# A ridge regression model of 158 biomarkers for the three stages of breast cancer

## Abstract

**Motivation:** Not all breast cancers are the same. Doctors use certain programs to collect stage and in order to formulate a treatment plan for patients. The stage of cancer is whether cancer cells have spread to the breast, if so , the cancer has spread in the body. In the breast cancer there are four major stages: stage i, stage ii, stage iii, stage iii and stage iv. This paper establishes a ridge regression model for predicting breast cancer stage.

**Results:** This study propose a ridge regression model to predict three-class(stage i, stage ii, stage iii and stage iv) accuracy achieved 81.6% with 158 selected features.

**Availability and Implementation:** In this paper, the data from breast cancer transcriptome of TCGA with 510 samples and 17814 features. Through the ridge regression model Stage i has 90 samples, of which 62 samples are accurate, the accuracy rate is 68.9%,Stage ii has 297 samples, of which 266 samples are accurate, the accuracy rate is 90%,Stage iii has 110 samples, Stage iv amount of data is too little only 13 samples, so stage iii and iv of samples together to discuss, of which 88 samples are accurate, the accuracy rate is 72%.

## 1 Introduction

Global incidence of breast cancer since the late 1970s has been the beginning of an upward trend.8 women in the United States will have 1 person suffering from breast cancer. This article explores the use of logistic regression to learn a problem of breast cancer stage(stage i, stage ii, stage iii and stage iv).Stage is usually expressed as a number on a scale of i through iii .Cancer stage is based on four characteristics:1)the size of the cancer;2)[whether the cancer is invasive or non-invasive](http://www.breastcancer.org/symptoms/diagnosis/invasive);3)whether cancer is in the lymph nodes;4)whether the cancer has spread to other parts of the body beyond the breast. The stage of the breast cancer can help you and your doctor understand your prognosis (the most likely outcome of the disease) and make decisions about treatment.

## 2 Material and Methods

### 2.1Datasets

BRCA transcriptome (agilentg4502a\_07\_3\_gene--leve3) with 17814 features data were downloaded from TCGA consortium Data Portal.clinical. According to tumor\_stage in project-TCGA-BRCA.2017-04-20T02\_01\_20.302397.json

the cancer is divided into three phases(stage i, stage ii, stage iii)and the stage X data is deleted. The number of samples is changed from 600 to 510,so the experimental study of 510 samples and 17814 features (the matrix of 510 columns and 17814 rows).

### 2.2 Using lasso regression to deal with six groups of two groups of data

The tumor\_stage was divided into four phases:stage i, stage ii,stage iii and stage iv, 90 samples in stage i,297 samples in stage ii, 123 samples in stage iii,There are 17814 features for each sample. Feature selection is a necessary step in this experiments due to the large numbers of features and small numbers of samples. Then, we choose lasso regression algorithm to do the feature selection. First, we divided the data sets into 6 classes and trained a binary model on each class. Second, some features were deleted if its score equals to zero. Third, we combine the features from the different classes and delete the same features, then got the total 905 features, which is regarded as the truly effective features to breast cancer classification issue. The result is predicting breast cancer stage three-class accuracy achieved 80.8% with 905 selected features. The confusion matrix under the optimal parameters is(alpha=0.01) :

[  54.   31.    5.  ]  
  [  8.   275.    14.  ]  
  [  2.    38.    83.  ]

Stage i has 90 samples, of which 54 samples are accurate, the accuracy rate is 60%,31 samples are predicted as stage ii and 5 samples are predicted as stage iii, the prediction error probability is 40%.Stage ii has 297 samples, of which 275 samples are accurate, the accuracy rate is 93%,8 samples are predicted as stage i and 14 samples are predicted as stage iii ,the prediction error probability is 7%.Stage iii contains 123 samples for the stage iii and stage iv ,of which 83 samples are accurate, the accuracy rate is 67.5%,2 samples are predicted as stage i and 38 samples are predicted as stage iii, the prediction error probability is 33%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| serial number | comparison of stage | sample numbers | feature numbers | acc |
| 1 | [i]vs[ii] | 387 | 276 | 0.97 |
| 2 | [i]vs[iii] | 213 | 180 | 1.00 |
| 3 | [ii]vs[iii] | 420 | 302 | 0.97 |
| 4 | [i,ii]vs[iii] | 497 | 321 | 0.95 |
| 5 | [i,iii]vs[ii] | 497 | 349 | 0.97 |
| 6 | [ii.iii]vs[i] | 510 | 303 | 0.98 |

### 2.3Create a ridge regression model to remove features

Ridge regression analysis is an essential statistical techniques to estimate the relationship among variables. After building a linear regression model, it is necessary to find out effective evaluation methods to measure the performance of the proposed model.

**2.3.1performance  measurement**

提出优化目标函数：opt=acc\*,acc表示预测逻辑回归三分类的准确性。（Goodness of Fit）拟合优度是指回归直线对观测值的拟合程度。拟合优度指标通常总结下的观测值和预期值之间的差异模型。的最大值为1，的值越接近1，说明回归直线对观测值的拟合程度越好，岭回归性能越好；反之的值越小，说明回归直线对观测值的拟合程度越差。评价性能测量标准有两个，分别是1）分类准确性（acc）高2）岭回归性能（）好；优化目标函数就是找到同时满足acc和都高的点。

**2.3.2循环删除算法**

在该算法中，我们选择岭回归作为回归模型。岭回归实质上是一种改良的最小二乘估计法，通过放弃最小二乘法的无偏性，以损失部分信息、降低精度为代价获得回归系数更为符合实际、更可靠的回归方法。又因为岭回归通过对最小二乘估计加入罚约束，使某些系数的估计为0，这为我们进行特征选择提供了基础。

利用 lasso模型进行特征选择之后，在 alpha = 0.1 的提交下下所求得的六个二分类的特征组合的并集就是循环删除算法所需要处理的特征子集，我们用字母 N 表示该特征子集的大小。我们设计的循环删除算法的思路如下： 在每一遍的循环中，利用岭回归对数据集进行处理，其回归系数作为评价特征的标准。我们给定 K 为每次循环后需要删除的特征个数。这意味着每一次循环中，我们会根据岭回归系数的绝对值对特征进行排序，其绝对值最低的K 个特征我们视为“冗余特征”，并将之删除。再下一次的循环中，我们要处理的特征个数就只剩下 N-K。假设循环n次之后，剩下的特征个数为 N- nK。

开始N=905

岭回归给特征打分

特征按分数排序

删除i个特征

（i取1、2、3、4、5）

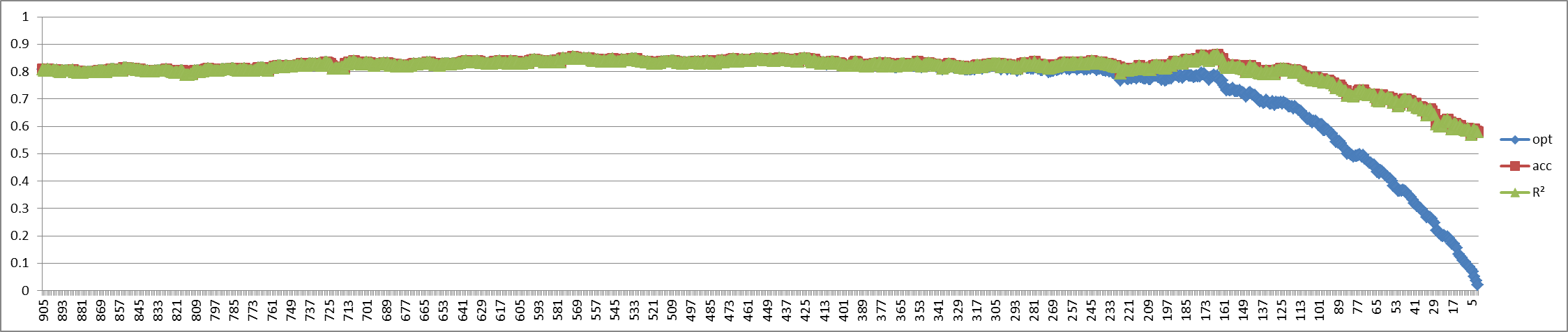
R＜＝0

R=N-i

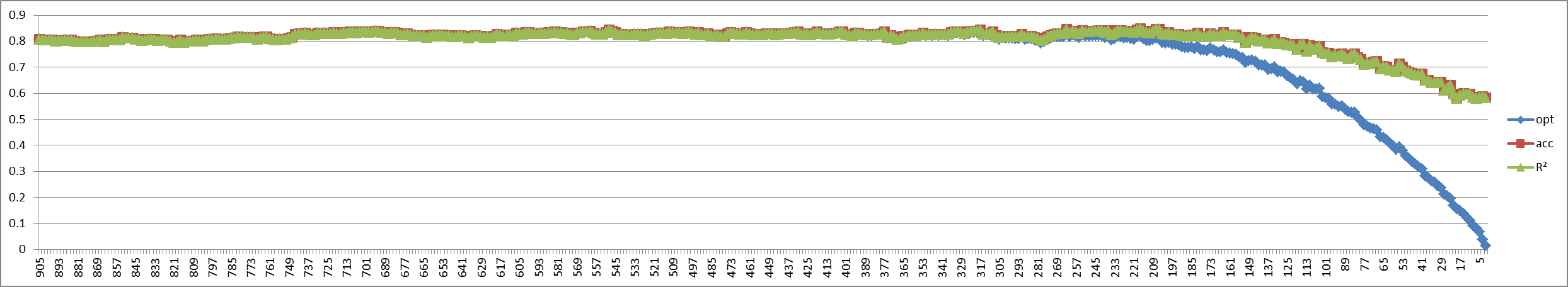
结束

**2.3.3测试、分析、结果**

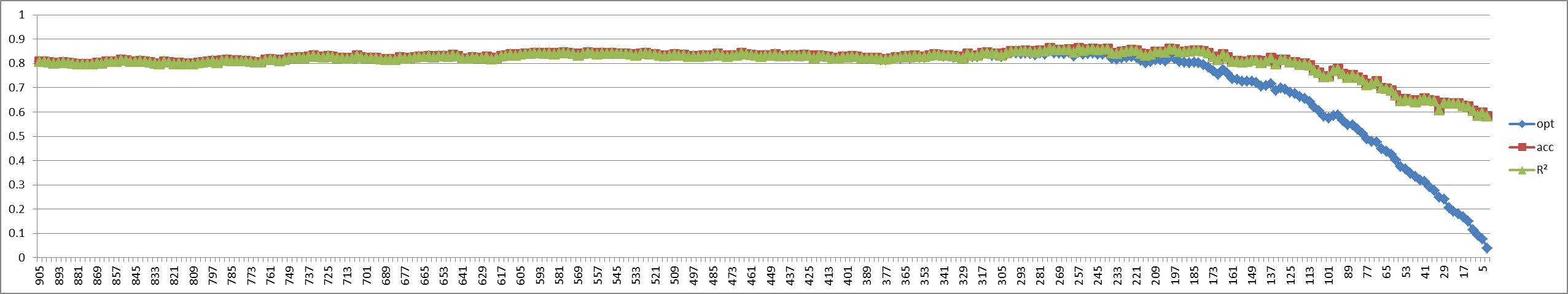
当K=1时，表示每次删除一个特征，初始有905个特征，第一次删除后剩余904个特征，循环删除的时候opt指标一直在上下波动，去直到删除到第673次剩余232个特征时遇到拐点，此处opt=0.8,acc=0.82, =0.82.拐点的定义：此点的opt大于0.8.余下的点opt均小于0.8；即P (k)>0.8且P(n>k)<0.8,（注：拐点定义是自己定义的）即删除673次时opt大于0.8,拐点之后的点opt均小于0.8。即如下图所示



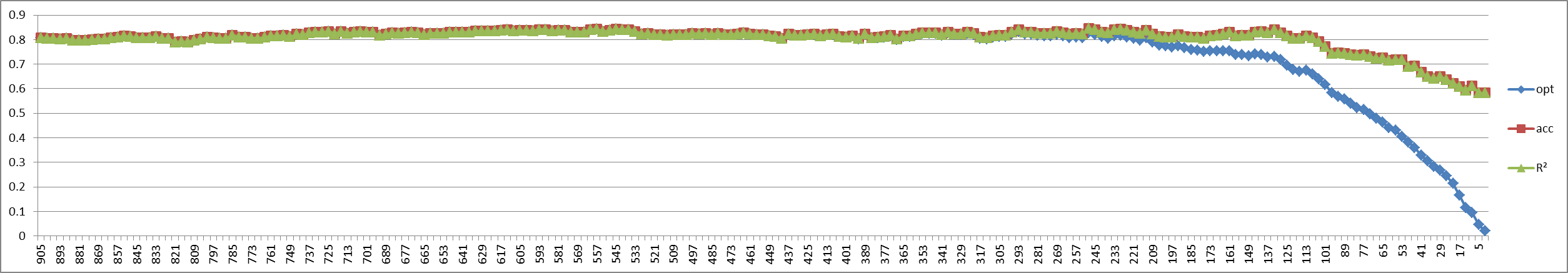
当K=2时，表示每次删除2个特征，初始有905个特征，第一次删除后剩余903个特征，循环删除的时候opt指标一直在上下波动，直到删除到第350次剩余205个特征时遇到拐点，此处opt=0.81,acc=0.85, =0.85拐点的定义：此点的opt大于0.8.余下的点opt均小于0.8；即P (k)>0.8且P(n>k)<0.8,（注：拐点定义是自己定义的）即删除350次时opt大于0.8,拐点之后的点opt均小于0.8。如下图所示



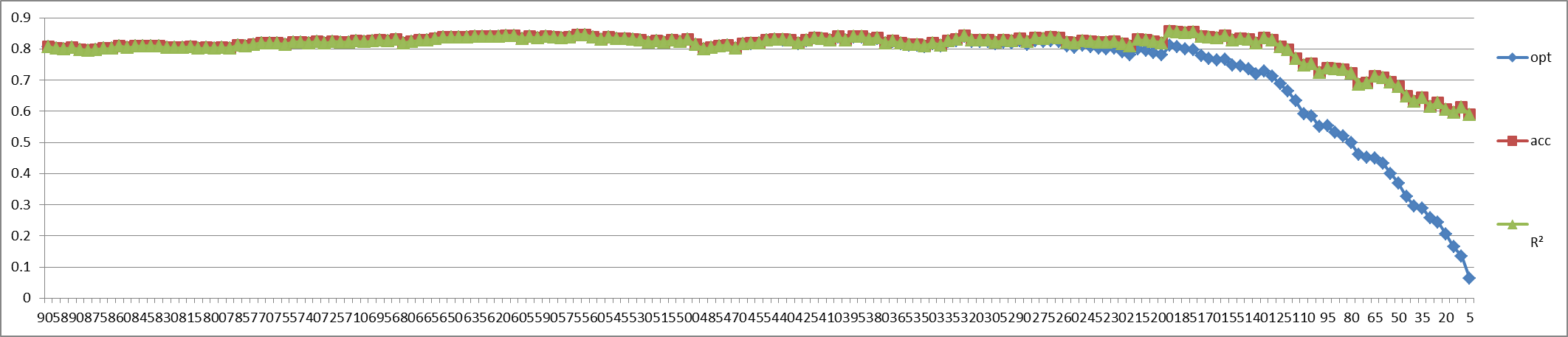
当K=3时，表示每次删除3个特征，初始有905个特征，第一次删除后剩余902个特征，循环删除的时候opt指标一直在上下波动，直到删除到第241次剩余182个特征时遇到拐点，此处opt=0.80,acc=0.85, =0.85拐点的定义：此点的opt大于0.8.余下的点opt均小于0.8；即P (k)>0.8且P(n>k)<0.8,（注：拐点定义是自己定义的）即删除241次时opt大于0.8,拐点之后的点opt均小于0.8。如下图所示



当K=4时，表示每次删除4个特征，初始有905个特征，第一次删除后剩余901个特征，循环删除的时候opt指标一直在上下波动，直到删除到第172次剩余221个特征时遇到拐点，此处opt=0.80,acc=0.83, =0.83拐点的定义：此点的opt大于0.8.余下的点opt均小于0.8；即P (k)>0.8且P(n>k)<0.8,（注：拐点定义是自己定义的）即删除172次时opt大于0.8,拐点之后的点opt均小于0.8。如下图所示



当K=5时，表示每次删除5个特征，初始有905个特征，第一次删除后剩余900个特征，循环删除的时候opt指标一直在上下波动，直到删除到第133次剩余240个特征时遇到拐点，此处opt=0.80,acc=0.82, =0.82。拐点的定义：此点的opt大于0.8.余下的点opt均小于0.8；即P (k)>0.8且P(n>k)<0.8,（注：拐点定义是自己定义的）即删除133次时opt大于0.8,拐点之后的点opt均小于0.8。如下图所示



## 3 Results

综上所述，每次循环删除2个特征，直到删除到第350次剩余205个特征时遇到拐点，此处opt=0.81,acc=0.85, =0.85,为最优.

k = 1, opt(first 100) = 0.8021379360446701

k = 2, opt(first 100) = 0.8106481671668401

k = 3, opt(first 100) = 0.81466728844522

k = 4, opt(first 100) = 0.8207967367833702

k = 5, opt(first 100) = 0.8212859454156

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| serial number | comparison of stage | sample numbers | opt |  | acc |
| 1 | [i]vs[ii] | 276 | 0.955 | 0.980 | 0.974 |
| 2 | [i]vs[iii] | 180 | 0.997 | 0.997 | 1.0 |
| 3 | [ii]vs[iii] | 302 | 0.953 | 0.979 | 0.974 |
| 4 | [i,ii]vs[iii] | 346 | 0.922 | 0.966 | 0.955 |
| 5 | [i,iii]vs[ii] | 351 | 0.948 | 0.77 | 0.971 |
| 6 | [ii.iii]vs[i] | 303 | 0.929 | 0.950 | 0.978 |
| 7 | [i]vs[ii]vs[iii] | 905 | 0.808 | 1.0 | 0.808 |

1. K取1,2,3,4,5；要前100个平均值

### Acknowledgement

### References