第 2 组的数据挖掘作业 3

SVM 分类

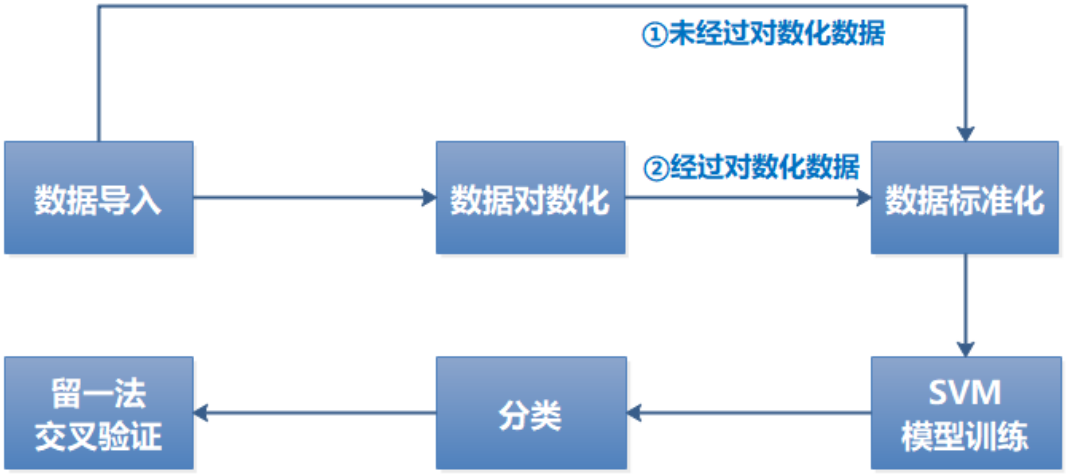
小组成员：高鹏昺 蒋世豪 李进雄 周亮 苏金涛 刘昊轩

完成人 15051317 高鹏昺

1. 利用全部特征进行分类

实验流程

实验采用了下图所示的数据处理、模型训练、模型检验流程。



实验首先导入 DLBCL_data_6285_77.txt 文件中的全部特征数据和 dlbcl_label.txt 文件中的标签数据。接着将特征数据分别在未经过对数化处理和经过对数化处理这 2 种情况下进行实验。数据在用于训练之前经过了 Z-score 标准化处理。每组实验又分别选择 linear、poly、sigmoid、rbf 这 4 种 SVM 核函数进行 SVM 模型训练。然后，使用训练完成的 SVM 模型对测试数据进行分类，这里是通过留一法(LeaveOneOut)交叉验证的，最后计算分类结果的正确率、错误率。

数据未经过对数化(log)处理的分类结果

SVM 核函数	分类错误数	正确率	错误率
linear	3	96.1039%	3.8961%
poly	18	76.6238%	23.3766%
sigmoid	6	92.2078%	7.7922%
rbf	11	85.7143%	14.2857%

数据经过对数化(log)处理后的分类结果

SVM 核函数	分类错误数	正确率	错误率
linear	2	97.4026%	2.5974%
poly	19	75.3247%	24.6753%
sigmoid	4	94.8052%	5.1948%
rbf	12	84.4156%	15.5844%

2. 特征选择后进行分类

利用全部的 6285 个特征显然数据的维度很高，因此我们在数据标准化之前加入特征选择的处理过程，选取分类效果较好的特征，降低数据的维度。本实验采用了 3 种不同的特征选择方法。

特征选择方法

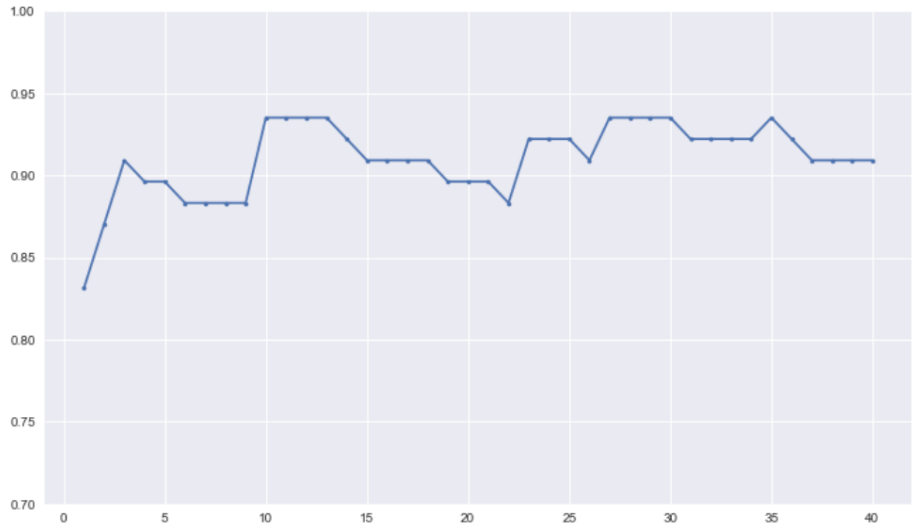
(1) Filter 过滤法中的卡方检验(chi2 test)

经典的卡方检验是检验定性自变量对定性因变量的相关性。假设自变量有 N 种取值，因变量有 M 种取值，考虑自变量等于i且因变量等于j的样本频数的观察值与期望的差距，构建统计量:

$$\chi^2 = \sum \frac{(A-E)^2}{E} = \sum_{i=1}^{k} \frac{(A_i - E_i)^2}{E_i}$$

这个统计量的含义简而言之就是自变量对因变量的相关性，然后我们可以结合卡方检验选取最好的k个特征。

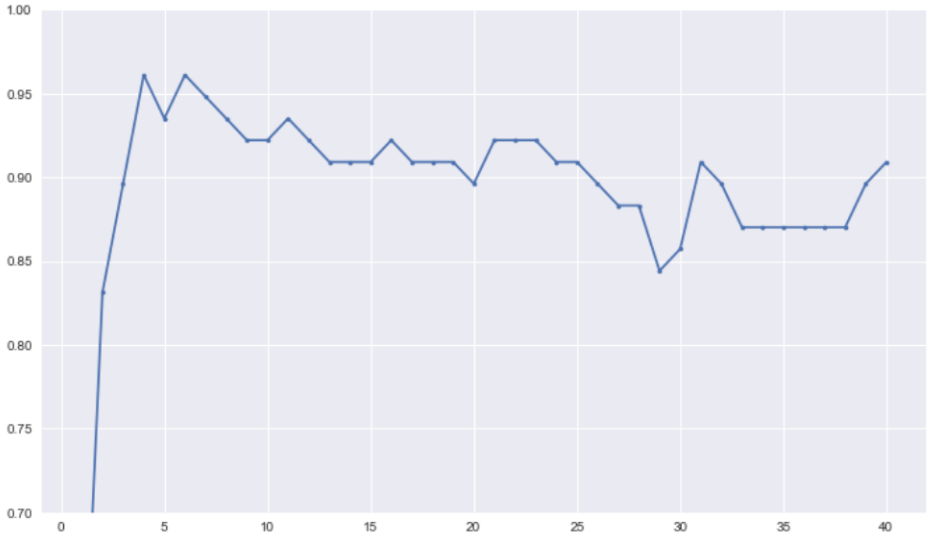
下图为未经过对数化(log)处理下，选取 1 到 10 个卡方检验效果最好的特征进行分类时的正确率。



下表表示最高分类正确率及选取的特征数。

SVM 核函数	选取特征数	分类错误数	正确率	错误率
linear	10	5	93.5065%	6.4935%

下图为经过对数化(log)处理下，选取 1 到 40 个卡方检验效果最好的特征进行分类时的正确率。



下表表示最高分类正确率及选取的特征数。

SVM 核函数	选取特征数	分类错误数	正确率	错误率
linear	4	3	96.1039%	3.8961%

(2) Embedded 集成法中的基于惩罚项的特征选择法

使用带惩罚项的基模型，除了筛选出特征外，同时也进行了降维。正则化是把额外的约束或者惩罚项加到已有模型（损失函数）上，以防止过拟合并提高泛化能力。实际上，L1 惩罚项降维的原理在于保留多个对目标值具有同等相关性的特征中的一个。

数据未经过对数化(log)处理

SVM 核函数	参数	选取特征数	分类错误数	正确率	错误率
linear	C=0.001	19	0	100.0000%	0.0000%

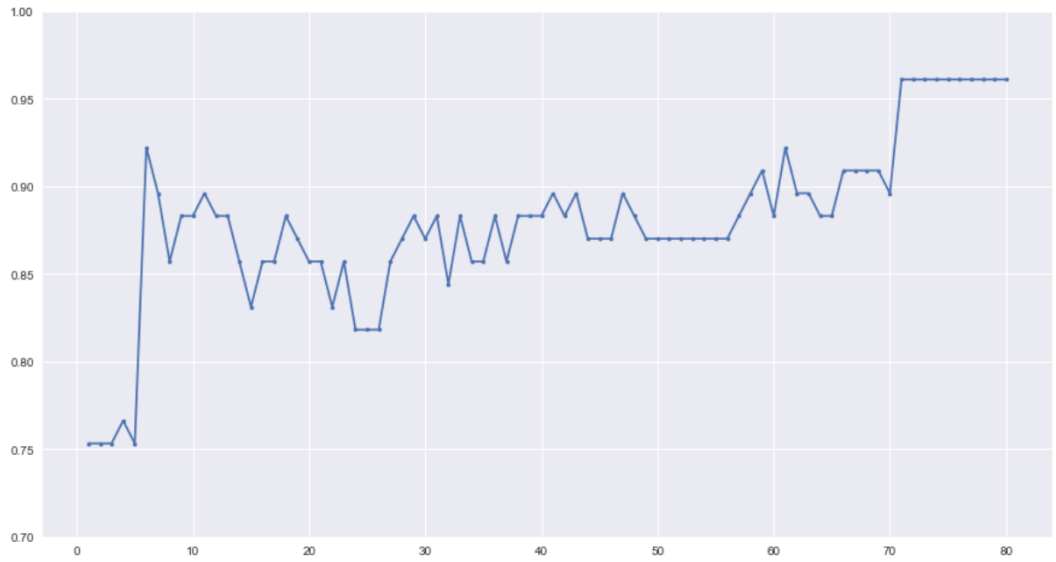
数据经过对数化(log)处理

SVM 核函数	参数	选取特征数	分类错误数	正确率	错误率
linear	C=0.25	33	0	100.0000%	0.0000%

(3) 数据降维:主成分分析法(PCA)

PCA 的思想是将n维特征映射到k维上$(k < n)$，这k维是全新的正交特征。这k维特征称为主成分，是重新构造出来的k维特征，而不是简单地从n维特征中去除其余$n - k$维特征。

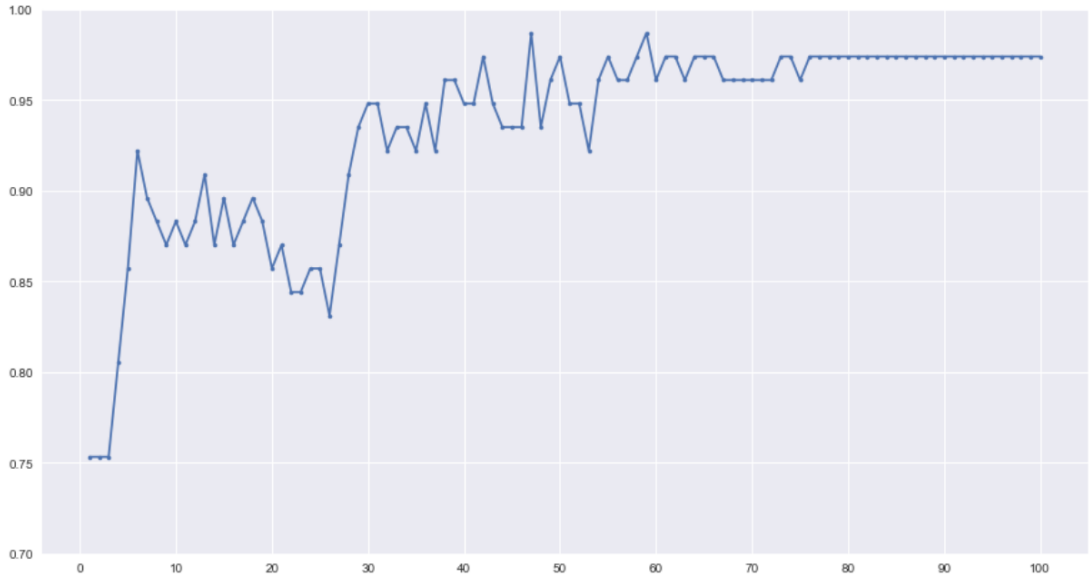
下图为未经过对数化(log)处理下，通过 PCA 降维到 1 到 80 个维度时的分类时的正确率。



下表表示最高分类正确率及降维后的特征数。

SVM 核函数	降维后特征数	分类错误数	正确率	错误率
linear	71	3	96.1039%	3.8961%

下图为经过对数化(log)处理下，通过 PCA 降维到 1 到 100 个维度时的分类时的正确率。



下表表示最高分类正确率及降维后的特征数。

SVM 核函数	降维后特征数	分类错误数	正确率	错误率
linear	47	2	98.7013%	1.2987%

3. 实验环境

编程语言： Python3.6　　Jupyter Notebook

主要工具： sklearn 库(其中的 SVM 模型使用了 libsvm 库)