COMP9414：人工智能第6b讲：文本分类

韦恩-沃布克
电邮：w.wobcke@unsw.edu.au

本讲座

- 文本分类的概率公式化
- 基于规则的文本分类
- 贝叶斯文本分类
  - 伯努利模型

文本分类应用

- 垃圾邮件检测
- 著述权分析
- 电子邮件分类/优先排序
- 新闻/科学文章主题分类
- 事件提取（事件类型分类）
- 情绪分析
- 推荐系统（使用产品评论）

电影评论/评级实例

  - 多项式奈何贝叶斯
- 评估分类器

.令人难以置信的失望.....。

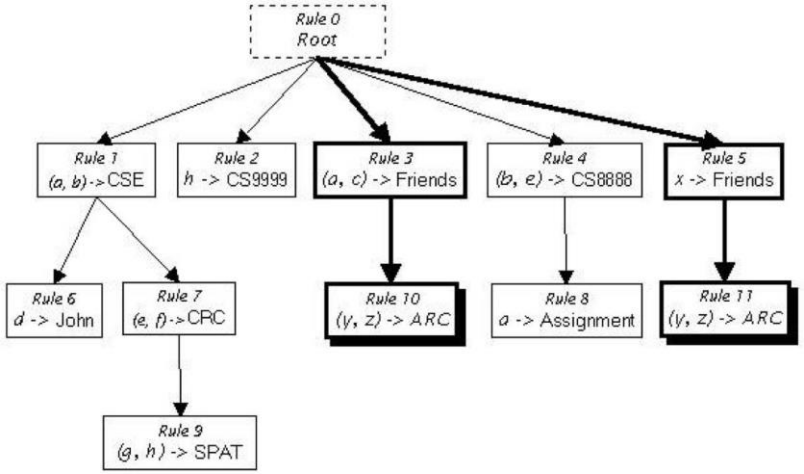
充满了古怪的人物和丰富的应用讽刺
，还有一些伟大的情节转折。

有史以来最伟大的螺旋式喜剧拍摄。

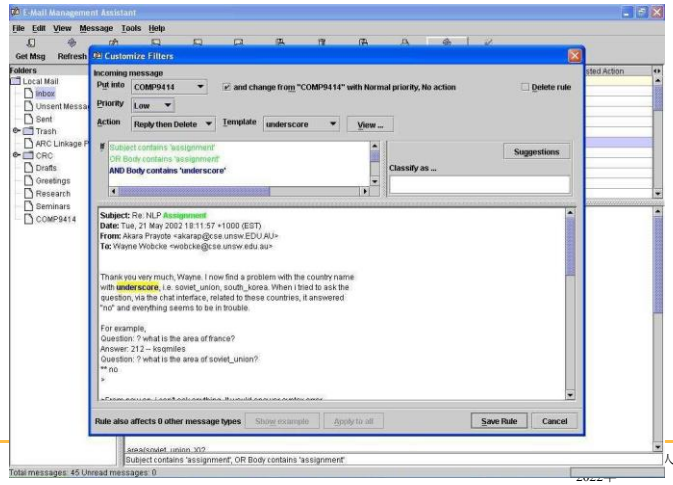
这很可悲。最糟糕的部分是拳击场面。

新南威
尔士大
学
©W.Wobcke等人，2019-
2022年

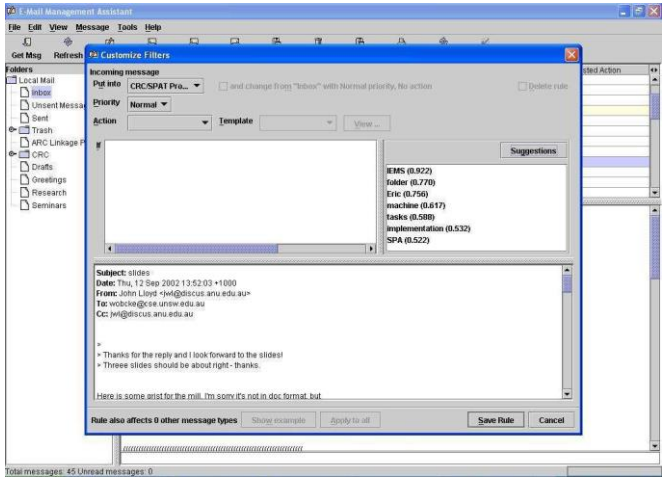
基于规则的方法



帮助用户定义规则



使用Naive Bayes推荐特征



监督学习

- 输入。一份文件（电子邮件、新闻报道、评论、**推特**）。
- 输出。从一个固定的班级集合中抽取一个班级
  - 所以文本分类是一个多类分类问题
  - .....有时是一个多标签的分类问题
- 学习问题
  - 输入。训练集的标记文件{$(d_1, c_1), - - -$
  - 输出。学习到的分类器，将d映射到预测的c类上

概率论的表述

- 事件。特征x的出现，类别c的文件的出现
- 给定文件$x_1, ---, x_n$， 选择c， 使$P(c|x_1, --, x_n$)达到最大。
- 应用贝叶斯规则
  - $P(c|x1,--, xn) = \frac{P(x1,--,xn|c)。P(c)}{P(x1--,xn)}$
  - 因此，最大化$P(x_1, ---, x_n |c)。P(c)$

特色工程

例子。SpamAssassin (垃圾电子邮件)

- 提到通用伟哥
- 网上药店
- 提及百万（美元）（（美元）NNN,NNN,NNN.NN）。
- 短语：打动人心 .. 女孩
- 来自：以许多数字开头
- 主题都是大写的
- HTML的文字与图像面积的比例很低
- 百分之百保证
- 声称可以将你从名单中删除http://spamassa

伯努利模型

最大化$P(x_1, ---, x_n |c)。P(c)$

- 特征是文件中是否存在w_i。
- 应用独立假设
  - $P(x_1, ---, x_n |c) = P(x_1 |c)。---.P(x_n |c)$
  - 词w（不）在c类中的概率与上下文无关
- 估计概率
  - $P(w|c)$ = #(w在c类中的文件)/#(c类中的文件)
  - $P(\neg w|c) = 1 - P(w|c)$
  - $P(c)$ = #(c类中的文件)/#文件

ssin.apache.org/old/tests 3 3 x.html

Naive Bayes 分类

	等级=1	*等级*=0
$P(Class)$	0.40	0.60
$P(w_1 \mid 类)$	0.75	0.50
$P(w_2 \mid 类)$	0.25	0.67
$P(w_3 \mid 类)$	0.50	0.33
$P(w_4 \mid 类)$	0.50	0.50

对有*w2*、*w3*、*w4*的文件进行分类

- $P(Class = 1 \mid \neg w1, w2, w3, w4)$
  $\approx ((1 - 0.75) * 0.25 * 0.5 * 0.5) * 0.4$
  $= 0.00625$
- $P(Class = 0 \mid \neg w1, w2, w3, w4)$
  $\approx ((1 - 0.5) * 0.67 * 0.33 * 0.5) * 0.6$
  $= 0.03333$

w1	*w2*	*w3*	*w4*	*级别*
1	0	0	1	1
0	0	0	1	0
1	1	0	1	0
1	0	1	1	1
0	1	1	0	0
1	0	0	0	0
1	0	1	0	1
0	1	0	0	1
0	1	0	1	0
1	1	1	0	0

新南威尔士大学 ©W.Wobcke等人，2019-2022年

文字袋模型

我喜欢这部电影!它很温馨，但有讽刺性的幽默。对话很好，冒险场面也很有趣。浪漫，同时对童话体裁的惯例感到好笑。我几乎会向任何人推荐它。我已经看过好几遍了，每当我有朋友还没有看过的时候，我总是很高兴再看一遍！"。

它	6
I	5
的	4
至	3
和	3
看到的	2
但	1
会	1
奇思妙想	1
时间	1
甜的	1
讽刺的	1
冒险	1
体裁	1
仙子	1
幽默	1
有	1
伟大的	1

奈何贝叶斯分类

最大化$P(x_1, - - - , x_n | c)$。$P(c)$

- 特征是单词在文件中的位置出现的情况
- 应用独立假设
  - $P(w_1, - - , w_n | c) = P(w_1 | c)$。$- - - .P(w_n | c)$
  - 词w在文件中的位置并不重要
- 估计概率
  - 设V为词汇表
  - 让 "文件 "c=类c中文件的串联
  - P（$w|c$）=#（w在文档c中）/$\Sigma_{w \in V}$#（w在文档c中）

拉普拉斯平滑

- 如果测试文件中的单词在训练中没有出现，怎么办？
- 那么$P(w|c)=0$，所以对c类的估计是0
- **拉普拉斯平滑**
  - 给未见过的单词分配小概率
  - P（$w|c$）=（#（w在文档c中）+1）/（$\Sigma_{w \in V}$#（w在文档c中）+|V|）
  - 不一定要加1，可以是0.05或一些参数α

MNB实例

- $P(c)$ = #(c类中的文件)/#文件

	词条	级别
d_1	中国人 北京人	c
d_2	中国 中国上海	c
d_3	中国澳门	c
d_4	东京 日本 中文	j
d_5	中国人 中国人 东京 日本	?

P（中文|c）=（5+1）/（8+6）=3/7 P（东京|c）=（0+1）/（8+6）=1/14 P（日本|c）=（0+1）/（8+6）=1/14

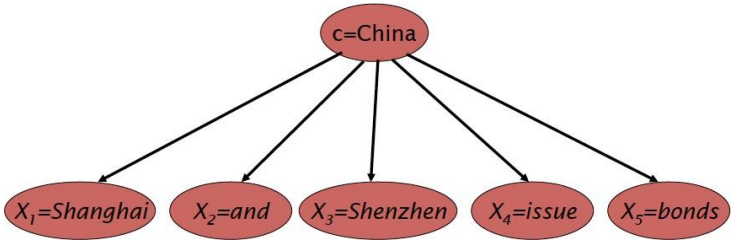
$P(中文|j) = (1+1)/(3+6) = 2/9$ $P(东京|j) = (1+1)/(3+6) = 2/9$ $P(日本|j) = (1+1)/(3+6) = 2/9$

为了对文件*d*进行分类[5]

- $P(c|d5) \propto [(3/7)^3 \cdot \mathbf{1/14} \cdot \mathbf{1/14}] \cdot 3/4$
  $\approx \mathbf{0}.0003$
- $P(j|d5) \propto [(2/9)^3 \cdot 2/9 \cdot 2/9] \cdot 1/4$
  $\approx \mathbf{0}.0001$
- 选择*c*类

新南威尔士大学
©W.Wobcke等人，2019-2022年

示例的图形模型



评估分类器

2×2应急表（单班*C*级）

	*c*类	不是*C*类
预测的*c*	真正的积极性	假阳性
预测的不是*c*	假阴性	真阴性

- 精度(P) = TP/(TP+FP) - 你想得到什么就有什么
  - ---但可能不会得到很多
- 召回率(R)=TP/(TP+FN) - 你得到你想要的东西

多个班级。每类指标

n×*n*混淆矩阵（每个实例在一个类别中）。

	预测的c_1	预测的c_2	---
*c*类$_1$	*c11*	*c12*	*c13*
*c*类$_2$	*c21*	*c22*	*c23*
---	*c31*	*c32*	*c33*

- 精度(*c*类$_i$) = c_{ii} /$\Sigma_j c_{ji}$
  - 预测为c_i的项目比例正确分类（如c_i）。
- 回顾（类c_i）= c_{ii} /Σ_j c_{ij}
  - *c*类项目的比例$_i$预测正确（如c_i）
- 准确度=Σ_i c_{ii} /Σ_i Σ_j c_{ij}

多类。微观/宏观平均法

n（每班一个）2×2应急表

- 微观平均数=对所有类别的汇总措施
  - 微型精度 = Σ_c TP_c /Σ_c (TP_c + FP_c)
  - 微召回 = Σ_c TP /Σ_c (TP + FN)
  - ---但你可能得到更多的（垃圾）。
- F1 = 2PR/(P+R) - 精度和召回率的谐波平均值

- 当每个实例具有并被赋予一个且仅有一个标签时也是如此
  - 被较大的阶级所支配

宏观平均数=每类措施的平均值

- 宏观精度 = $\Sigma^{1}_{c}\ \mathrm{TP}_c / (\mathrm{TP}_c + \mathrm{FP}_c)$
- macro-recall = $\frac{1}{n}\Sigma^{1}_{c}\ \mathrm{TP}_c / (\mathrm{TP}_c + \mathrm{FN}_c)$
- 以小班为主导
- 对不平衡的数据更公平，例如情感分析。

摘要：奈何贝叶斯

- 非常快，存储要求低
- 对不相关的特征具有鲁棒性
- 不相关的特征相互抵消而不影响结果
- 在各领域非常好，有许多同样重要的特点
  - 决策树在这种情况下会受到碎片化的影响--
    尤其是在数据很少的时候。
- 如果独立假设成立，则是最优的
  如果假设的独立性是正确的，那么它就是问题的贝叶斯最优分类器。
- 良好可靠的文本分类基线