实验二：k近邻算法
——分类和回归

PPT制作：李彦良，黄行昌
出题人　：詹雪莹，王耀威

- 要记住，几乎所有的有监督的机器学习模型，都遵循着这样的步骤：给出带有标签的数据集，进行模型训练，学习到新模型后，再给出不带有标签/正确答案的数据集，用于预测结果。可以类比上课与考试。
- 无监督以后再说。

k-NN处理分类问题

- 输出：类标签
- 分类：多数投票原则

Document number	The sentence words	emotion
train 1	I buy an apple phone	happy
train 2	I eat the big apple	happy
train 3	The apple products are too expensive	sadnesss
test 1	My friend has an apple	?

Table 2.1: example of classification dataset

数据集

Document number	The sentence words	emotion
train 1	I buy an apple phone	happy
train 2	I eat the big apple	happy
train 3	The apple products are too expensive	sadnesss
test 1	My friend has an apple	?

处理成one-hot矩阵

Document number	I	buy	an	apple	...	friend	has	emotion
train 1	1	1	1	1	...	0	0	happy
train 2	1	0	0	1	...	0	0	happy
train 3	0	0	0	1	...	0	0	sadness
test 1	0	0	1	1	...	1	1	?

计算test1与每个train的距离：

欧氏距离：

$$d(train1, test1) = \sqrt{(1-0)^2 + (1-0)^2 + \cdots + (0-1)^2} = \sqrt{6};$$

$$d(train2, test1) = \sqrt{(1-0)^2 + (1-0)^2 + \cdots + (0-1)^2} = \sqrt{8};$$

$$d(train3, test1) = \sqrt{(0-0)^2 + (0-0)^2 + \cdots + (0-1)^2} = \sqrt{9};$$

（也可以使用其他距离度量方式）

若k=1，test1的标签即为train1的标签happy；

若k=3，test1的标签为train1,train2,train3的标签中数量较多的，即为happy。

k-NN处理回归问题

- 输出：属于该标签的概率

Document number	The sentence words	the probability of happy
train 1	I buy an apple phone	0.8
train 2	I eat the big apple	0.6
train 3	The apple products are too expensive	0.1
test 1	My friend has an apple	?

Table 2.3: example of regression dataset

数据集

Document number	The sentence words	the probability of happy
train 1	I buy an apple phone	0.8
train 2	I eat the big apple	0.6
train 3	The apple products are too expensive	0.1
test 1	My friend has an apple	?

处理成one-hot矩阵

Document number	I	buy	an	apple	...	friend	has	probability
train 1	1	1	1	1	...	0	0	0.8
train 2	1	0	0	1	...	0	0	0.6
train 3	0	0	0	1	...	0	0	0.1
test 1	0	0	1	1	...	1	1	?

计算test1与每个train的距离，把该距离的倒数作为权重，计算test1属于该标签的概率：

$$P(test1\ is\ happy) = \frac{train1\ probability}{d(train1, test1)} + \frac{train2\ probability}{d(train2, test1)} + \frac{train3\ probability}{d(train3, test1)}$$
$$= 0.47$$

思考：为什么是倒数呢？

不同距离度量方式

- 距离公式：

  L_p距离(所有距离的总公式)：

  - $L_p(x_i, x_j)=\left\{\sum_{i=1}^{n}\left|x_i^{(l)}x_j^{(l)}\right|^p\right\}^{\frac{i}{p}}$

  - $p=1$：曼哈顿距离；
  - $p=2$：欧式距离，最常见。
  - （思考：在矩阵稀疏程度不同的时候，这两者表现有什么区别，为什么？）

- 余弦角公式：

  $\cos\left(\vec{A},\vec{B}\right)=\frac{\vec{A}\cdot\vec{B}}{\left|\vec{A}\right|\left|\vec{B}\right|}$，其中$\vec{A}$和$\vec{B}$表示两个文本向量；

  余弦值作为衡量两个个体间差异的大小的度量，为正且值越大，表示两个文本差距越小，为负代表差距越大，请大家自行脑补两个向量余弦值。

更多实验方法提高准确率

- 使用不同的距离度量方法
- 尝试改变k的值(这里规定一下，上限不能超过64)
- 在回归问题中对权重进行归一化（2种方式）

  防止计算出来的概率值大于1

  PS：关于k的经验公式：一般取$k=\sqrt{N}$，N为训练集实例个数，大家可以尝试一下

训练集，验证集，测试集的区别

- 训练集（training set）：给出了标准答案，相当于平时练习。用来训练模型或确定模型参数的，如k-NN中权值的确定等。
- 验证集（validation set）：给出标准答案，用来确定网络结构或者控制模型复杂程度的参数，修正模型。相当于模拟考试。
- 测试集（testing set）：没有给出标准答案，用于检验最终选择最优的模型的性能如何。相当于期末考试。
- 一个典型的划分是训练集占总样本的50%，而其它各占25%，三部分都是从样本中随机抽取。
- 本次实验用于分类的数据集只有训练集和测试集。用于回归的数据集给出了训练集，验证集和测试集。
- validation.xlsx文件用于在验证集上一个结果的评估，使用相关系数，大家把500个验证集上的预测结果，粘贴在Predict工作表中，右边会产生结果。Standard工作表不要修改内容。

实验任务

- 分类（使用准确率进行衡量结果）
  - 必须实现1-NN，使用欧氏距离，在246个训练集文本上进行训练，再在1000个测试文本上进行预测，得到所有测试文本的情感预测结果，与标准答案（test文本第二列）进行对比，将准确率记录在实验报告上
  - 可以尝试调K的值和使用不同的距离度量方式来提高准确率，并在报告中写下你的发现，或者有其他的新方法也请记录在报告中

- 回归（使用相关系数进行衡量结果）
  - 使用k-NN处理回归问题，得出所有500个测试文本属于每个情感（一共6种）的概率，计算出在验证集上的相关系数（使用validation.xlsx文件进行计算），并记录在实验报告中，尝试优化参数提升在验证集上的准确率（如果有）
  - 提交在测试集的预测结果，命名为"学号_姓名拼音_regression.txt"，提交到相关FTP文件夹中，result文件内部格式为

    testx P1 P2 P3 P4 P5 P6 （x为test文本序列号，P为对应情感概率）

    提示：请记得检查你们6种情感概率相加是否为1

注意事项

1、作业提交地址

ftp://my.ss.sysu.edu.cn/~ryh

2、命名方式

- 实验报告：请按照模板写，提交为：学号_拼音名字.pdf。
- 实验代码：同一个算法请尽量写成一份代码，提交为：学号_拼音名字.xxx，后缀视使用语言而定，除非你有多个文件需要提交，否则不要压缩！。
- 实验结果：看前面一页。

3、编程语言可用c++, python, matlab, java，不能使用现成库->这里指的是，不能使用直接调用所要求实现算法的库，不是说不能用STL！，否则扣分

4、 提交截止时间

2016年09月25日23：59：59前提交至FTP对应文件夹，否则视为迟交