 P-R曲线
  - 3.3.3 ROC曲线、AUC

本章内容

- 3.1 引言
- 3.2 基于传统机器学习的文本分类方法
  - 3.2.1 文本表示
  - 3.2.2 特征选择
  - 3.2.3 分类算法
    - 监督学习
    - 无监督学习
    - 半监督学习
- 3.3 文本分类性能评估
  - 3.3.1 准确率、宏平均/微平均的召回率、精确率和F_1值
  - 3.3.2 P-R曲线
  - 3.3.3 ROC曲线、AUC

分类算法

- 监督学习（关于分类算法的基本概念）
  - 训练数据
    - 人工标注
  - 模型表示
    - 用参数进行建模(构建目标函数)
  - 学习算法
    - 最大似然估计，计算最大后验概率(生成式模型)
    - 梯度下降法、牛顿法(判别式模型)
  - 推断
    - 决策/预测规则

分类算法

- 监督学习
  - ✓ 生成式模型
    - ➢ 朴素贝叶斯（Naïve Bayes）
  - ✓ 判别式模型
    - ➢ 线性判别函数（Linear Discriminate Function）
    - ➢ 支持向量机（Support Vector Machine）
    - ➢ 最大熵模型（Maximum Entropy）
- 无监督
- 半监督学习

贝叶斯决策理论

- 贝叶斯理论

$$P(B \mid A) = \frac{P(A,B)}{P(A)} = \frac{P(B)P(A \mid B)}{P(A)}$$

- 贝叶斯决策理论

$$P(c_j \mid \mathbf{x}) = \frac{P(c_j, \mathbf{x})}{P(\mathbf{x})} = \frac{P(c_j)P(\mathbf{x} \mid c_j)}{P(\mathbf{x})}$$

$$c^* = \underset{j=1,\ldots,C}{\operatorname{argmax}} P(c_j \mid \mathbf{x}) = \underset{j=1,\ldots,C}{\operatorname{argmax}} P(c_j)P(\mathbf{x} \mid c_j)$$

朴素贝叶斯分类器

- 学习难点 $P(x|c_j)$ =???
- 多项式分布朴素贝叶斯假设

$$P(\boldsymbol{X} \mid c_j) \approx P([w_1, \cdots, w_n] \mid c_j) \approx \prod_{k=1}^{n} P(w_k \mid c_j) = \prod_{i=1}^{M} P(w_i \mid c_j)^{N(w_i)}$$

- 多项式分布朴素贝叶斯决策规则 (模型)

$$P(c_j \mid \mathbf{x}) = \frac{P(\mathbf{x}, c_j)}{P(\mathbf{x})} \propto P(\mathbf{x}, c_j) = P(c_j) \prod_{i=1}^{M} P(w_i \mid c_j)^{N(w_i)}$$

$$c^* = \underset{j=1,\ldots,C}{\operatorname{argmax}} P(c_j) \prod_{i=1}^{M} P(w_i \mid c_j)^{N(w_i)}$$

为什么是生成式

多项式分布NB模型中的参数估计

- 最大似然估计

$$P(c_j) \approx \frac{1 + N(c_j)}{C + N_{\text{all}}} \qquad P(w_i \mid c_j) \approx \frac{1 + N(w_i, c_j)}{M + \sum_{i'=1}^{M} N(w_{i'}, c_j)}$$

- 多项式分布NB模型一个例子

Feature Set = [计算机, 排球, 运动会, 高校, 大学]

$P(c_j)$	P(教育)=0.5	P(体育)=0.5
$P(w_i \mid c_j)$	P(计算机|教育)=0.3	P(计算机|体育)=0.1
	P(排球|教育)=0.1	P(排球|体育)=0.3
	P(运动会|教育)=0.1	P(运动会|体育)=0.3
	P(高校|教育)=0.2	P(高校|体育)=0.1
	P(大学|教育)=0.3	P(大学|体育)=0.2

多项式分布NB模型决策的例子

- “北京 理工 大学 是 理工 为主 工理文 协调 发展 的 全国 重点 高校”

Feature Set = [计算机, 排球, 运动会, 高校, 大学]

$$\mathbf{x} = [0,0,0,1,1]^T$$

$$P(\text{教育})P(\mathbf{x} \mid \text{教育}) = 0.5 \times 0.3 \times 0.2 = 0.03$$

$$P(\text{体育})P(\mathbf{x} \mid \text{体育}) = 0.5 \times 0.1 \times 0.2 = 0.01$$

$$P(\text{教育} \mid \mathbf{x}) = \frac{0.03}{0.03 + 0.01} = 0.75$$

$$P(\text{体育} \mid \mathbf{x}) = 0.25$$

多项式分布NB模型决策的例子

- “复旦 大学 排球 队 获得 本届 大学生 运动会 排球 比赛 冠军”

Feature Set = [计算机, 排球, 运动会, 高校, 大学]

$$\mathbf{x} = [0,2,1,0,1]^T$$

$$P(\text{教育})P(\mathbf{x} \mid \text{教育}) = 0.5 \times 0.1^2 \times 0.1 \times 0.3 = 0.00015$$

$$P(\text{体育})P(\mathbf{x} \mid \text{体育}) = 0.5 \times 0.3^2 \times 0.3 \times 0.2 = 0.0027$$

$$P(\text{教育} \mid \mathbf{x}) = \frac{0.00015}{0.00015 + 0.0027} = 0.0526$$

$$P(\text{体育} \mid \mathbf{x}) = 0.9474$$

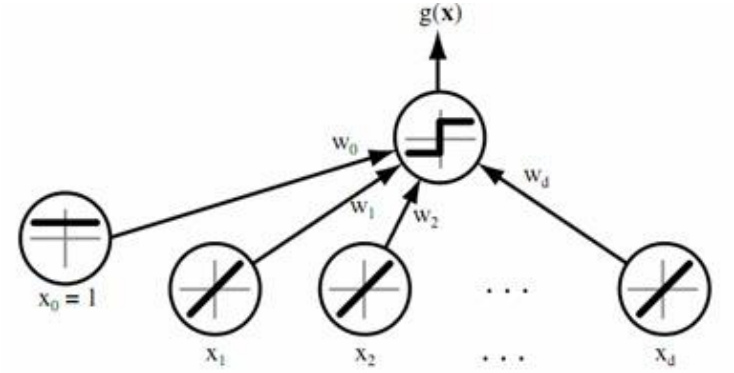
分类算法

- 监督学习
  - ✓ 生成式模型
    - ➢ 朴素贝叶斯（Naïve Bayes）
  - ✓ 判别式模型
    - ➢ 线性判别函数（Linear Discriminate Function）
    - ➢ 支持向量机（Support Vector Machine）
    - ➢ 最大熵模型（Maximum Entropy）
- 无监督
- 半监督学习

线性判别函数

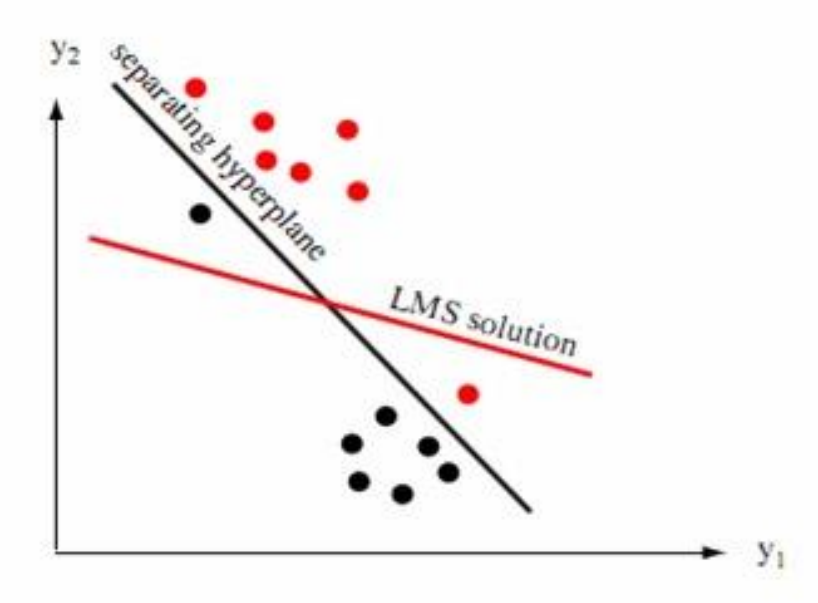
- 模型表示

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = \sum_{l=1}^{M} w_l x_l + w_0$$



线性判别函数对应一个线性决策面

线性判别函数的学习准则



- 感知器准则
- 最小均方差 (LMS)
- 交叉熵 (CE)
- 最小分类错误率 (MCE)
- …

哪个分类面更优?

选择哪个学习准则?

本章内容

- 3.1 引言
- 3.2 基于传统机器学习的文本分类方法
  - 3.2.1 文本表示
  - 3.2.2 特征选择
  - 3.2.3 分类算法
- 3.3 文本分类性能评估
  - 3.3.1 准确率、宏平均/微平均的召回率、精确率和F_1值
  - 3.3.2 P-R曲线
  - 3.3.3 ROC曲线、AUC

文本分类性能评估

- 假设一个文本分类任务共有M个类别，类别名称分别为$C_1, \ldots, C_M$。在完成分类任务以后，对于每一类都可以统计出真正例、真负例、假正例和假负例四种情形的样本数目。
  - **真正例（True Positive, TP）：**模型正确预测为正例（即模型预测属于该类，真实标签属于该类）。
  - **真负例（True Negative, TN）:** 模型正确预测为负例（即模型预测不属该类，真实标签不属该类）。
  - **假正例（False Positive, FP）：**模型错误预测为正例（即模型预测属于该类，真实标签不属该类）。
  - **假负例（False Negative, FN）：**模型错误预测为负例（即模型预测不属该类，真实标签属于该类）。

文本分类性能评估

类别	TP	FP	FN	TN
C_1	TP_1	FP_1	FN_1	TN_1
C_2	TP_2	FP_2	FN_2	TN_2
…				
C_M	TP_M	FP_M	FN_M	TN_M

所有类别的微观统计值

文本分类性能评估

- 召回率、精确率和F_1值

$$P_i = \frac{TP_i}{TP_i + FP_i}$$

$$R_i = \frac{TP_i}{TP_i + FN_i}$$

$$F_1 = \frac{2PR}{P + R}$$

本章内容

- 3.1 引言
- 3.2 基于传统机器学习的文本分类方法
  - 3.2.1 文本表示
  - 3.2.2 特征选择
  - 3.2.3 分类算法
- 3.3 文本分类性能评估
  - 3.3.1 准确率、宏平均/微平均的召回率、精确率和F_1值
  - 3.3.2 P-R曲线
  - 3.3.3 ROC曲线、AUC

文本分类性能评估

- 正确率

$$Acc=\frac{\#Correct}{N}$$

N为样本总数，$\#Correct$为其中被模型正确预测的样本数

- 宏平均的召回率、精确率和F_1值定义分别为：

$$\text{Macro\_P} = \frac{1}{C}\sum_{j=1}^{C}\frac{TP_i}{TP_i + FP_i} \qquad \text{Macro\_}R = \frac{1}{C}\sum_{j=1}^{C}\frac{TP_i}{TP_i + FN_i}$$

$$\text{Macro\_}F_1 = \frac{2 \times \text{ Macro\_P} \times P \text{ Macro\_R}}{\text{Macro\_P} + \text{ Macro\_R}}$$

文本分类性能评估

- 微平均的召回率、精确率和F_1值定义分别为：

$$\text{Micro\_P} = \frac{\sum_{j=1}^{C} TP_i}{\sum_{j=1}^{C} (TP_i + FP_i)} \qquad \text{Micro\_R} = \frac{\sum_{j=1}^{C} TP_i}{\sum_{j=1}^{C} (TP_i + FN_i)}$$

$$\text{Micro\_F } F_1 = \frac{2 \times \text{ Micro\_P } \times \text{ Micro\_ } R}{\text{Micro\_P } + \text{ Micro\_R}}$$

在二分类且类别互斥的情况下，$Micro\_P$、$Micro\_R$、$Micro_F_1$都与正确率Acc相等。

文本分类性能评估

二分类的分类结果示例

预测/真实	正类(+)	负类(−)	全部
正类(+)	250	20	270
负类(−)	50	180	230
全部	300	200	500

	TP	FP	FN	TN	Recall	Precision	F_1	Acc
正类(+)	250	20	50	180	0.8333	0.9259	0.8772	0.8600
负类(−)	180	50	20	250	0.9000	0.7826	0.8372	
宏平均					0.8667	0.8543	0.8605	
微平均					0.8600	0.8600	0.8600	

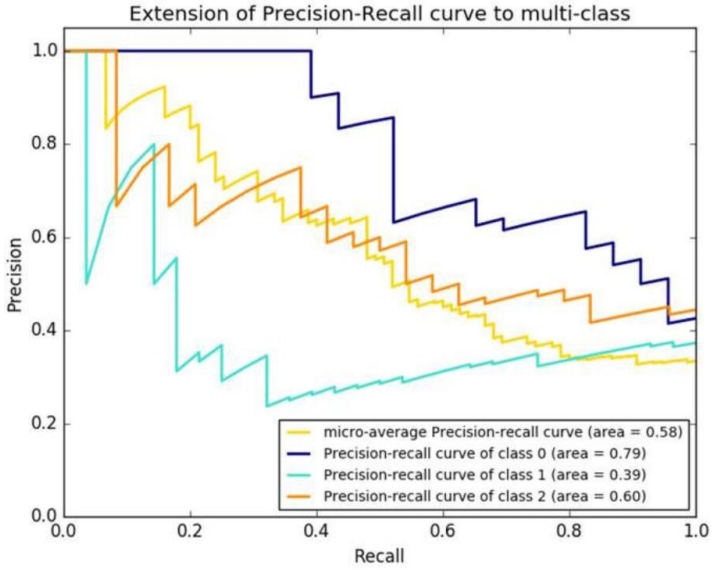
针对上表分类结果的评估指标

本章内容

- 3.1 引言
- 3.2 基于传统机器学习的文本分类方法
  - 3.2.1 文本表示
  - 3.2.2 特征选择
  - 3.2.3 分类算法
- 3.3 文本分类性能评估
  - 3.3.1 准确率、宏平均/微平均的召回率、精确率和F_1值
  - 3.3.2 P-R曲线
  - 3.3.3 ROC曲线、AUC

文本分类性能评估

- P-R曲线
  - 通过调整分类器的阈值，将按输出排序的样本序列分割为两部分，大于阈值的预测为正类，小于阈值的预测为负类，从而得到不同的召回率和精确率。如设置阈值为0时，召回率为1；设置阈值为1时，则召回率为0。以召回率作为横轴、精确率作为纵轴，可以绘制出精确率-召回率（precision-recall, PR）曲线

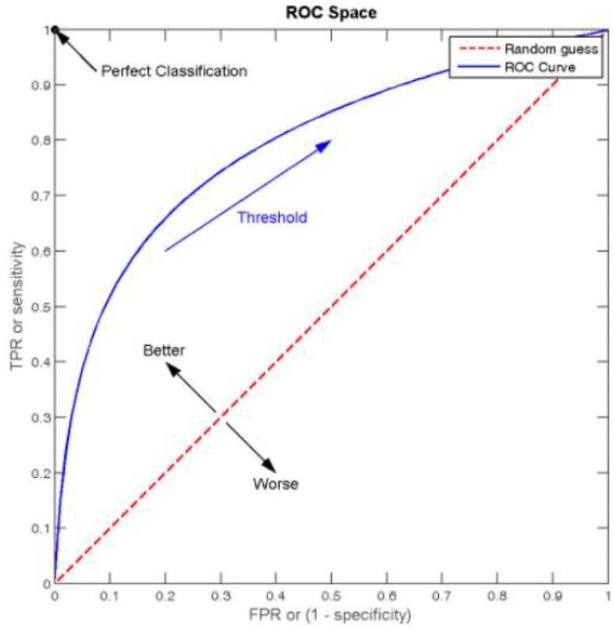


本章内容

- 3.1 引言
- 3.2 基于传统机器学习的文本分类方法
  - 3.2.1 文本表示
  - 3.2.2 特征选择
  - 3.2.3 分类算法
- 3.3 文本分类性能评估
  - 3.3.1 准确率、宏平均/微平均的召回率、精确率和F_1值
  - 3.3.2 P-R曲线
  - 3.3.3 ROC曲线、AUC

文本分类性能评估

- ROC曲线、AUC
  - 以假正率（false positive rate）作为横坐标，以真正率（true positive rate）（即召回率）作为纵坐标，绘制出的曲线称为ROC（receiver operating characteristic）曲线。ROC曲线下的面积称为AUC（area under ROC curve），AUC曲线越靠近左上方越好。AUC值越大，说明分类器性能越好。



本章小节

- 文本分类
  - 基于传统机器学习的方法
    - 文本表示
    - 特征选择
    - 分类算法
  - 文本分类性能评估
    - 准确率
    - 宏平均的召回率、精确率和F_1值
    - 微平均的召回率、精确率和F_1值
    - **P-R**曲线
    - **ROC**曲线、**AUC**

欢迎提问！