数据挖掘实验报告

数据的分类预测

姓　　名：汤凯(24320142202489)

王崇菲(24320142202492)

陈东东(24320142202402)

指导老师：刘昆宏

实验地点：海韵实验楼

完成时间：2017.10.17

一. 实验目的

1、使用 scikit-learn 包中的 tree 与贝叶斯，对数据进行模型训练，尽量了解其原理及运用。

2、分析 tree 和贝叶斯在实验中的性能比较，并基于此比较两者的特点。

二. 实验内容

分别使用 scikit-learn 文本分类器和朴素贝叶斯分类器做预测，训练数据模型，输出准确度。

三. 实验步骤以及结果

1、伯努利朴素贝叶斯

实验步骤：

（1） 按行读取文件，并将读取的第一行数据作为 Y，其余数据转置作为训练数据 X;

（2） 创建 BernoulliNB()类的实例

（3） 调用 fit()方法进行训练数据

（4） 调用 predict()方法进行预测

（5） 计算并输出正确率

实验结果：

Leukemia1 训练及测试结果如下：

真实值、预测值、目前正确个数

('B_cell', 'B_cell', 1)
('B_cell', 'B_cell', 2)
('B_cell', 'B_cell', 3)
('B_cell', 'B_cell', 4)
('B_cell', 'B_cell', 5)
('B_cell', 'B_cell', 6)
('B_cell', 'B_cell', 7)
('B_cell', 'B_cell', 8)
('B_cell', 'B_cell', 9)
('B_cell', 'B_cell', 10)
('B_cell', 'B_cell', 11)
('B_cell', 'B_cell', 12)
('B_cell', 'B_cell', 13)
('B_cell', 'B_cell', 14)
('B_cell', 'B_cell', 15)
('B_cell', 'B_cell', 16)

('B_cell', 'B_cell', 17)
('B_cell', 'B_cell', 18)
('B_cell', 'B_cell', 19)
('T_cell', 'B_cell', 19)
('AML', 'AML', 20)
('AML', 'AML', 21)
('AML', 'AML', 22)
('AML', 'AML', 23)
('AML', 'B_cell', 23)
('AML', 'B_cell', 23)
('AML', 'AML', 24)
('AML', 'B_cell', 24)
('AML', 'B_cell', 24)
('AML', 'B_cell', 24)
('AML', 'B_cell', 24)
('AML', 'B_cell', 24)
('AML', 'B_cell', 24)
('AML', 'B_cell', 24)
总数：34 正确数：24 正确率：0.705882

Leukemia2 训练及测试结果如下：
真实值、预测值、目前正确个数
('ALL', 'ALL', 1)
('ALL', 'ALL', 2)
('ALL', 'ALL', 3)
('ALL', 'ALL', 4)
('MLL', 'MLL', 5)
('MLL', 'MLL', 6)
('MLL', 'MLL', 7)
('AML', 'AML', 8)
('AML', 'AML', 9)
('AML', 'AML', 10)
('AML', 'AML', 11)
('AML', 'AML', 12)
('AML', 'AML', 13)
('AML', 'AML', 14)
('AML', 'AML', 15)
总数：15 正确数：15 正确率：1.000000

Breast 训练及测试结果如下：
真实值、预测值、目前正确个数
('lumina', 'lumina', 1)
('lumina', 'lumina', 2)
('lumina', 'lumina', 3)

('lumina', 'lumina', 4)
('lumina', 'lumina', 5)
('lumina', 'lumina', 6)
('lumina', 'lumina', 7)
('lumina', 'lumina', 8)
('lumina', 'lumina', 9)
('lumina', 'lumina', 10)
('lumina', 'lumina', 11)
('lumina', 'lumina', 12)
('ERBB2', 'lumina', 12)
('ERBB2', 'lumina', 12)
('ERBB2', 'lumina', 12)
('basal', 'basal', 13)
('basal', 'basal', 14)
('basal', 'ERBB2', 14)
('normal', 'normal', 15)
('normal', 'lumina', 15)
('normal', 'lumina', 15)
('normal', 'lumina', 15)
('normal', 'normal', 16)
('cell_lines', 'cell_lines', 17)
('cell_lines', 'cell_lines', 18)
('cell_lines', 'cell_lines', 19)
('cell_lines', 'cell_lines', 20)
('cell_lines', 'cell_lines', 21)
('cell_lines', 'cell_lines', 22)
('cell_lines', 'cell_lines', 23)
总数：30　正确数：23　正确率：0.766667

GCM 训练及测试结果如下：
真实值、预测值、目前正确个数
('Breast', 'Bladder', 0)
('Breast', 'Bladder', 0)
('Breast', 'Bladder', 0)
('Prostate', 'Uterus', 0)
('Prostate', 'Mesothelioma', 0)
('Lung', 'Mesothelioma', 0)
('Lung', 'Lymphoma', 0)
('Lung', 'Mesothelioma', 0)
('Colorectal', 'Colorectal', 1)
('Colorectal', 'Mesothelioma', 1)
('Colorectal', 'Colorectal', 2)
('Lymphoma', 'Lymphoma', 3)
('Lymphoma', 'Lymphoma', 4)

('Lymphoma', 'Lymphoma', 5)
('Lymphoma', 'Lymphoma', 6)
('Lymphoma', 'Lymphoma', 7)
('Lymphoma', 'Lung', 7)
('Bladder', 'Renal', 7)
('Bladder', 'Ovary', 7)
('Bladder', 'Melanoma', 7)
('Melanoma', 'Bladder', 7)
('Melanoma', 'Melanoma', 8)
('Uterus', 'Uterus', 9)
('Uterus', 'Colorectal', 9)
('Leukemia', 'Leukemia', 10)
('Leukemia', 'Leukemia', 11)
('Leukemia', 'Leukemia', 12)
('Leukemia', 'Lymphoma', 12)
('Leukemia', 'Leukemia', 13)
('Leukemia', 'Leukemia', 14)
('Renal', 'Uterus', 14)
('Renal', 'Bladder', 14)
('Renal', 'Uterus', 14)
('Pancreas', 'Uterus', 14)
('Pancreas', 'Bladder', 14)
('Pancreas', 'Mesothelioma', 14)
('Ovary', 'Bladder', 14)
('Ovary', 'Ovary', 15)
('Ovary', 'Mesothelioma', 15)
('Mesothelioma', 'Mesothelioma', 16)
('Mesothelioma', 'Mesothelioma', 17)
('Mesothelioma', 'Mesothelioma', 18)
('CNS', 'CNS', 19)
('CNS', 'CNS', 20)
('CNS', 'CNS', 21)
('CNS', 'CNS', 22)
总数：46　正确数：22　正确率：0.478261

2、决策树

实验步骤：

（1） 按行读取文件按行读取文件，并将读取的第一行数据作为 Y，其余数据转置作为训练数据 X;
（2） 创建 tree.DecisionTreeClassifier()分类器
（3） 调用 fit()方法进行训练数据
（4） 调用 predict()方法进行预测

（5） 计算并输出正确率

实验结果：

Leukemia1 训练及测试结果如下：
真实值,预测值,目前正确个数
('B_cell', 'B_cell', 1)
('B_cell', 'B_cell', 2)
('B_cell', 'B_cell', 3)
('B_cell', 'B_cell', 4)
('B_cell', 'B_cell', 5)
('B_cell', 'B_cell', 6)
('B_cell', 'B_cell', 7)
('B_cell', 'B_cell', 8)
('B_cell', 'B_cell', 9)
('B_cell', 'B_cell', 10)
('B_cell', 'B_cell', 11)
('B_cell', 'B_cell', 12)
('B_cell', 'B_cell', 13)
('B_cell', 'B_cell', 14)
('B_cell', 'B_cell', 15)
('B_cell', 'B_cell', 16)
('B_cell', 'T_cell', 16)
('B_cell', 'B_cell', 17)
('B_cell', 'B_cell', 18)
('T_cell', 'AML', 18)
('AML', 'AML', 19)
('AML', 'AML', 20)
('AML', 'AML', 21)
('AML', 'AML', 22)
('AML', 'AML', 23)
('AML', 'AML', 24)
('AML', 'AML', 25)
('AML', 'AML', 26)
('AML', 'AML', 27)
('AML', 'AML', 28)
('AML', 'T_cell', 28)
('AML', 'AML', 29)
('AML', 'AML', 30)
('AML', 'B_cell', 30)
总数：34 正确数：30 正确率：0.882353

Leukemia2 训练及测试结果如下：

真实值,预测值,目前正确个数
('ALL', 'ALL', 1)
('ALL', 'ALL', 2)
('ALL', 'ALL', 3)
('ALL', 'ALL', 4)
('MLL', 'MLL', 5)
('MLL', 'MLL', 6)
('MLL', 'MLL', 7)
('AML', 'AML', 8)
('AML', 'AML', 9)
('AML', 'AML', 10)
('AML', 'AML', 11)
('AML', 'MLL', 11)
('AML', 'AML', 12)
('AML', 'AML', 13)
('AML', 'AML', 14)
总数：15　正确数：14　正确率：0.933333

Breast 训练及测试结果如下：
真实值,预测值,目前正确个数
('lumina', 'lumina', 1)
('lumina', 'lumina', 2)
('lumina', 'normal', 2)
('lumina', 'ERBB2', 2)
('lumina', 'lumina', 3)
('lumina', 'lumina', 4)
('lumina', 'lumina', 5)
('lumina', 'lumina', 6)
('lumina', 'lumina', 7)
('lumina', 'lumina', 8)
('lumina', 'lumina', 9)
('lumina', 'normal', 9)
('ERBB2', 'cell_lines', 9)
('ERBB2', 'basal', 9)
('ERBB2', 'lumina', 9)
('basal', 'basal', 10)
('basal', 'basal', 11)
('basal', 'ERBB2', 11)
('normal', 'basal', 11)
('normal', 'normal', 12)
('normal', 'lumina', 12)
('normal', 'ERBB2', 12)
('normal', 'normal', 13)
('cell_lines', 'cell_lines', 14)

('cell_lines', 'lumina', 14)
('cell_lines', 'cell_lines', 15)
('cell_lines', 'cell_lines', 16)
('cell_lines', 'cell_lines', 17)
('cell_lines', 'cell_lines', 18)
('cell_lines', 'cell_lines', 19)
总数：30　正确数：19　正确率：0.633333

GCM 训练及测试结果如下：
真实值,预测值,目前正确个数
('Breast', 'Breast', 1)
('Breast', 'Leukemia', 1)
('Breast', 'Pancreas', 1)
('Prostate', 'Uterus', 1)
('Prostate', 'Bladder', 1)
('Lung', 'Lung', 2)
('Lung', 'Mesothelioma', 2)
('Lung', 'Lung', 3)
('Colorectal', 'Colorectal', 4)
('Colorectal', 'Colorectal', 5)
('Colorectal', 'Colorectal', 6)
('Lymphoma', 'Lymphoma', 7)
('Lymphoma', 'Renal', 7)
('Lymphoma', 'Lymphoma', 8)
('Lymphoma', 'Lymphoma', 9)
('Lymphoma', 'Lymphoma', 10)
('Lymphoma', 'Lymphoma', 11)
('Bladder', 'Ovary', 11)
('Bladder', 'Bladder', 12)
('Bladder', 'Bladder', 13)
('Melanoma', 'Melanoma', 14)
('Melanoma', 'Melanoma', 15)
('Uterus', 'Pancreas', 15)
('Uterus', 'Pancreas', 15)
('Leukemia', 'Leukemia', 16)
('Leukemia', 'Leukemia', 17)
('Leukemia', 'Prostate', 17)
('Leukemia', 'Prostate', 17)
('Leukemia', 'Leukemia', 18)
('Leukemia', 'Leukemia', 19)
('Renal', 'Ovary', 19)
('Renal', 'Pancreas', 19)
('Renal', 'CNS', 19)
('Pancreas', 'Colorectal', 19)

('Pancreas', 'Breast', 19)
('Pancreas', 'Colorectal', 19)
('Ovary', 'Bladder', 19)
('Ovary', 'Bladder', 19)
('Ovary', 'Bladder', 19)
('Mesothelioma', 'Mesothelioma', 20)
('Mesothelioma', 'Renal', 20)
('Mesothelioma', 'Melanoma', 20)
('CNS', 'CNS', 21)
('CNS', 'CNS', 22)
('CNS', 'Renal', 22)
('CNS', 'CNS', 23)
总数：46　正确数：23　正确率：0.500000

四．总结

本次实验结果的分析，我们主要想从三个方面来进行分析，首先是决策树、贝叶斯算法本身，然后就是他们两个算法的比较。

1、决策树：

从实验结果来看，我们的决策树训练出来的模型对不同的数据他们的效果还是有所不同的，尤其是在类别很多的时候，而且每次的运行，对于同一组数据训练出来的模型预测的效率也是不一样的，我们以对 Leukemia1 数据的预测为例：

```
Leukemia1训练及测试结果如下：
真实值,预测值,目前正确个数
B_cell B_cell 1
B_cell B_cell 2
B_cell B_cell 3
B_cell B_cell 4
B_cell B_cell 5
B_cell B_cell 6
B_cell B_cell 7
B_cell B_cell 8
B_cell B_cell 9
B_cell B_cell 10
B_cell B_cell 11
B_cell B_cell 12
B_cell B_cell 13
B_cell B_cell 14
B_cell B_cell 15
B_cell B_cell 16
B_cell AML 16
B_cell B_cell 17
B_cell B_cell 18
T_cell AML 18
AML AML 19
AML AML 20
AML AML 21
AML AML 22
AML T_cell 22
AML AML 23
AML AML 24
AML AML 25
AML AML 26
AML AML 27
AML AML 28
AML T_cell 28
AML T_cell 28
AML B_cell 28
总数：34  正确数：28  正确率：0.823529
```

```
Leukemia1训练及测试结果如下：
真实值,预测值,目前正确个数
B_cell B_cell 1
B_cell B_cell 2
B_cell AML 2
B_cell T_cell 2
B_cell B_cell 3
B_cell B_cell 4
B_cell B_cell 5
B_cell B_cell 6
B_cell B_cell 7
B_cell B_cell 8
B_cell B_cell 9
B_cell B_cell 10
B_cell B_cell 11
B_cell B_cell 12
B_cell T_cell 12
B_cell B_cell 13
B_cell B_cell 14
B_cell B_cell 15
B_cell B_cell 16
T_cell AML 16
AML AML 17
AML AML 18
AML AML 19
AML AML 20
AML T_cell 20
AML AML 21
AML AML 22
AML AML 23
AML AML 24
AML AML 25
AML AML 26
AML B_cell 26
AML B_cell 26
AML AML 27
总数：34  正确数：27  正确率：0.794118
```

由此我们可以看出决策树还是存在一定的缺点的（以下只列举部分）：

（1）决策树的结果可能是不稳定的，因为在数据中一个很小的变化可能导致生成一个完全不同的树。

（2）当类别太多时，错误可能就会增加的比较快。

当然决策树还是有自身的优点的（以下只列举部分）：

（1）对于决策树，数据的准备往往是简单或者是不必要的 . 其他的技术往往要求先把数据一般化，比如去掉多余的或者空白的属性。

（2）在相对短的时间内能够对大型数据源做出可行且效果良好的结果。

（3）效率高，决策树只需要一次构建，反复使用，每一次预测的最大计算次数不超过决策树的深度。

2、贝叶斯

在实验中，我们发现，相对于决策树出现的每次运行预测效率不一样的情况是不存在的，伯努利朴素贝叶斯每次的运行结果都是一致的。

在此我们也总结了一部分关于贝叶斯算法的优缺点：

优点：

（1）朴素贝叶斯模型发源于古典数学理论，有稳定的分类效率。

（2）对小规模的数据表现很好，能个处理多分类任务，适合增量式训练，

尤其是数据量超出内存时，我们可以一批批的去增量训练。

（3）对缺失数据不太敏感，算法也比较简单，常用于文本分类。

缺点：

（1）需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可以有很多种，因此在某些时候会由于假设的先验模型的原因导致预测效果不佳。

（2）由于我们是通过先验和数据来决定后验的概率从而决定分类，所以分类决策存在一定的错误率。

（3）对输入数据的表达形式很敏感。

3、决策树与贝叶斯的比较

和决策树相比，贝叶斯模型发源于古典数学理论，有着坚实的数学基础，以及稳定的分类效率，同时贝叶斯模型所需估计参数很少，对缺失数据不太敏感，算法也比较简单。理论上，贝叶斯模型与其他分类方法相比较具有最小的误差率。