Homework 4:
电影评论情感分类

大数据原理与技术 (Spring 2025)

22336226 王泓沣
Lectured by: Changdong Wang
Sun Yat-sen University

1 问题描述

使用 IMDB 影评数据集，使用 TF-IDF 进行特征提取并用逻辑回归/SVM/RNN 实现分类，分析准确率差异。

2 数据描述

IMDB 数据集包含 50000 个带标签样本，划分为训练集和测试集，比例为 1:1。每一个样本为一句英文影评，标签为积极 (pos) 或消极 (neg)。

3 Method

3.1 TF-IDF 特征提取

TF(Term Frequency): 一个词在当前样本中出现的频率，若词 t 在样本 d 中出现的次数为 $f_{t,d}$，则

$$TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

IDF(Inverse Document Frequency): 衡量一个词在整个数据集中是否普遍出现，若数据集中包括 N 个样本，其中词 t 出现的样本数为 n_t，则

$$IDF(t) = \log(\frac{N}{n_t})$$

最终得到 TF-IDF 评分

$$TF - IDF(t,d) = TF(t,d) \times IDF(t)$$

3.2 逻辑回归

在文本特征空间中寻找线性决策边界

$$P(y=1|x) = \sigma(w^\top x + b) = \frac{1}{1+e^{-(w^\top x+b)}}$$

3.3 SVM

在文本特征空间中寻找最优超平面

$$\min \frac{1}{2}\|w\|^2$$
$$s.t. y_i(w_i x_i + b) \geqslant 1$$

3.4 RNN

RNN 接受一个序列 $[x_1, x_2, ..., x_n]$ 作为输入，在时间步 t，更新隐藏状态 h_t

$$h_t = f(W_x x_t + W_h h_{t-1} + b)$$

其中 W_x 和 W_h 分别是输入和隐藏状态的权重矩阵

4 Result

Method	Accuracy
LogisticRegression	0.8807
SVM	0.8727
RNN	0.7778

Table 1: IBDB 二分类结果

IMDB 数据集文本数据特征清晰，是一个中小型数据集，因而 RNN 效果存在一定程度过拟合，效果不佳。