**CS 497 Data Mining Spring 2011**

# Test#1 (*100 points in total*)

**This is a home taken test. Test#1 is due on Wednesday, March 23, at the beginning of class, hard copy submission. The test contains 6 problems and worth 100 points.**

Given lenses dataset and its attributes description.

**Attribute Information:**

**Last attribute – Classification Attribute** -- 3 Classes   
1 : the patient should be fitted with hard contact lenses,   
2 : the patient should be fitted with soft contact lenses,   
3 : the patient should not be fitted with contact lenses.

**Other attributes, given by order and their values:**  
1. age of the patient: (1) young, (2) pre-presbyopic, (3) presbyopic   
2. spectacle prescription: (1) myope, (2) hypermetrope   
3. astigmatic: (1) no, (2) yes   
4. tear production rate: (1) reduced, (2) normal

Dataset:

1 1 1 1 1 3

2 1 1 1 2 2

3 1 1 2 1 3

4 1 1 2 2 1

5 1 2 1 1 3

6 1 2 1 2 2

7 1 2 2 1 3

8 1 2 2 2 1

9 2 1 1 1 3

10 2 1 1 2 2

11 2 1 2 1 3

12 2 1 2 2 1

13 2 2 1 1 3

14 2 2 1 2 2

15 2 2 2 1 3

16 2 2 2 2 3

17 3 1 1 1 3

18 3 1 1 2 3

19 3 1 2 1 3

20 3 1 2 2 1

**Problem 1.** **[20 points]** Use 1R method to construct classification rules for the predicted attribute **recommendation** (last attribute)**. Show all your work!**

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Rules | Errors | Total Errors |
| Age | Young → 4  Pre-presbyopic → 5  Presbyopic → 3 | 4/8  3/8  1/4 | 8/20 |
| Prescription | Myope → 7  Hypermetrope → 5 | 5/12  3/8 | 8/20 |
| Astigmatic | No → 6  Yes → 6 | 4/10  4/10 | 8/20 |
| Tear Rate | Reduced → 10  Normal → 2 | 0/10  8/10 | 8/20 |

The above table is based on the “no lenses” class. Errors reflected are for either soft or hard contact lens prescription.

**Problem 2. [20 points]** Construct Naïve Bayes classifier. **( Compute all needed probabilities )**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | | | | Prescription | | | | Astigmatic | | | | Tear Rate | | | | Lenses | | |
|  | H | S | N |  | H | S | N |  | H | S | N |  | H | S | N | H | S | N |
| Y | 2 | 3 | 3 | M | 3 | 2 | 7 | Y | 4 | 0 | 6 | R | 0 | 0 | 10 | 4 | 4 | 12 |
| PP | 1 | 2 | 5 | H | 1 | 2 | 5 | N | 0 | 4 | 6 | N | 4 | 4 | 2 |  |  |  |
| P | 1 | 0 | 3 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Y | 2/8 | 3/8 | 3/8 | M | 3/12 | 2/12 | 7/12 | Y | 4/10 | 0/10 | 6/10 | R | 0/10 | 0/10 | 10/10 | 4/20 | 4/20 | 12/20 |
| PP | 1/8 | 2/8 | 5/8 | H | 1/8 | 2/8 | 5/8 | N | 0/10 | 4/10 | 6/10 | N | 4/10 | 4/10 | 2/10 |  |  |  |
| P | 1/4 | 0/4 | 3/4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**Likelihood of '1': 1/4 x 1/8 x 0/10 x 4/10 = 0**

**Likelihood of '2': 0/4 x 2/8 x 4/10 x 4/10 = 0**

**Likelihood of '3': 3/4 x 5/8 x 6/10 x 2/10 = 0.056**

3 2 1 2 **3**

Likelihood of '1': 1/8 x 0/10 = 0

Likelihood of '2': 2/8 x 0/10 = 0

Likelihood of '3': 3/4 x 10/10 = 0.75

**?** 2 **?** 1 **3**

**Problem 3. [20 points]** Calculate the Gain(S,A1) and Gain(S,A2) for the dataset S and attributes A1 and A2**. Show all your work.** (Show each step of the calculation.)

A1 = Age

Age = Young: info([2,3,3]) = entropy(2/8,3/8,3/8) = -2/8log(2/8) – 3/8log(3/8) – 3/8log(3/8) = 0.470 bits

Age = Pre-Presbyopic: info([1,2,5]) = entropy(1/8,2/8,5/8) = -1/8log(1/8) – 2/8log(2/8) – 5/8log(5/8) = 0.391 bits

Age = Presbyopic: info([1,0,3]) = entropy(1/4,0/4,3/4) = -1/4log(1/4) – 0log(0) – 3/4log(3/4) = 0 bits

Expected Info: info([2,3,3],[1,2,4],[1,0,3]) = (8/20 x 0.470) + (8/20 x 0.391) + (4/20 x 0) = 0.344 bits

gain(Age) = info([8,8,4]) – info([2,3,3],[1,2,4],[1,0,3]) = 0.458-0.344 = 0.144 bits

A2 = Spectacle Prescription

Spectacle Prescription = Myope: info([3,2,7]) = entropy(3/12,2/12,7/12) = -3/12log(3/12) – 2/12log(2/12) – 7/12log(7/12) = 0.417 bits

Spectacle Prescription = Hypermetrope: info([1,2,5]) = entropy(1/8,2/8,5/8) = -1/8log(1/8) – 2/8log(2/8) – 5/8log(5/8) = 0.391 bits

Expected Info: info([3,2,7],[1,2,5]) = (12/20 x 0.417) + (8/20 x 0.391) = 0.254 bits

gain(Spectacle Prescription) = info([12,8]) – info([3,2,7],[1,2,5]) = 0.292 – 0.254 = 0.038 bits

**Problem 4. [10 points]** Calculate the Gain\_ratio(S,A1) and Gain\_ratio(S,A2) for the dataset S and attributes A1 and A2**. Show all your work.** (Show each step of the calculation. )If you use ID3 algorithm, which of these two attributes you will chose for a root node of a decision tree? **Explain your answer!**

**Gain Ratio(S, A1): info([8,8,4]) = 0.458 bits. Ratio = 0.144/0.458 = 0.314 bits**

**Gain Ratio(S, A1): info([12,8]) = 0.292 bits. Ratio = 0.038/0.292 = 0.130 bits**

**If I were to use ID3 I would choose Age as the root node for my decision tree because it has the highest gain ratio and therefore the greatest chance for accurate predictions.**

**Problem 5. [20 points]** Construct classification rule for class=”3” using Simple Covering Algorithm. **Show all stages/steps of rule construction and all your calculations!**

Phase 0:

Rule: If true then Recommendation = None

Training set: Entire Data Set

Phase 1:

Attribute = Attribute Value n/t

Age = Young 3/8

Age = Pre-Presbyopic 5/8

Age = Presbyopic 3/4

Spectacle Prescription = Metrope 7/12

Spectacle Prescription = Hypermetrope 5/8

Astigmatic = Yes 6/10

Astigmatic = No 6/10

Tear Production Rate = Normal 2/10

Tear Production Rate = Reduced 10/10

Choose Attribute = Attribute Value with the largest n/t value, in this case Tear Production Rate = Reduced.

**Rule: If Tear Production Rate = “Reduced” then Recommendation = “None”.**

**Using this rule the dataset would be reduced to all instances having recommendation = “none” and therefore the rule growing would stop. The classification rule for class “none” is:**

**If Tear Production Rate = “Reduced” then Recommendation = “None”**

**Problem 6. [10 points]** GivenCPU Performance data set. The task is to predict the Estimated Relative Performance (ERP) based on a number of attributes, namely MYCT, MMIN, MMAX, CACH, CHMIN, CHMAX. The linear regression model for the dataset is

ERP = -55.9 + 0.0489MYCT + 0.153MMIN + 0.0056MMAX + 0.6410 CACH – 0.27CHMIN + 1.48CHMAX.

Predict ERP for the following instance:

MYCT MMIN MMAX CACH CHMIN CHMAX ERP

125 256 600 256 16 128 ?

If the the PRP (Performance real value) is 198, what is the prediction error for the instance using the given linear regression model? **Show all your work!**

**ERP** = -55.9 + 0.0489(125) + 0.153(256) + 0.0056(600) + 0.6410(256) – 0.27(16) + 1.48(120) **= 330.12**