文本情感分析

作者 林映廷

摘要 随着我国信息化程度的不断提高，各种网络平台上的信息爆炸式增长，信息系统的应用能够有效分析舆论趋势，能够有效的保障网络内容安全，如果能从这些大量的信息中挖掘出具有价值的规律或关系，就能够为推荐系统提供可靠依据.基于此，本文使用了线性回归器来分辨不同评论的态度，并通过 SGD 来减小损失函数.同时，为了对没有标签的数据进行分析，代码也对 k-means 聚类进行了实现.

关键词 k-means 聚类，多元线性回归，随机梯度下降

Text Sentiment Analysis

Lin Ying-ting

Abstract With the continuous improvement of the degree of informatization in our country, the explosive growth of information on various network platforms, the application of information systems can effectively analyze the trend of public opinion and effectively ensure the security of network content. The law or relationship of value can provide a reliable basis for the recommendation system. Based on this, this paper uses a linear regression to distinguish the attitudes of different reviews, and uses SGD to reduce the loss function. At the same time, in order to analyze the data without labels , the code also implements k-means clustering.

Key words k-means clustering, MLR, SGD

1 引言

实现斯坦福 CS221 的作业中 Sentiment Analysis 的代码部分.

1.1 背景

情感是一种心理状态，通常会导致人们的行为方式和计算理性相冲突，是人类等高等生物区分于计算机的显著属性，人类的大脑具有意识的维度，著名的情感学家Scherer将情感定义为组成过程，Scherer指出，情绪在适应生物体生命中频繁发生和典型模式的重大事件方面发挥着重要作用，情绪范围很难界定，而愤怒、喜悦、恐惧、悲伤等功利主义情绪相对频繁出现.情感在人工智能领域有着重要的研究价值.情感分析，又称情感计算、意见挖掘，最早起源于Picard提出的“情感计算”概念，Picard指出，情绪在人类的思维、推断和决策中发挥着重要作用，情感计算机可以通过识别人类情感来提高决策能力.

在该论文中，我们将构建一个线性分类器，该分类器获取电影评论信息后，猜测它们是“正面”还是“负面”.该分类器的优化方法主要是随机梯度下降，减少该分类器的训练时间.

但是由于数据采集后对大量数据进行标注需要消耗大量人力。所以为了处理无标签的数据，实现了多维向量的K-means聚类.

为了更好地理解SGD和K-means聚类，代码部分没有使用numpy，sklearn等常用机器学习相关的包.

2 方法

2.1 数据处理及特征提取

本文的数据处理及特征提取的方式主要包含两种。

一种是处理用空格隔开单词的语言（如：英文）时，直接得到不同单词出现的次数，并基于此得到对应的特征向量，在函数 *extractWordFeatures* 中实现，效果如图 1 所示：

```
@param string x:
@return dict: feature vector representation of x.
Example: "I am what I am" --> {'I': 2, 'am': 2, 'what': 1}
```

图 1

一种是处理没用空格隔开单词的语言时，无法直接得到不同单词出现的次数，而在该实验使用了一种较为简单方法，就是直接以几个字母在一起之后的集合，做为一个单词。然后，用不同单词的数量做为该评论的特征，使用函数 *extractCharacterFeatUres* 实现，其中具体是使用几个字母做为一个单词，需要通过不断地实验，通过最后的实验效果之后确定.如图 2 所示：

图 2

2.2 模型建立

2.2.1 K-means 聚类

K-means 算法是以距离为数据对象间相似性度量的标准，即数据对象间的距离越小，相似性越高，越有可能在同一个类内.对于任意给定的含有 n 个数据对象的非空数据集，K-means 算法需要人为先制定初始聚类中心的数目即 k 的值，根据 k 值个数确定初始聚类中心的个数，即从待运算的数据集中随机选取相应个数的点，然后利用距离公式计算数据对象与初始聚类中心之间的最短距离，按照就近原则，将数据对象分配到距离最短的聚类中心所在的类中，最后对聚类中心点的位置不断修正.因为聚类算法的函数是收敛函数，算法每次迭代后都会参照给定的聚类目标函数或聚类效果评价准则，对目标函数值的数值进行递减运算.
对于任意给定样本$D=\{x_1,x_2,x_3\ldots,x_m\}$，经过 K-means 算法得到的聚类的类为$C=\{C_1,C_2,C_3\ldots,C_k\}$，采用欧式距离作为衡量指标时，得到的平方误差为

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}||x-\gamma_i||^2$$

$$\gamma_i=\frac{1}{|C_i|}\sum_{x\in C_i}x$$

其中γ_i是 C_i的均值向量，有时也称为质心。E 的值越小越说明类中数据对象相似度越高，我们的目标是取 E 的最小值，即误差平方和最小。对γ_i求导，当$C_i=k$时，γ_i为 E 函数在γ_i的最优解，且 E 值最小。对 x 而言，求导后，当$x=\gamma_i$时，E 最小。每次迭代 E 函数的值都会降低，最终趋向于收敛。

欧式距离公式为：

$$dist_{ed}(x_i,x_j)=(\sum_{u=1}^{n}|x_i-x_j|^2)^{\frac{1}{2}}$$

2.2.2 线性分类器

多元线性回归所考虑的零假设是所有变量都是不相关的，所以在训练时，可以应用 PCA（Peres-Neto 等，2005），提取特征减小不同变量间的相关性。在该实验中使用一组特征向量和表达情感之间的关系由一次线性方程组表示。一般方程如下：

$$\hat{Y}=P_0+P_1X_1+\cdots+P_nX_n$$

可以使用最小二乘的方法或者梯度下降法最小化损失函数，在该实验中使用的方法是利用随机梯度下降法，降低损失函数.其中多元线性回归的损失函数为：

$$loss=\frac{1}{2m}\sum_{i=1}^{m}(y^i-\sum_j w_j x_j^{(i)})^2$$

2.2.3 SGD

随机梯度下降 1847 年提出。每次选择一个 mini-batch，而不是全部样本，使用梯度下降来更新模型参数.参数更新方法如下：

$$\boldsymbol{W}\leftarrow\boldsymbol{W}-\mu\frac{\partial loss}{\partial \boldsymbol{W}}$$

它和使用全部的样本来更新模型参数相比能够增加模型的训练速度，但仍然有自适应学习率、容易卡在梯度较小点等问题。

3 结果与分析

在完成代码编写后，使用 *python grader.py* 来检验代码的正确性，在提供的检测案例中，basic 和 hidden 两个部分。一部分用来验证逻辑的正确性，一部分来验证代码的鲁棒性。最终，利用提供的验证程序得到代码部分的得分。

其中，由于在编写 SGD 部分代码时，是每一轮都遍历每一个例子，每单个例子输入都会改变权重，所以在其中一个简单案例中，第一轮就使训练集和测试集的误差都为 0，不符合第一轮迭代不能误差都为 0 的要求，

```
Note that the hidden test cases do not check for correctness.
They are provided for you to verify that the functions do not crash and run within the time limit.
Points for these parts not assigned by the grader (indicated by "--").
========== END GRADING [8/9 points + 0/0 extra credit]
```

所以被扣除 1 分.

References

1. 陈国伟,张鹏洲,王婷,叶前坤.多模态情感分析综述[J].中国传媒大学学报(自然科学版),2022,29(02):70-78.DOI:10.16196/j.cnki.issn.1673-4793.2022.02.009.胡新荣,陈志恒,刘军平,彭涛,何儒汉,何凯.基于 SGD 的决策级融合维度情感识别方法[J].郑州大学学报(理学版),2022,54(04):49-54.DOI:10.13705/j.issn.1671-6841.2021299.
2. 施文幸,曹诗韵.基于萤火虫 K-means 聚类的电力用户画像构建和应用[J].计算机系统应用.2021,30(08):281-287.DOI:10.15888/j.cnki.csa.008055.