# 实验报告二

学号：20191060239

姓名：李品鑫

实验名称：分类算法的实现与应用

【实验目标】：

实现贝叶斯分类算法、决策树分类算法，并在UCI数据集“Breast Cancer”上进行应用。数据集见文件夹breast-cancer。

【实验任务】：

1. 阐述贝叶斯分类算法的原理，编程实现贝叶斯分类算法。

2. 阐述决策树分类算法的原理，编程实现决策树分类算法。

3. 原始数据描述：对Breast Cancer数据集的属性信息进行描述。

4. 数据集划分：使用合适的方式将数据集划分为训练集、测试集两部分。

5. 将Breast Cancer数据集中的Class作为分类标号，其他属性作为描述属 性，在训练集上对已实现的算法进行训练，对测试集中的数据进行分类任 务（假定测试集中数据对象的Class分类标号未知，使用分类算法预测数据 对象的Class分类标号），并使用至少一种评价指标对分类算法的性能进行 评估。

【实验要求】：

（1）独立完成实验，拒绝抄袭。撰写实验报告，需包含实验目的，实验原理及过程，实验结果展示等。

（2）对实验任务中的1和2，可根据个人能力选做其中一个或两个都做，任务1的难度低于任务2。（评分会考虑任务难度，鼓励有余力的同学两个算法都实现）。

（3）可根据自己能力借助合适的工具（如sklearn库）完成实验，但需清楚算法原理。鼓励自己动手编程实现整个算法，评分时，自己动手编程实现>借助工具。

【实验过程】：

1. （对应任务3.）原始数据描述：对Breast Cancer数据集的属性信息进行描述。

根据breast-cancer.names文件的描述，总结出以下三点对数据集的描述

（1）该数据集为南斯拉夫卢布尔雅那肿瘤研究所大学医学研究中心提 供的乳腺癌数据集。

（2）数据集共有286个样本，包含两类，201例为其中一类（“no-recurrence-events”），85例为另一类（“recurrence-events”）。

（3）数据集包含了2类（分别为“no-recurrence-events”，“recurrence-events”），9个属性（“age”，“menopause”，“tumor-size”，“inv-nodes”，“node-caps”，“deg-malig”，“breast”，“breast-quad”，“irradiat”）

（4）

1. Class: no-recurrence-events, recurrence-events

2. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.

3. menopause: lt40, ge40, premeno.

4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44,

45-49, 50-54, 55-59.

5. inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,

27-29, 30-32, 33-35, 36-39.

6. node-caps: yes, no.

7. deg-malig: 1, 2, 3.

8. breast: left, right.

9. breast-quad: left-up, left-low, right-up, right-low, central.

10. irradiat: yes, no.

2. （对应任务4.）数据集划分：使用合适的方式将数据集划分为训练集、测试集两部分。

采用留出法划分数据集，约80%为训练集、20%为测试集。

3. （对应任务1. 任务5.）完成任务：阐述贝叶斯分类算法的原理，编程实现贝叶斯分类算法。并使用至少一种评价指标对分类算法的性能进行评估。

贝叶斯算法是利用特征向量与贝叶斯公式

其中 为分类的类别， 为一个待分类项， 为特征向量的分量。因在计算中可被视为常数，故而可将公式简化为：

利用该公式依次计算待判别样本属于全部分类类别的概率值，得到其中

最大的值

概率值最大的类别即可为分类结果。

贝叶斯分类matlab代码如下：

close all;

clear all;

clc;

%读取数据

filename = 'breast-cancer\breast-cancer.txt';

fileID = fopen(filename);

data = textscan(fileID,'%s%s%s%s%s%s%s%s%s%s','Delimiter',',');

fclose(fileID);

attrib = [data{1,1}, data{1,2}, data{1,3}, data{1,4}, data{1,5}, data{1,6}, data{1,7}, data{1,8}, data{1,9}, data{1,10}];

X1=floor(201\*0.8);

X2=floor(85\*0.8);

c1=attrib(1:X1,:);c2=attrib(202:201+X2,:);

c3=attrib(X1+1:201,:);c4=attrib(201+X2:286,:);

tdata=cat(1,c3,c4);

n\_age=[];n\_menopause=[];n\_tumor\_size=[];

n\_inv\_nodes=[];n\_node\_caps=[];n\_deg\_malig=[];

n\_breast=[];n\_breast\_quad=[];n\_irradiat=[];

n1\_age=[];n1\_menopause=[];n1\_tumor\_size=[];

n1\_inv\_nodes=[];n1\_node\_caps=[];n1\_deg\_malig=[];

n1\_breast=[];n1\_breast\_quad=[];n1\_irradiat=[];

%计算各概率

age = {"10-19","20-29","30-39","40-49","50-59","60-69","70-79","80-89","90-99"};

for i =1:9

n\_age(1,i)=length(find(c1(:,2)==age{i}))/X1;

n1\_age(1,i)=length(find(c2(:,2)==age{i}))/X2;

end

menopause={"lt40","ge40","premeno"};

for i =1:3

n\_menopause(1,i)=length(find(c1(:,3)==menopause{i}))/X1;

n1\_menopause(1,i)=length(find(c2(:,3)==menopause{i}))/X2;

end

tumor\_size={"0-4","5-9","10-14","15-19","20-24","25-29","30-34","35-39","40-44","45-49","50-54","55-59"};

for i =1:12

n\_tumor\_size(1,i)=length(find(c1(:,4)==tumor\_size{i}))/X1;

n1\_tumor\_size(1,i)=length(find(c2(:,4)==tumor\_size{i}))/X2;

end

inv\_nodes={"0-2","3-5","6-8","9-11","12-14","15-17","18-20","21-23","24-26","27-29","30-32","33-35","36-39"};

for i =1:13

n\_inv\_nodes(1,i)=length(find(c1(:,5)==inv\_nodes{i}))/X1;

n1\_inv\_nodes(1,i)=length(find(c2(:,5)==inv\_nodes{i}))/X2;

end

node\_caps={"yes","no"};

for i =1:2

n\_node\_caps(1,i)=length(find(c1(:,6)==node\_caps{i}))/X1;

n1\_node\_caps(1,i)=length(find(c2(:,6)==node\_caps{i}))/X2;

end

deg\_malig={"1","2","3"};

for i =1:3

n\_deg\_malig(1,i)=length(find(c1(:,7)==deg\_malig{i}))/X1;

n1\_deg\_malig(1,i)=length(find(c2(:,7)==deg\_malig{i}))/X2;

end

breast={"left","right"};

for i =1:2

n\_breast(1,i)=length(find(c1(:,8)==breast{i}))/X1;

n1\_breast(1,i)=length(find(c2(:,8)==breast{i}))/X2;

end

breast\_quad={"left-up","left-low","right-up","right-low","central"};

for i =1:5

n\_breast\_quad(1,i)=length(find(c1(:,9)==breast\_quad{i}))/X1;

n1\_breast\_quad(1,i)=length(find(c2(:,9)==breast\_quad{i}))/X2;

end

irradiat={"yes","no"};

for i =1:2

n\_irradiat(1,i)=length(find(c1(:,10)==irradiat{i}))/X1;

n1\_irradiat(1,i)=length(find(c2(:,10)==irradiat{i}))/X2;

end

a={age,menopause,tumor\_size,inv\_nodes,node\_caps,deg\_malig,breast,breast\_quad,irradiat};

n={n\_age,n\_menopause,n\_tumor\_size,n\_inv\_nodes,n\_node\_caps,n\_deg\_malig,n\_breast,n\_breast\_quad,n\_irradiat};

n1={n1\_age,n1\_menopause,n1\_tumor\_size,n1\_inv\_nodes,n1\_node\_caps,n1\_deg\_malig,n1\_breast,n1\_breast\_quad,n1\_irradiat};

%

sum=0;

p1=X1/(X1+X2);

p2=X2/(X1+X2);

for k =i:length(tdata(:,1))

for i =1:9

for j=1:length(n{i})

if tdata{k,i+1}==a{1,i}{1,j}

p1=p1\*n{1,i}(1,j);

end

end

end

for i =1:9

for j=1:length(n1{i})

if tdata{k,i+1}==a{1,i}{1,j}

p2=p2\*n1{1,i}(1,j);

end

end

end

if p1>p2

result="no-recurrence-events"

else

result="recurrence-events"

end

if result ==tdata{k,1}

sum=sum+1;

end

end

right\_rate=sum/length(tdata(:,1));

利用准确率对该贝叶斯分类器进行评估。

对测试集的判断准确率如下：



大于0.5，故认为有效。