# 生物统计课程论文

李昌昊

2021304120016

1. **数据集介绍**

本实验的数据集包括一个分类数据集和一个回归数据集。分类数据集是由ChIP-seq（ChIP-seq，指的是结合位点分析法，作为研究体内蛋白质与DNA相互作用。染色质免疫共沉淀技术（Chromatin Immunoprecipitation，ChIP）也称结合位点分析法，是研究体内蛋白质与DNA相互作用的有力工具，通常用于转录因子结合位点或组蛋白特异性修饰位点的研究）得到一共有13212个样本，其中正样本6606个（pos.fa）,负样本6606个（neg.fa），每个样本包含了对应的DNA序列。

回归数据集包含两个文件，一个是x-2k\_sequence.fa，其中包含了序列和对应的名称，序列长度为2000，一共包含55734个样本。y.csv中包含的序列名称和对应TPM的分数。

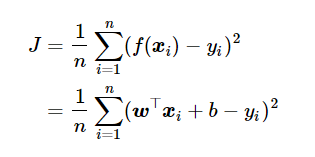
1. **实验目的**

对于分类数据集，使用k-mer对样本DNA序列进行特征提取，用Z\_score作为真实值，分别用支持向量机（SVM）、K近邻算法（KNN）、Logistic回归（LR）和随机森林（RF）这机器学习算法进行建模。并比较不同模型的性能。

对于回归数据集，使用k-mer对样本DNA序列进行特征提取，用正负标签作为真实值，用LASSO、Ridge回归、支持向量机回归（SVR）和随机森林（RF）这四种算法进行建模。并比较不同模型的性能。

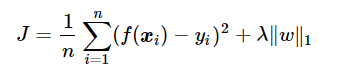
1. **机器学习算法简介**
   1. **LASSO和岭回归**

线性回归很简单，用线性函数拟合数据，用 mean square error (MSE) 计算损失（cost），然后用梯度下降法找到一组使MSE最小的权重。下图为cost function(MSE)：

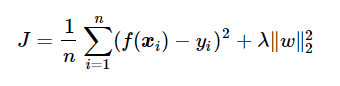


lasso 回归和岭回归（ridge regression）其实就是在标准线性回归的基础上分别加入 L1 和 L2 正则化，其中正则化的主要作用是防止过拟合，对模型添加正则化项可以限制模型的复杂度，使得模型在复杂度和性能达到平衡。L1正则化是指权值向量中各个元素的绝对值之和，例如|w1| + |w2|。L2正则化是指权值向量中各个元素的平方和然后再求平方根。

Lasso加入 L1 正则化的公式为：



岭回归加入 L2 正则化的公式为：



* 1. **支持向量机**

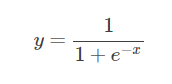
支持向量机（support vector Machine,SVM）是由Cortes和Vapnik于1995年首先提出的，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。它是建立在统计学习理论的VC维理论和结构风险最小原理基础上的，根据有限的样本信息在模式的复杂性（即对特定训练样本的学习精度，accurary）和学习能力（即无错误地识别任意样本的能力）之间寻求最佳折衷，以期获得最好的推广能力。SVM实质上是一个类分类器，是一个能够将不同类样本在样本空间分隔的超平面。换句话说，给定一些标记（label）号的训练样本，SVM算法输出一个最优化的分隔超平面。

* 1. **随机森林**

随机森林(random forest,RF)就是通过集成学习的思想将多棵树集成的一种算法，它的基本单元是决策树，而它的本质属于机器学习的一大分支——集成学习（Ensemble Learning）方法。随机森林的出现是为了解决决策树泛化能力比较弱的特点，因为决策树就有一棵树，它的决策流只有一条， 泛化能力弱。而随机森林就比较好解决了这个问题。随机森林主要有训练可以高度并行化，对于大数据时代的大样本训练速度有优势、由于可以随机选择决策树节点划分特征，这样在样本特征维度很高的时候，仍然能高效的训练模型、在训练后，可以给出各个特征对于输出的重要性、采用了随机采样，训练出的模型的方差小，泛化能力强、相对于Boosting系列的Adaboost和GBDT， RF实现比较简单、对部分特征缺失不敏感、在某些噪音比较大的样本集上，RF模型容易陷入过拟合、取值划分比较多的特征容易对RF的决策产生更大的影响，从而影响拟合的模型的效果等特点。

* 1. **Logistic回归**

Logistic回归(Logistic Regression,LR)是一种广义线性模型，用于分类问题。在线性回归模型的基础上，使用Sigmoid函数，将线性模型的结果压缩到[0,1]之间，使其拥有概率意义，它可以将任意输入映射到[0,1]区间，实现值到概率转换。下图为Sigmoid函数：



Logistic回归模型的适用条件：

1、因变量为二分类的分类变量或某事件的发生率，并且是数值型变量。

2. 残差和因变量都要服从二项分布。二项分布对应的是分类变量，所以不是正态分布，进而不是用最小二乘法，而是最大似然法来解决方程估计和检验问题。

3. 自变量和Logistic概率是线性关系

4. 各观测对象间相互独立

**3.5 K邻近算法**

K最近邻(k-Nearest Neighbor，KNN)分类算法，是一个理论上比较成熟的方法，也是最简单的机器学习算法之一。该方法的思路是：在特征空间中，如果一个样本附近的K个最近(即特征空间中最邻近)样本的大多数属于某一个类别，则该样本也属于这个类别。KNN算法不仅可以用于分类，还可以用于回归。通过找出一个样本的K个最近邻居，将这些邻居的属性的平均值赋给该样本，就可以得到该样本的属性。更有用的方法是将不同距离的邻居对该样本产生的影响给予不同的权值(weight)，如权值与距离成反比。　该算法在分类时有个主要的不足是，当样本不平衡时，如一个类的样本容量很大，而其他类样本容量很小时，有可能导致当输入一个新样本时，该样本的K个邻居中大容量类的样本占多数。 该算法只计算“最近的”邻居样本，某一类的样本数量很大，那么或者这类样本并不接近目标样本，或者这类样本很靠近目标样本。无论怎样，数量并不能影响运行结果。可以采用权值的方法（和该样本距离小的邻居权值大）来改进。

1. **回归数据集实验步骤**
   1. **提取特征**

将x-2k\_sequence.fa文件中的DNA序列使用kmer算法进行特征提取(k=5,当K=6的时候运行时间过长)，作为输入x。y.csv对应的标签取log（）作为真实值。再将提取的特征值保存到lch.npy中（数据量较大，python中npy文件读取速度较快），按照9:1划分训练集和测试集，使用训练集进行十折交叉验证寻找最优参数。得出最优参数后，再用训练集进行训练，测试集进行测试，其中测试集始终不参加模型寻参和模型训练。

* 1. **用不同的机器学习方法训练数据**

分别用LASSO、Ridge回归、支持向量机回归（SVR）和随机森林（RF）这四种算法进行训练，使用十折交叉验证挑选最优参数，使用皮尔逊相关系数（PCC）和R2作为评价指标评估不同的模型效果。

LASSO寻参，在（0.00001，1）之间，步长0.05中寻找参数，得到在在alpha=0.00001时模型取得最好值；Ridge在[0.01,0.001,0.0001,0.00001]中寻参得到在alpha = 0.1时模型取得最好值；随机森林寻参{'max\_features':range(275,365,10) "n\_estimators": [100,150,200,350] 'max\_depth':range(20,50,3) 'min\_samples\_split':range(2,41,3)}，当n\_estimators=200,max\_features=315,max\_depth=29,min\_samples\_split=11时模型取得最好的结果。其中n\_estimators为决策树的数量，max\_features为最大特征数，max\_depth为最大深度，min\_samples\_split为节点可分的最小样本数。SVM寻参"kernel": ("linear", 'rbf'), "C": np.logspace(-3, 3, 7), "gamma": np.logspace(-3, 3, 7)，当C=10.0,gamma=1000.0,kernel='rbf'时模型取得的结果最好

* 1. **回归数据集结果分析**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | LASSO | Ridge | SVR | RF |
| PCC | 0.6278 | 0.6331 | **0.6751** | 0.6248 |
| R2 | 0.3921 | 0.3922 | **0.4539** | 0.3566 |

由结果可以看出，LASSO和Ridge的性能差不多，SVR在PCC和R2中的性能最好，得出SVR的性能最好。

1. **分类数据集**
   1. **提取特征**

将neg.fa、pos.fa两个文件中的DNA序列使用kmer算法进行特征提取(k=5)，作为输入x。对应的标签作为真实值。再将正样本和负样本的训练集合并，按照9:1划分训练集和测试集，使用训练集进行十折交叉验证寻找最优参数。得出最优参数后，再用训练集进行训练，测试集进行测试，其中测试集始终不参加模型寻参和模型训练。

* 1. **用不同的机器学算法进行寻参**

KNN寻参，在（1,30）之间寻找参数，当K=15的时候取得模型的ACC性能最好；SVM寻参在{kernel:[‘rbf’、‘linear’、‘poly’],gamma:[ 1e-2,1e-3, 1e-4],C:[1、10、100、1000]}中寻找最优参数，当C=1000,gamma=0.01,kernel=’linear’的时候模型的ACC性能最好；Logistic回归寻参，在{penalty:[ ‘L1’,‘L2’]，C:[ 1e-03,1e-02,1e-01,1e+00,1e01,1e+02,1e+03]}中寻找参数，当C=1000.0,penalty='l2'时模型的ACC性能最好。因为每个参数对随机森林模型的影响成都不一样，所以我们按照n\_estimators（决策树的个数）>max\_features（最大特征数）>max\_depth（决策树最大深度）>min\_samples\_split（节点可分的最小样本数）这个顺序寻参，得到在n\_estimators=560，max\_features=65，min\_samples\_split=30，max\_depth=30时模型的AUC取值最高。

* 1. **分类数据集结果分析**

十折寻参后，测试集的结果如下：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | AUC | ACC | Precision | Recall | F1 |
| SVM | 0.9703 | 0.9054 | 0.8887 | 0.9253 | 0.9066 |
| RF | 0.9671 | 0.9077 | 0.8892 | **0.9298** | **0.9090** |
| KNN | **0.9729** | **0.9084** | **0.9239** | 0.8887 | 0.9059 |
| LR | 0.9711 | 0.9077 | 0.8903 | 0.9283 | 0.9089 |

由上表可以看出，KNN算法在AUC, ACC, Precision三个指标上都有最高值，RF在Recall,F1这两个指标上面最好。综合分析最终应选取KNN算法来建立最优模型

1. **讨论**

在回归数据集中，SVM的PCC和R2都是最好的,所以回归任务中最好的方法时SVM,LASSO和Ridge的性能差不多。在分类数据集中KNN时最好的，它的精度时最好的达到了0.9239，总体来看，分类任务中这几个方法差不多，KNN和RF相对好一些

1. **附录**

代码见：[yyzhou-gh/biostatistics: course assignment of biostatistics (github.com)](https://github.com/yyzhou-gh/biostatistics)