Lab 2: Regression Models based on kNN and NB

1. Objective

Know how to implement and compare regression models based on kNN and NB.

已知 246 篇训练文本对应的公众在 anger (愤怒)、disgust (厌恶)、fear (害怕)、joy (高兴)、sad (悲伤)、surprise (惊讶)这六种情感上的概率值，预测 1000 篇测试文本对应的公众在上述六种情感上的概率值各是多少？

2. Dataset

Dataset_words.txt：文档 ID、词（以空格分隔）。该数据集包含 246 篇训练文本及 1000 篇测试文本的词列表。

Dataset_words_anger.txt：文档 ID、词（以空格分隔）、公众“愤怒”的概率。该数据集包含 246 篇训练文本，每篇训练文本既有词列表，也有标准答案（即公众“愤怒”的概率值）；另有 1000 篇测试文本，每篇测试文本只有词列表，其公众“愤怒”的概率值需要大家预测。

Dataset_words_disgust.txt：类似于“Dataset_words_anger.txt”，只是第三列为“厌恶”。

Dataset_words_fear.txt：类似于“Dataset_words_anger.txt”，只是第三列为“害怕”。

Dataset_words_joy.txt：类似于“Dataset_words_anger.txt”，只是第三列为“高兴”。

Dataset_words_sad.txt：类似于“Dataset_words_anger.txt”，只是第三列为“悲伤”。

Dataset_words_surprise.txt：类似于“Dataset_words_anger.txt”，只是第三列为“惊讶”。

gold_train 文件夹：246 篇训练文本在六种情感上的标准答案，供参考。

AILab 文件夹：运行其中的 RunResult.bat 得到相关系数值，请阅读其中的 readme.txt。

该数据集是 SemEval-2007 的国际竞赛用数据（http://nlp.cs.swarthmore.edu/semeval/），截止到目前，国际上在这个数据集上表现较佳的方法被称为 SWAT，其性能如下：

Anger	24.51
Disgust	18.55
Fear	32.52
Joy	26.11
Sadness	38.98
Surprise	11.82

例如，SWAT 方法在 anger 这种情感上，其预测的概率值和真实值之间的相关系数为 0.2451。

参考论文：Katz, P., Singleton, M., & Wicentowski, R. (2007). SWAT-MP: The SemEval-2007 Systems for Task 5 and Task 14. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, ACL (pp. 308-313).

3. kNN【本部分提交的截止时间为 10 月 28 日 23:00，鼓励当场提交】

kNN 是 k 最近邻的简称。当 k 取值为 1 时，即采用最相近的那篇训练文本的标准答案进行预测，该方法参考“Lab 1.pdf”中的“3. Processes”以及“4. More methods”。

请采用你在 Lab 1 中实现的代码，运行在本次实验（Lab 2）的数据集上，记录 anger (愤怒)、disgust (厌恶)、fear (害怕)、joy (高兴)、sad (悲伤)、surprise (惊讶)这六种情感上的相关系数。将你认为**最好**的一个结果（比如以上六种相关系数值的平均值最大的一组）上传到 FTP。

【上传文件】

(1) 实验结果文件：参考 99999999.txt，先写上 knn，然后空格，然后是你的方法在 anger (愤怒)、disgust (厌恶)、fear (害怕)、joy (高兴)、sad (悲伤)、surprise (惊讶)这六种情感上的相关系数值（以空格分隔），将该文件上传到 FTP 的 Lab 2 results 目录中。

(2) 实验报告文件：如 99999999.doc 或 99999999.pdf，在实验报告中阐述你的实验方法，并将该文件上传到 FTP 的 Lab2 reports 目录中。

【备注】由于数据集规模稍大，请充分预留好代码运行的时间。

4. NB【本部分提交的截止时间为 11 月 04 日 23:00，鼓励当场提交】

NB 是朴素贝叶斯的简称。下面以一个简单的数据集为例，阐述基于 NB 的回归/预测模型：

DocumentID	Words (split by space)	joy
train1	sheva delight us	0.6
train2	goal delight for sheva	0.7
test1	sheva goal	?

上述三篇文本的词列表如下：

DocumentID	Words (split by space)
train1	sheva delight us
train2	goal delight for sheva
test1	sheva goal

基于实验一（Lab 1.pdf）中的“4.1 更改向量中值的表示方法”，可以将上述两篇训练文本（train1 和 train2），以及一篇测试文本（test1）转为如下向量格式：

DocumentID	sheva	delight	us	goal	for
train1	0.33	0.33	0.33	0	0
train2	0.25	0.25	0	0.25	0.25
test1	0.5	0	0	0.5	0

接下来，就是基于 NB 的回归模型如何在已知 train1 和 train2 的标准答案（即公众感到“joy”的概率）分别为 0.6 和 0.7 的前提下，预测 test1 对应的公众感到“joy”的概率值。

为了便于模型的推导，我们将文本 train1 记为 d1，train2 记为 d2，test1 记为 d3，情感 joy 记为 j，词 sheva 记为 s，词 goal 记为 g。我们要估计的是 p(d3, j)，即文本 test1 和情感 joy 的联合概率值（也就是说，100 名用户看到 test1，会有多大的可能性将它关联到 joy）。由于 d3 包含 s 和 g 两个词，因此要估计的 p(d3, j)近似等于 p(s, g, j)，即词 sheva、goal 和情感 joy 的联合概率值。基于 NB 的回归模型利用所有的训练文本及其标准答案来估计这个值：

p(s, g, j) = p(d1, s, g, j) + p(d2, s, g, j)……概率的加法法则，参照 Lec 2 课件第 9 页

其中，p(d1, s, g, j) = p(d1, j, s, g) = p(d1, j)p(s | d1, j)p(g | d1, j, s)……概率的乘法法则，同上

上式中，p(d1, j) = 0.6；p(s | d1, j) = 0.33；p(g | d1, j, s) = 0，这里假设给定每篇文本和情感的前提下，词与词之间是独立的，也就是说 p(g | d1, j, s) = p(g | d1, j) = 0。

所以，p(d1, s, g, j) = 0.6*0.33*0 = 0。

同理，p(d2, s, g, j) = 0.7*0.25*0.25 = 0.044。

所以，p(s, g, j) = 0 + 0.044 = 0.044。即，基于 NB 的回归模型会将 test1 对应的公众感到“joy”的概率值预测为 0.044。

【编程实现的一点技巧】

首先，基于实验一（Lab 1.pdf）中的“4.1 更改向量中值的表示方法”，得到所有训练文本和测试文本的向量文件，如下所示：

DocumentID	sheva	delight	us	goal	for
train1	0.33	0.33	0.33	0	0
train2	0.25	0.25	0	0.25	0.25
test1	0.5	0	0	0.5	0

然后，对于每一篇测试文本，比如 test1，采用下述流程计算其预测的概率值：

(1) 读取 test1 的词向量，即(0.5, 0, 0, 0.5, 0)，输出向量值大于 0 的维度，这里是第一维（sheva）和第四维（goal）；

(2) 读取全部训练集的词向量，即 train1 的(0.33, 0.33, 0.33, 0, 0)和 train2 的(0.25, 0.25, 0, 0.25, 0.25)；以及全部训练集的标准答案（存放在 gold_train 文件夹中），即 train1 的 0.6 和 train2 的 0.7；

(3) 将 0.6 乘以 train1 向量的第一维，再乘以 train1 向量的第四维，以及 0.7 乘以 train2 向量的第一维，再乘以 train2 向量的第四维的总和，作为 test1 的概率预测值。即，0.6*0.33*0 + 0.7*0.25*0.25 = 0.044。

【备注】步骤(1)中向量值大于 0 的维度若有 10 个，则步骤(3)中连乘的维度也是这 10 个。

【上传文件】

(1) 实验结果文件：参考 99999999.txt。在实现了“3. kNN”的基础上，在第二行先写上 nb，然后空格，然后是该方法在 anger (愤怒)、disgust (厌恶)、fear (害怕)、joy (高兴)、sad (悲伤)、surprise (惊讶)这六种情感上的相关系数值（以空格分隔），将该文件上传到 FTP 的 Lab 2 results 目录中。

(2) 实验报告文件：如 99999999.doc 或 99999999.pdf，在实验报告中阐述你的代码截图、实验结果对比及思考，并将该文件上传到 FTP 的 Lab2 reports 目录中。