第4章：文本分类

课件制作：曲彦儒、屠可伟
讲解人：曲彦儒

文本分类示例

影评情感分类

哪些是好评，哪些是差评？

- 真·一出好戏。剧情表演俱佳...
- 中国不适合拍这种片。
- 意外的好看，也可能是因为预告片感觉不伦不类导致期望值不高...
- 提前看的，无聊...
- 荒诞中揭示人性，导演处女作应该说太用心了…
- 电影前期口碑营销做得太狠了，正所谓希望越大失望也越大。

文本分类示例

文本分类应用

- 主题分类
- 垃圾邮件过滤
- …

定义

文本分类

- 输入：
  - 文档 d
  - 给定的类别集合 $C = \{c^1, c^2, ..., c^K\}$
- 输出：
  - 对目标文档 d 的一个预测类别 $c \in C$

文本分类方法

基于机器学习的文本分类

- 生成式模型（建模 $p(x,y)$）
  - 朴素贝叶斯（Naïve Bayes）
- 判别式模型（建模 $p(y|x)$）
  - 逻辑斯谛回归（logistic regression）
  - 支持向量机（support vector machines）
  - 神经网络 （neural networks）
  - 决策树（decision tree）

4.2 基于机器学习的文本分类

逻辑斯谛回归

- 将文本表示为特征向量
  - TF-IDF，*n*元语法，词嵌入，etc.
- 计算概率
  - 对特征向量线性求和，使用 softmax 函数归一化

$$P(y=c|x)=\frac{\exp(w_c^{\mathrm{T}}x+b_c)}{\sum_{j=1}^{K}\exp(w_j^{\mathrm{T}}x+b_j)}$$

- 使用最大似然估计和随机梯度下降训练

4.3 分类结果评价

测试集 Test Set

一组独立于训练集，用作分类性能评价的数据。

使用混淆矩阵计算评价指标

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

- 准确度（Accuracy）计算分类正确的样本占全部样本的比例

分类器结果 \ 真实标签	正	负
正	真阳性TP (true positive)	假阳性FP (false positive)
负	假阴性FN (false negative)	真阴性TN (true negative)

准确度的问题

在测试数据上通过准确度衡量分类器的性能存在哪些问题呢？

- 类别不均衡
  - 例如：如果99%的邮件都不是垃圾邮件，那么永远预测“否”可以得到99%的准确率
- 无法考虑类别间的相对重要性与不同误差的代价
- 测试数据本身的随机性

精度与召回

二分类中的精度/召回

- 二分类（binary classification）中给定类别集合中有且仅有两个类别，即 $|C| = 2$，假定其中一个类别 $t \in C$ 为目标类别。
- 精度（Precision）与召回（Recall）分别从返回的目标类别集合是否精确，以及能否尽可能多的找出属于目标类别的样本，这两方面评价分类器的性能。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{F}_1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

分类器结果	真实标签：正	真实标签：负
正	真阳性TP (true positive)	假阳性FP (false positive)
负	假阴性FN (false negative)	真阴性TN (true negative)

精度与召回

多分类中的精度/召回

在多分类问题中（$|C| > 2$），如何综合考虑多个目标类别，整体衡量分类器的分类性能呢？

宏平均 Macroaveraging

- 计算分类器在每一类上的精度/召回，然后取平均

微平均 Microaveraging

- 把分类器在各类别上的统计求和，然后计算精度/召回

多分类精度/召回示例

分类器预测结果

	truth yes	truth no
classifier yes	10	20
classifier no	20	950

类别 c^1 上的预测结果

	truth yes	truth no
classifier yes	30	40
classifier no	30	900

类别 c^2 上的预测结果

宏平均 Macroaveraging

- 宏平均精度 $\mathrm{P} = \frac{0.33+0.43}{2} = 0.38$

微平均 Macroaveraging

- 微平均精度 $\mathrm{P} = 40/100 = 0.4$

	truth yes	truth no
classifier yes	40	60
classifier no	50	1850

类别 c^1 和 c^2 上的整体预测结果

总结

文本分类总结

- 定义：预测给定文档的所属类别
- 模型：生成式（generative），判别式（discriminative）
- 评价指标：准确度（Accuracy），精度（Precision），召回（Recall）