人工智能基础

编程作业 2

http://staff.ustc.edu.cn/~linlixu/ai2014spring/

实验截止时间：2014/6/22

助教：

王臻　wang1231991@126.com
李亦锬　daniyitan@gmail.com
仲小伟　zhxwmessi@gmail.com

实验目的：

本次实验考虑机器学习中传统的监督学习问题，基于两个经典应用数据集：垃圾邮件和手写数字图片，并结合课上介绍的相应学习算法，在数据集上分别进行实验，以加强对相关算法原理及应用的理解。

数据集介绍：

1. 邮政系统手写数字数据集(usps)，十分类问题。每个样本对应了一个 16*16 的灰度图像，图片内容为 0~9 中某个数字，每个样本图片可以用 imshow 函数在 matlab 里显示。
2. UCI 垃圾邮件数据集(spam)，二分类问题。每封邮件相应的特征描述为 54 维的向量，向量每个维度分别代表词库中的某个单词，对应维度值为 1 代表该维度所对应的单词在此邮件中出现，为 0 则代表没有出现。

注：我们对每个数据集进行了一定的划分，保留了每个数据集的一部分作为对实验结果的评价之一，余下的部分放在了课程主页上供下载。

训练与测试

在监督学习中，训练数据带有标号，在训练的过程中需要从训练数据 traindata 和其对应的标号 trainlabel 中学习相应的分类模型。

在测试过程中，用学习到的模型对测试集中的数据 testdata 作预测，并将预测结果与测试数据的真实标签 testlabel 进行比较，从而度量分类模型的性能。

$$\text{Accuracy} = \frac{\sum_{i \in test\ set} I(\text{predict}_i = testlabel_i)}{\#of\ test\ size}$$

实验要求：

Part1. 实现一个朴素贝叶斯分类器(15%)

提交一个 Matlab 函数 nbayesclassifier，函数形式为

```
function [ypred,accuracy]= nbayesclassifier (traindata,
                    trainlabel, testdata, testlabel, threshold)
```

其中 threshold 为用于判断类别的后验概率的阈值，即如果 P(spam|email) > threshold 则判别为 spam。要求函数返回对测试数据的预测 ypred，以及通过与真实标号比较计算得到的分类正确率 accuracy。ypred 与 trainlabel 和 testlabel 形式相同。

Part 2.实现一个最小二乘分类器(引入规范化项后)(10%)

1. 对引入了 L2 规范化项之后的最小二乘分类问题进行推导。即求解以下优化问题：

$$min_w(X\boldsymbol{w}-y)^2+\lambda\|\boldsymbol{w}\|^2$$

2. 基于 1 中的结果，实现并提交一个 Matlab 函数 lsclassifier

```
function [ypred,accuracy] = lsclassifier(traindata, trainlabel,
                              testdata, testlabel, lambda)
```

Part 3.实现一个支持向量机分类器 (15%)

提交一个 Matlab 函数 softsvm

```
function  [ypred,accuracy]  =  softsvm(traindata,  trainlabel,
                               testdata, testlabel , sigma, C)
```

其中 C 为 soft margin SVM 的控制参数，sigma 为控制核函数的参数，当 sigma=0 时，使用线性核函数$\mathrm{K}(x_i,x_j)=x_i{}^Tx_j$，其他情况则使用 RBF 核函数$\mathrm{K}(x_i,x_j)=e^{-\frac{\|x_i-x_j\|^2}{sigma^2}}$

注意：手写数字识别是一个多分类问题，基于对应数据集中含有 10 种不同的数字，因此该问题的类别数目为 10。因此这里需要将传统的二分类 SVM 算法扩展至多分类。SVM 多分类方法建议采用 One vs All (One against All)，该方法的描述如下：对于每一个类别的数据学习一个 SVM 模型$\mathrm{f_k}(\mathbf{x})=\mathbf{w}_\mathrm{k}^\mathrm{T}\varphi(\mathbf{x})+\mathrm{b_k}$，假设数据中有 4 个类别 1、2、3、4，那么分别以（1 作为正例，234 作为反例）、（2 作为正例，134 作为反例）、（3 作为正例，124 作为反例）、（4 作为正例，123 作为反例）学习 4 个分类器，对于新来的样例进行判别时，以返回值最大的分类器对应的正例类别作为预测的类别：$\hat{\mathrm{y}}=\arg\max_\mathrm{k}\mathrm{f_k}(\mathbf{x})$。实验数据中有 10 个类别，请扩展以上方法，应用于 10 类情况。**(对于每个分类器，请采用一样的 C 和 sigma 参数)**

对于手写数字识别数据，trainlabel 和 testlabel 取值范围为{1,2,3,4,5,6,7,8,9,10}，其中标号为 1 的样例实际上为数字 0 的手写图片，标号为 2 的样例则实质上为数字 1 的手写图片，而其他标号类似，为保持统一，ypred 的取值范围同样为{1,2,3,4,5,6,7,8,9,10}

Part 4.在不同数据集上使用交叉验证选择各个算法的参数(15%)

实现交叉验证（代码需要提交），在各个数据集上：

- 使用 5-fold 交叉验证为每个算法挑选适当的参数(Naïve Bayes 中的 threshold，最小二乘法中的 Lambda，SVM 中的 sigma 和 C)；
- 对每一个算法：
  - 返回一个矩阵，表示每一个参数（参数组合）在每一个 fold 上的正确率（若有 10 个参数，则返回 10x5 的矩阵）；
  - 挑选在 5 个 fold 中平均正确率最高的参数（参数组合）

在实验报告中需要记录交叉验证的结果，即对于每个参数(参数组合)在 5 个 fold 上的平均正确率。
注：Naïve Bayes 方法只需在 spam 数据上实验，即实验部分包括以下内容：

Algorithms	Naïve Bayes	Least Squares	SVM
Spam (2 classes)	√	√	√
USPS (10 classes)		√	√

Part 5.实验报告(25%)

总结以上的实验结果，并对实验结果进行分析。

Part 6.实验测试结果评价(20%)

对于这部分，保存每个算法在相应数据集上对应的最佳参数并提交。如对于分类算法，需要保存 Naïve Bayes 和 SVM 在相应数据集上使用 5-fold 交叉验证得到的参数(Naïve Bayes 的 threshold，Least Squares 的 lambda，SVM 的 sigma 和 C)，保存文件名统一为“数据集名”_parameters.mat。我们将会基于你们的算法代码以及最优参数，在保留下来的一部分数据上进行测试，并度量各个算法的性能。

备注：

1.矢量化编程是提高算法速度的一种有效方法，其思想就是尽量使用高度优化的

数值运算操作来实习学习算法。例如，假设$x \in R^n$和 $y \in R^n$为向量，需要计算 $z = x^T y$，在 Matlab 中可以用以下方式实现：

```
z = 0;
for i = 1 : n
    z = z + x(i) * y(i);
end
```

或者可以更简单的写为：

```
z = x' * y;
```

很显然，第二段程序代码不仅简单，而且运行速度更快。

通常，一个编写 Matlab 程序的诀窍是：**代码中尽可能避免显示的 for 循环**

2.SVM 求解二次优化问题可以使用 Matlab 函数 quadprog，可以输入 help quadprog 查看函数使用帮助。

3.实验 Part4 部分当训练数据比例占整个数据集比例较大时，程序会运行的比较慢，所以一定要注意尽量不要在程序中使用 for 循环，大部分运行耗时的 for 循环都可以被矢量化。

4.提交格式为“学号_姓名.rar”，除了包含必须的.m 文件之外，还需要把在 part4 中用 5-fold 交叉验证得到的各个函数对应正确率最好的参数保存到“数据集名”_parameter.mat 文件中同时提交，该.mat 文件中应该只有 4 个参数（分别名为 threshold, lambda, sigma, C）。

5.Naive Bayes 算法中的 threshold 的取值可以从[0.5 0.6 0.7 0.75 0.8 0.85 0.9]中取值；最小二乘分类器中的 lambda 可以从[1e-4 0.01 0.1 0.5 1 5 10 100 1000 5000 10000]中取值；SVM 中的参数有高斯核参数 sigma 以及 C，其中 sigma 的取值范围由数据决定：假设数据集为$\{(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n)\}$，令 $d = \frac{\sum_{i,j}(\mathbf{x_i} - \mathbf{x_j})^2}{n^2}$，则 sigma 从[0.01d 0.1d d 10d 100d]中取值，C 可以取[1 10 100 1000]。