



文本处理实践
第四讲

文本自动分类

文本信息处理的基本任务

关键任务

自动分词
命名实体识别
词性标注
句法分析
语义分析
篇章分析

应用型任务

机器翻译
文本分类
情感分析
信息检索与过滤
自动问答
信息抽取
自动文摘
人机对话

中文自动分词

各种工具各有优缺点：<http://blog.csdn.net/hello9050/article/details/7889658>

“Jieba”：Python中文分词组件

<http://ictclas.nlpir.org/nlpir/>

词性标注 (POS Tagging)



文本分类

- 简介
- 评价指标

- 分类的概念

分类：对于给定一个对象，从一个事先定好的分类体系中挑出一个（或者多个）最适合该对象的类别。

分类体系：分类体系一般人工构造，常见为层次结构

分类模式：

2类问题，属于或不属于(binary)

多类问题，多个类别(multi-class)，可拆分成2类问题

一个对象可以属于多类(multi-label)

- 分类的概念

分类：对于给定一个对象，从一个事先定好的分类体系中挑出一个（或者多个）最适合该对象的类别。

分类体系：分类体系一般人工构造，常见为层次结构

分类模式：

2类问题，属于或不属于(binary)

多类问题，多个类别(multi-class)，可拆分成2类问题

一个对象可以属于多类(multi-label)

- Text Categorization (TC)
- 在给定的分类体系下，根据文本的内容自动地确定文本关联的类别。数学角度来看，文本分类是一个映射的过程，它将未标明类别的文本映射到已有的类别中，该映射可以是一一映射或一对多的映射。
- 应用
 - 门户网站（网页）
 - 图书馆（电子资料）
 - 情报/信息部门（情报处理）
 - 政府、企业等（电子邮件）

• Text Categorization (TC)

- 原标题：个税改革进入倒计时
- 资料图。

免征额有上调空间，综合与分类相结合，专项扣除考虑家庭因素，税率可适当下调，这正是备受关注的个税改革可着力的四个方向法治周末记者 赵晨熙

关乎每个人“钱袋子”的个人所得税改革，在每年两会上都会成为众人关注的热点问题。

“目前，个人所得税的改革方案正在研究设计和论证中，总体思路是实行综合与分类相结合，方案总体设计、实施分步到位，逐步建立起适合我国国情的个人所得税制。”

.....



- 基本步骤
- 定义分类体系
- 将预先分类过的文档作为训练集
- 从训练集中得出分类模型（需要测试过程，不断细化）
- 用训练得出的分类模型对其它文档加以分类

文本分类 简介

- 评价指标——错误率与精度（即适用于二分类，也适用于多分类问题）
- 错误率：分类错误的样本数占样本总数的比例
- 精度：分类正确的样本数占样本总数的比例
- 给定样例集 $D = \{(x_1, y_1), \dots, (x_m, y_m)\}$
- 期中 y_i 是 x_i 的真实标记。要评估学习器 f 的性能，就要把学习器预测结果 $f(x)$ 与真实标记 y 进行比较。
- 对于样例集 D ，分类错误率：
$$E(f; D) = \frac{1}{m} \sum_{i=1}^m f((x_i) \neq y_i)$$
- 精度：
$$\text{acc}(f; D) = \frac{1}{m} \sum_{i=1}^m f((x_i) = y_i) = 1 - E(f; D)$$

文本分类 简介

- 评价指标——查准率，查全率与F1

「准确率」 (P, precision) , 「召回率」 (R, recall) , F - Measure()

对于二分类问题，可将样例根据其真实类别与学习器预测类别的组合划分为真正例 (true positive) , 假正例 (false positive) , 真反例 (true negative) , 假反例 (false negative) 。

分类结果混淆矩阵

$$P = \frac{TP}{TP + FP} \quad R = \frac{TP}{TP + FN}$$

真实情况	预测结果	
	正例	反例
正例	TP	FN
反例	FP	TN

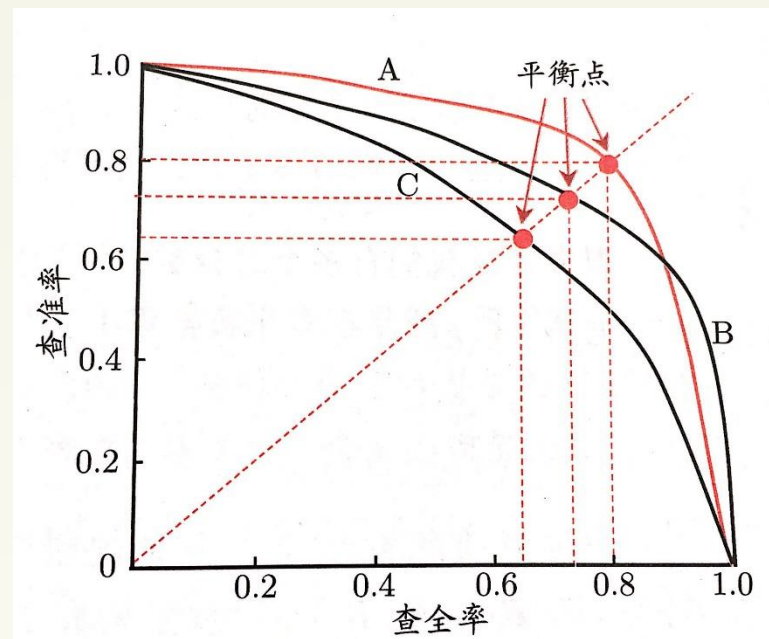
文本分类 简介

- 评价指标——查准率，查全率与F1
- P-R 曲线：P 和R是一对矛盾的度量。一般来说，查准率高时，查全率往往偏低。
- 在很多情况下，可以根据学习器的预测结果对样例进行排序，排在前面的是学习器认为“最可能”是正例的样本。排在最后的则是学习器认为“最不可能”的正例的样本。按此顺序逐个把样本作为正例进行预测，则每次可以计算出当前的P和R，以R为横轴，P为纵轴作图，可以得到P-R曲线。

A和C？A性能优于C。

A和B？曲线下面积大小。

平衡点（Break-Even Point，简称BEP）， $P=R$ 时的取值



- 评价指标——查准率，查全率与F1

$$F1 = \frac{2 * P * R}{P + R}$$

在一些应用中，P和R的重视程度有所不同。因此，用F1度量的一般形式来表达偏好。

$$F_{\beta} = \frac{(1 + \beta^2) * P * R}{\beta^2 * P + R}$$

取值为1时，退化为标准的F1

>1,更关注R

<1,更关注P

- 评价指标——查准率，查全率与F1
- 多次测试，多个数据集，或者是多分类任务——我们希望在n个二分类混淆矩阵上综合考察P和R
- 宏查准率，宏查全率： $macro - P = \frac{1}{n} \sum_{i=1}^n p_i$, $macro - R = \frac{1}{n} \sum_{i=1}^n R_i$,

还可先将各混淆矩阵的对应元素进行平均，分别记为 $\overline{TP}, \overline{FP}, \overline{TN}, \overline{FN}$ 在基于这些平均值算出微查准率，微查全率。

$$micro - P = \frac{\overline{TP}}{\overline{TP} + \overline{FP}}, \quad micro - R = \dots \dots$$

完成文本分类作业。

对测试数据集得到的分类结果进行评价指标的计算。要求计算P, R, F1