人工智能实验

Week10　文本情感分类

实验任务

- 实验内容
  - **二选一实现**：k-NN 分类器、朴素贝叶斯分类器
  - 在给定文本数据集完成文本情感分类训练，在测试集完成测试，计算准确率。
- 实验要求
  - **不可**直接调用机器学习库中的分类器算法（仅可用于和自己的方法对比准确率）
  - **可**使用各种提取文本特征的辅助工具，如 OneHotEncoder、TfidfVectorizer 等
- 其他说明
  - 可以对比不同方法（不同编码、不同参数、自己实现和调库）或进行优化得到更高的准确率
  - 由于是多分类问题，准确率在 20% 以上就差不多了，最高也不会超过 40%
  - 时限：5.9**（周四）**晚 23:59，命名：E6_学号

文本信息情感分类

train.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

```
documentId emotionId emotion words
1 5 sad mortar assault leav at least dead
2 4 joy goal delight for sheva
3 4 joy nigeria hostag fear dead is freed
4 3 fear bomber kill shopper
5 6 surprise veget not fruit slow brain declin
6 4 joy pm havana deal a good experi
7 4 joy kate is marri doherti
8 6 surprise nasa revisit life on mar question
9 4 joy happi birthdai ipod
10 4 joy alonso would be happi to retir with three titl
11 4 joy madonna s new tot happi at home in london
12 5 sad nicol kidman ask dad to help stop husband s drink
13 4 joy unit find good connect in win
14 4 joy runwai make good without make nice
15 5 sad we were arrog and stupid over iraq sai us diplomat
16 5 sad bad reason to be good
17 6 surprise madonna s new babi s daddi didn t realiz adopt wa for good
```

第 1 行，第 35 列 100% Windows (CRLF) UTF-8

分类问题

- 文本的情感分类任务:
  - 输入：文本
  - 输出：类标签

Document number	Sentence words	Emotion
train 1	Step by step, we succeed	joy
train 2	I step on shit	sad
train 3	I trip on step	sad
train 4	The trip is shit	sad
test 1	We succeed	?

编码特征选取

Step by step, we succeed

X	step	by	we	succeed	I	on	shit	trip	the	is
One hot	1	1	1	1	0	0	0	0	0	0
TF	0.4	0.2	0.2	0.2	0	0	0	0	0	0
TF-IDF	0	0.06	0.06	0.06	0	0	0	0	0	0

- 编码方式：遍历两遍输入，第一遍数有多少个词确定向量长度，第二遍编码
- One hot：构建词表，化成词向量，单词出现设1，否则设0。
- TF：词频=**某词出现数**（频数）**/本文档词语总数**（标准化为频率）
- TF-IDF：TF和IDF的乘积
  - TF：词频
  - IDF：逆文档频率=**log(文档总数/(包含该词的文档数+1))** 【+1避免分母为0】
- 处理仅在测试集出现的词：
  - 可忽略，也可考虑其他的处理方式，如映射到随机向量（自行了解）

One hot 编码

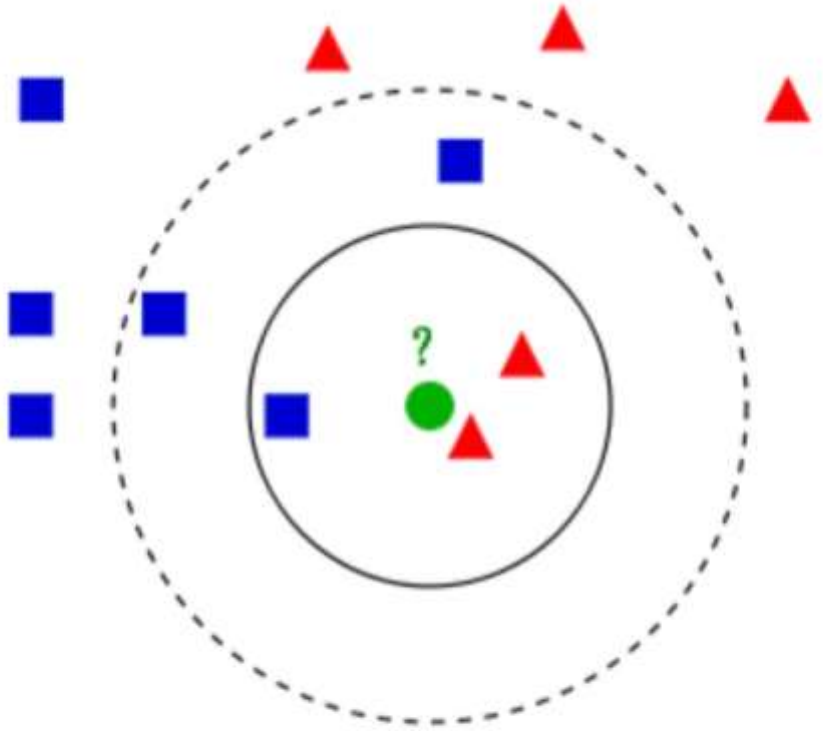
- 之前的例子的 one hot 矩阵

Document number	step	by	we	succeed	I	on	shit	trip	the	is
train 1	1	1	1	1	0	0	0	0	0	0
train 2	1	0	0	0	1	1	1	0	0	0
train 3	1	0	0	0	1	1	0	1	0	0
train 4	0	0	0	0	0	0	1	1	1	1
test 1	0	0	1	1	0	0	0	0	0	0

TF-IDF

- 如果某个词或短语t_i在一篇文章d_j中出现的频率 TF 高，并且包含这个词语的文章出现的频率 DF 低，则认为此词或者短语具有很好的类别区分能力，适合用来分类
- $TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$
  - $n_{i,j}$表示词条t_i在文档d_j出现的次数
- $IDF_i = lg \frac{|D|}{\left|\{j \mid t_i \in d_j\}\right| + 1}$
  - $|D|$为文档总数
  - $\left|\{j \mid t_i \in d_j\}\right|$为$t_i$出现的文档数
- $TFIDF_{i,j} = TF_{i,j} \times IDF_i$，即TF-IDF是TF和IDF的乘积

k-NN



半径大小 表示 K值大小

- k-nearest neighbours **classifier**:

$$f(q) = maj\left(g\left(\Phi_{X,k}(q)\right)\right)$$

- 其中:
  - $\Phi_{X,k}(q)$: 返回训练集X中距离q最近的k个样本
  - $g(\cdot)$: 返回（训练）样本的标签
  - $maj(\cdot)$: 返回众数

k-NN处理分类问题

1. 处理成 one hot 矩阵（或别的特征）
2. 相似度计算：计算test1与每个train的距离

- L_p距离(闵氏距离)：$L_p(x_i, x_j)=\left\{\sum_{l=1}^{n}\left|x_i^{(l)}-x_j^{(l)}\right|^p\right\}^{\frac{1}{p}}$
  - $p=1$，曼哈顿距离
  - $p=2$，欧氏距离（一般说的距离）
  - $p=\infty$ 是什么距离？

- 余弦相似度：$\cos\left(\vec{A},\vec{B}\right)=\frac{\vec{A}\cdot\vec{B}}{\left|\vec{A}\right|\left|\vec{B}\right|}$，其中$\vec{A}$和$\vec{B}$表示两个文本特征向量
  - 值越大，两个向量越相似

3. 类别计算：最相似的k个样本之标签的众数

- 之前的例子，若k=1，test1的标签即为train1的标签joy；
- 若k=4，test1的标签为train1,train2,train3,train4的标签中数量较多的，即为sad。

k-NN参数设置

- 通过验证集对参数（k值）进行调优
  - k过大：学习样本更多，会引入更多的噪音 → 可能存在欠拟合的情况;
  - k过小：参考样本少 → 容易出现过拟合的情况
  - 经验公式：一般取$k=\sqrt{N}$，N为训练集实例个数
    - k取N时，实际只会输出训练集的众数（joy，三成左右）
    - 大家可以尝试一下取不同的k

- 权重归一化（感兴趣的同学可以了解）

Name	Formula	Explain
Standard score	$X' = \frac{X-\mu}{\sigma}$	μ is the mean and σ is the standard deviation
Feature scaling	$X' = \frac{X-X_{min}}{X_{max}-X_{min}}$	X_{min} is the min value and X_{max} is the max value

朴素贝叶斯

- 要求某个特征组合T出现在某个类别c_j的概率 $P(c_j|T)$
- 贝叶斯公式

$$P(c_j|T) = \frac{P(T|c_j)P(c_j)}{P(T)}$$

- 假定特征之间相互独立

$$P(c_j|T) = \frac{\prod_{t_i \in T} P(t_i|c_j)\, P(c_j)}{\prod_{t_i \in T} P(t_i)}$$

- 因此使用$P(t_i|c_j)$、 $P(t_i)$和$P(c_j)$可以计算样本属于类别c_j的概率

朴素贝叶斯处理分类问题

- 之前的例子

$$P(c_{joy}|we\ succeed) = \frac{P(we|c_{joy})P(succeed|c_{joy})P(c_{joy})}{P(we)P(succeed)}$$

$$P(c_{sad}|we\ succeed) = \frac{P(we|c_{sad})P(succeed|c_{sad})P(c_{sad})}{P(we)P(succeed)}$$

- 比较两个概率的大小
  - 从训练集中计算出$P(c_{joy})$、$P(we|c_{joy})$、$P(succeed|c_{joy})$、 $P(c_{sad})$、$P(we|c_{sad})$、 $P(succeed|c_{sad})$
  - 分母一致，无需参与比较

朴素贝叶斯处理分类问题

- 关键是求出$P(t_i|c_j)$，单词t_i在情感类别c_j下出现的概率
- 在TF-IDF特征下

$$P(t_i|c_j) = \frac{\sum_{class(d_k)=c_j} TFIDF_{i,k}}{\sum_{u \in U_{all}} \sum_{class(d_k)=c_j} TFIDF_{u,k}}$$

即t_i的TF-IDF值的总和在c_j类的文章d_k的占比，U_{all}为总单词集

- 然而对训练集中未出现的单词，上面的计算会为0
- Laplace平滑：为每个单词的权重加$\lambda(\lambda \geq 0)$

$$P(t_i|c_j) = \frac{\lambda + \sum_{class(d_k)=c_j} TFIDF_{i,k}}{\sum_{u \in U_{all}} (\lambda + \sum_{class(d_k)=c_j} TFIDF_{u,k})}$$