垃圾邮件分类

在这个问题中，我们将使用朴素贝叶斯算法和支持向量机（SVM）构建一个垃圾邮件分类器。近年来，电子媒体上的垃圾信息已经成为一个越来越令人担忧的问题。在这里，我们将构建一个分类器来区分真实消息和垃圾消息。这次作业，我们将建立一个用于检测短信垃圾消息的分类器。我们将使用由 Tiago A. Almeida 和 José María Gómez Hidalgo 开发的 SMS 垃圾短信数据集，该数据集可以在 http://www.dt.fee.unicamp.br/~tiago/smsspamcollection 公开获取。

我们已将此数据集分为训练集和测试集，并将它们包含在此作业中，分别命名为 *data/ds6_spam train.tsv* 和 *data/ds6_spam test.tsv*。有关该数据集的更多详细信息，请参阅 *data/ds6_readme.txt*。请不要重新分发这些数据集文件。此作业的目标是从头开始构建一个分类器，通过短信消息的文本区分垃圾消息和非垃圾消息。

（1）实现用于处理垃圾消息的代码，将其转换为可以输入到机器学习模型中的 numpy 数组。通过完成我们提供的 *src/p06_spam.py* 中的 get_words、create_dictionary 和 transform_text 函数来实现这一点。请注意每个函数的相应注释，其中包含有关所需的具体处理的说明。

我们提供的代码将运行你的函数，并将结果字典保存到 *output/p06_dictionary*，将结果训练矩阵的样本保存到 *output/p06_sample_train_matrix*。

（2）在这个问题中，你将实现一个朴素贝叶斯分类器，用于使用多项事件模型和拉普拉斯平滑进行垃圾邮件分类（有关拉普拉斯平滑的详细信息，请参考朴素贝叶斯的课堂笔记）。

通过完成 *src/p06_spam.py* 中的 fit_naive_bayes_model 和 predict_from_naive_bayes_model 函数来编写你的实现。*src/p06_spam.py* 需要能用于训练一个朴素贝叶斯模型，计算你的预测准确性，然后将结果保存到 *output/p06_naive_bayes_predictions*。

注意：如果你按照直接的方式实现朴素贝叶斯，你会发现计算得到的$p(x|y) = \prod_i p(x_i|y)$往往等于零。这是因为$p(x|y)$，它是许多小于 1 的数的乘积，是一个非常小的数。标准计算机表示实数的方法无法处理太小的数，而是将它们四舍五入为零（这被称为“下溢”）。你需要找到一种方法来计算朴素贝叶斯预测的类标签，而不需要显式表示诸如$p(x|y)$这样的非常小的数。

提示：你可以考虑使用对数。

（3）直觉上，某些标记可能特别表明一条短信属于特定的类别。我们可以通过查看以下内容来试图非正式地了解标记i对于“垃圾邮件”类的表现有多明显：

$$\log \frac{p(x_j = i|y = 1)}{p(x_j = i|y = 0)} = \log \left(\frac{P(\text{token } i|\text{email is SPAM})}{P(\text{token } i|\text{email is NOTSPAM})} \right)$$

完成提供的代码中的 get_top_five_naive_bayes_words 函数，使用上述公式以获取最具指示性的 5 个标记。提供的代码将打印出结果的指示性标记，然后将它们保存到 *output/p06_top_indicative_ words*。

（4）支持向量机（SVM）是课堂上讨论的另一种机器学习模型。我们在 *src/svm.py* 中提供了一个 SVM 实现（使用径向基函数（RBF）核），你不需要修改那段代码。

训练使用 RBF 核参数的支持向量机的一个重要部分是选择适当的核半径。完成计算最佳 SVM 半径的工作，编写代码以计算在验证数据集上最大化准确性的最佳 SVM 半径。提供的代码将使用你的计算最佳 SVM 半径的结果，然后将最佳半径写入到 *output/p06_optimal_radius*。

注意：要完成本次作业，你需要将附件中提供的 *src/p06_spam.py* 代码补充完整，完成后运行该脚本，即可在 */output* 中得到题目中所提及的结果文件。请将补充完整的代码以及结果文件一同打包上传，**请不要**将代码写到 word 或者 pdf 中提交。