**华中科技大学**

**《生物信息资源与实践》实验报告**

|  |  |
| --- | --- |
| 实验名称： | 多基因donor位点预测分析 |
| 院（系）： | 电子信息与通信学院 |
| 专业班级： | 提高1501 |
| 姓名： | 周克坚 |
| 学号： | U201513485 |
| 时间： | 2018.12.11 |
| 地点： | 华中科技大学南一楼西211 |
| 实验成绩： |  |
| 指导教师： | 彭凯 |

2018 年 12 月 11 日

# 实验要求

训练集和测试集为多个基因donor剪接位点信息，其中括号中为donor剪接位点的序列位置，熟悉两个文件并学习使用两种算法进行预测分析。

1.熟悉贝叶斯模型，使用WAM算法对donor位点进行预测。

a) 选用合适的阈值，进行预测，计算sn(sensitivity)和sp(specificity)值

b) 选用多个阈值进行计算，绘制sn和sp曲线，查看最好的阈值取值

2. 使用SVM方法对训练集和测试机进行预测，计算出sn和sp值

# 二、实验原理

1.利用贝叶斯模型进行donor位点预测：

贝叶斯决策基本思想：

* 先计算先验概率和联合概率
* 利用贝叶斯公式计算条件概率
* 根据条件概率的分布选择合适的阈值进行决策分类

实验中使用WAM模型对9bp序列基因片段进行判断，对于任一DNA片段X = x1x2x3 ... x-1x，它作为位点信号的WAM判别函数为：

2.利用SVM对donor位点进行预测：

已知：n个样本，(X1, y1), (X2, y2)…(Xn, yn)

给定核函数



求解



结果：



本实验中选取的核为高斯核



对样本进行训练后，对测试集的正例和负例进行预测，对于得到的结果进行评估。

# 三、实现过程

## 1、数据预处理

对所给的训练集和测试集中的序列数据进行预处理，将donor位点进行提取，并保存在一个单独的文本文件中。对于非donor位点，采用背景序列提取，对于每个非donor片段随机获得同样为9bp的基因序列。所得的非donor位点序列也保存在一个单独的文本文件中。

在进行SVM训练前，需要将”A,T,C,G”序列转换为向量。本实验中以{'a’:0, ‘c’:1, ‘g’:2, ‘t’:3}的映射方法进行向量转换。并将donor位点序列标记为1，非donor位点标记为0。并将处理后的数据单独保存为svm训练与测试数据文本

## 2、贝叶斯分类器

根据WAM模型的判别函数：

首先对训练集数据进行数学统计，得到各项概率。

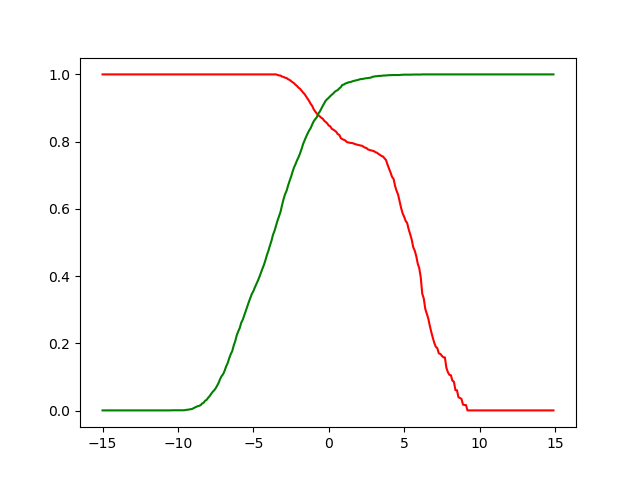
再对测试集各数据进行判别函数的计算，取不同阈值对测试集数据进行预测。并对得到的结果计算sn(sensitivity)和sp(specificity)值，并绘制曲线图。

## 3、SVM分类器：

调用python scikit-learn包中的SVM函数，对训练集进行训练，所取的核函数为高斯，各项参数见附页二中的代码，对训练集进行训练后得到SVM分类器，并对测试集进行预测。从而得到sn(sensitivity)和sp(specificity)。

# 四、实验结果

## 贝叶斯分类器：



红色是sn ,绿色sp,横轴是阈值

## SVM分类器

SN: 0.973555

SP: 0.909709

## 结果分析:

可见贝叶斯分类器根据实际需求调整threshold，能够以不同sn和sp值的要求对数据进行预测，综合两者考虑的话阈值在0附近的分类效果最好。SVM分类器训练结果准确率较高，但SP值偏低。