CS4445 Data Mining and Knowledge Discovery in Databases. B Term 2012

Solutions Exam 1 - November 19, 2012

Prof. Carolina Ruiz
Department of Computer Science
Worcester Polytechnic Institute

NAME: _Prof. Carolina Ruiz____

Problem I: **(/15 points)** Data Preprocessing

Problem II: **(/10 points)** Model Evaluation

Problem III: **(/80 points)** Classification

TOTAL SCORE: (/100 points)

Instructions:

- Show your work and justify your answers
- Use the space provided to write your answers
- Ask in case of doubt

Problem I. Data Preprocessing [15 points]

1. **[5 points]** What is the difference between simple random sampling and stratified random sampling?

Simple random sampling draws data instances at random using a uniform distribution (that is each data instance is equally likely to be chosen), while stratified random sampling draws data instances at random according to the distribution of the target attribute (so that the subsample preserves the distribution of the target attribute).

2. **[5 points]** Briefly describe a similarity between Correlation Based Feature Selection (CFS) and Principal Components Analysis (PCA).

They both are dimensionality reduction techniques.

3. **[5 points]** Briefly describe a difference between Correlation Based Feature Selection (CFS) and Principal Components Analysis (PCA).

CFS is a feature selection technique (that is, it selects some of the original data attributes to keep and removes the rest), while PCA is a feature extraction techniques (that is, it creates new attributes from the original ones).

Problem II. Model Evaluation [10 points].

The ***leave-one-out*** test method is defined as a particular case of *n*-fold cross-validation by taking *n* to be equal to the number of data instances in the dataset.

1. **[5 points]** Use your knowledge of *n*-fold cross validation to briefly explain how the ***leave-one-out*** test method works. How is the accuracy reported by this test method computed?

Leave-one-out works as follows:

```
For k = 1 to n  (where n is the number of data instances in the dataset)
    Use data instances $i_1$, $i_2$, …, $i_{(k-1)}$, $i_{(k+1)}$, …, $i_n$ as the training set to construct a model $M_k$.
    Use the data instance $i_k$ to test $M_k$. Let $acc_k$ be the accuracy of $M_k$ on instance $i_k$.
Output the average of $acc_1$, …, $acc_n$.
```

2. **[5 points]** Briefly describe an advantage and a disadvantage of this test method.

Advantage: Uses as much data as possible to construct each model while still testing each model with a data instance not used during its construction. Since the process is repeated for each data instance, each data instance is used as the test instance at some point.

Disadvantage: The process might take a long time, especially is the dataset contains a large number of data instances.

Problem III. Classification [80 points]

Consider the following training dataset consisting of 12 data instances:

AGE	SPECTACLE-PRESCRIPTION	ASTIGMATISM	TEAR-PRODUCTI0N-RATE	CONTACT LENSES
pre-presbyopic	myope	no	normal	soft
young	myope	no	normal	soft
presbyopic	hypermetrope	no	normal	soft
presbyopic	myope	yes	normal	hard
pre-presbyopic	myope	yes	normal	hard
young	myope	yes	normal	hard
presbyopic	myope	no	normal	none
pre-presbyopic	hypermetrope	yes	normal	none
presbyopic	hypermetrope	yes	reduced	none
pre-presbyopic	myope	yes	reduced	none
young	hypermetrope	no	reduced	none
pre-presbyopic	myope	no	reduced	none

Let CONTACT LENSES be the classification target. In this problem, we will investigate different model construction techniques and how the resulting models classify the following test data instance:

AGE	SPECTACLE-PRESCRIPTION	ASTIGMATISM	TEAR-PRODUCTI0N-RATE	CONTACT LENSES
?	myope	yes	normal	?

Note: *For all model construction techniques below, if you need to break ties use the class values in the order they appear in the dataset: soft, hard, and none.*

1. **ZeroR:**
   a. **[4 points]** Construct the ZeroR model for the training dataset. Explain your work.

The distribution of the classes is: *soft*: 3/12, *hard*: 3/12, *none*: 6/12.

Hence, the ZeroR model always predicts the majority class: *none*.

   b. **[1 points]**: What CONTACT LENSES class does ZeroR predict for the test data instance?

It predicts class = *none.*

2. **OneR**:
   a. **[15 points**:] Construct the OneR model over the training dataset. Show all candidate one-attribute rules. Remember that OneR uses classification accuracy over the training set as the metric to select the best among candidate one-attribute rules. Show your work in detail.

A candidate one-attribute rule is created for each attribute:

- Candidate rule for Age:

If	AGE = young	then	class = soft
	AGE = presbyopic	then	class = none
	AGE = pre- presbyopic	then	class = none

Classification accuracy of this rule over the training dataset: 6/12

- Candidate rule for PRESCRIPTION:

If	PRESCRIPTION = myope	then	class = hard
	PRESCRIPTION = hypermetrope	then	class = none

Classification accuracy of this rule over the training dataset: 6/12

- Candidate rule for ASTIGMATISM:

If	ASTIGMATISM = yes	then	class = hard
	ASTIGMATISM = no	then	class = soft

Classification accuracy of this rule over the training dataset: 6/12

- Candidate rule for TEAR-PROD-RATE:

If	TEAR-PROD-RATE = normal	then	class = soft
	TEAR-PROD-RATE = reduced	then	class = none

Classification accuracy of this rule over the training dataset: 7/12

Since its accuracy over the training set is the highest among all the candidate rules, the TEAR-PROD-RATE rule is chosen as the OneR model.

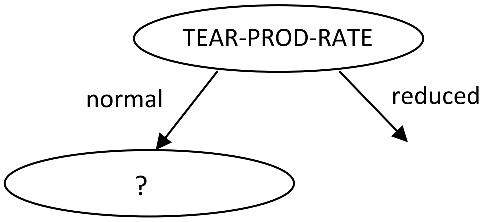
   b. **[2 points]**: What CONTACT LENSES value does OneR predict for the test data instance? Explain.

Since TEAR-PROD-RATE of the test instance is normal, then OneR predicts class = soft for this instance.

3. **Decision Trees**:
    a. **[20 points**:] Construct (part of) the ID3 decision over the training dataset. To save time, follow these instructions:
        i. Assume that TEAR-PROD-RATE is the root note of the tree (no need to show why this is the case).
        ii. Do NOT consider the AGE attribute (assume that it won't appear in the tree).
        iii. *At each step, construct just the one branch needed to classify the test data instance*.
        iv. Assume that the minimum number of instances in a leaf node is 1.
        v. Use the following log values in your entropy calculations.

x	1/2	1/3	2/3	1/4	3/4	1/5	2/5	3/5	1/6	5/6	1/7	2/7	3/7	4/7	3/8	1
log$_2$(x)	-1	-1.5	-0.6	-2	-0.4	-2.3	-1.3	-0.7	-2.5	-0.2	-2.8	-1.8	-1.2	-0.8	-1.4	0

Show all your work neatly.



In the node under construction, there are 8 data instances with TEAR-PROD-RATE= normal:

AGE	SPECTACLE-PRESCRIPTION	ASTIGMATISM	TEAR-PRODUCTI0N-RATE	CONTACT LENSES
pre-presbyopic	myope	no	normal	soft
young	myope	no	normal	soft
presbyopic	hypermetrope	no	normal	soft
presbyopic	myope	yes	normal	hard
pre-presbyopic	myope	yes	normal	hard
young	myope	yes	normal	hard
presbyopic	myope	no	normal	none
pre-presbyopic	hypermetrope	yes	normal	none

Since some of them are soft, hard, and none, we need to split this node. Ignoring AGE, there are two possible attributes to split the node: SPECTACLE-PRESCRIPTION and ASTIGMATISM. Let's calculate the entropy of each of them with respect to the 8 data instances with TEAR-PROD-RATE=normal.

Entropy of SPECTACLE-PRESCRIPTION (values: myope and hypermetrope) with respect to the 8 data instances with TEAR-PROD-RATE=normal:
= ENTROPY([2,3,1],[1,0,1])

= (6/8)* ENTROPY([2,3,1]) + (2/8)*ENTROPY([1,0,1])

= $(6/8)*[-(2/6)\log_2(2/6) – (3/6)\log_2(3/6)-(1/6)\log_2(1/6)] \quad + \quad (2/8)*[-(1/2)\log_2(1/2) – 0 – (1/2)\log_2(1/2)]$

= (3/4)*[-(1/3)(-1.5) – (1/2)(-1) – (1/6)(-2.5)] + (1/4)*[- (1/2)(-1) – (1/2)(-1)]

= 1.0625 + 0.25 = 1.3125

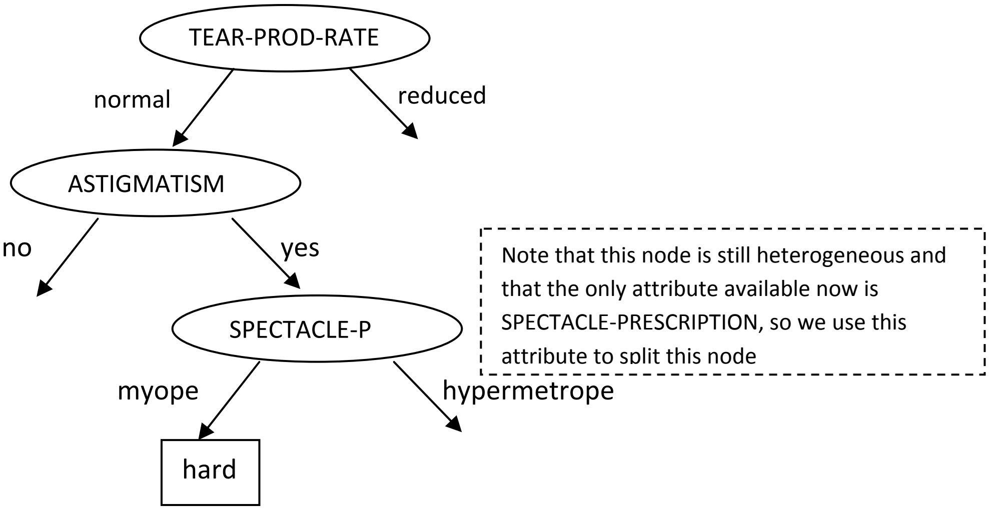
Entropy of ASTIGMATISM (values: no and yes) with respect to the 8 data instances with TEAR-PROD-RATE=normal:
= ENTROPY([3,0,1],[0,3,1])

= (4/8)* ENTROPY([3,0,1]) + (4/8)*ENTROPY([0,3,1])

= $(4/8)*[-(3/4)\log_2(3/4) – 0 - (1/4)\log_2(1/4)] \quad + \quad (4/8)*[-0 -(3/4)\log_2(3/4) – (1/4)\log_2(1/4)]$

= [-(3/4)(-0.4) – (1/4)(-2)] = 0.8

Since the entropy of ASTIGMATISM is lower than that of SPECTACLE-PRESCRIPTION on this set of instances, ASTIGMATISM is chosen to split the node.



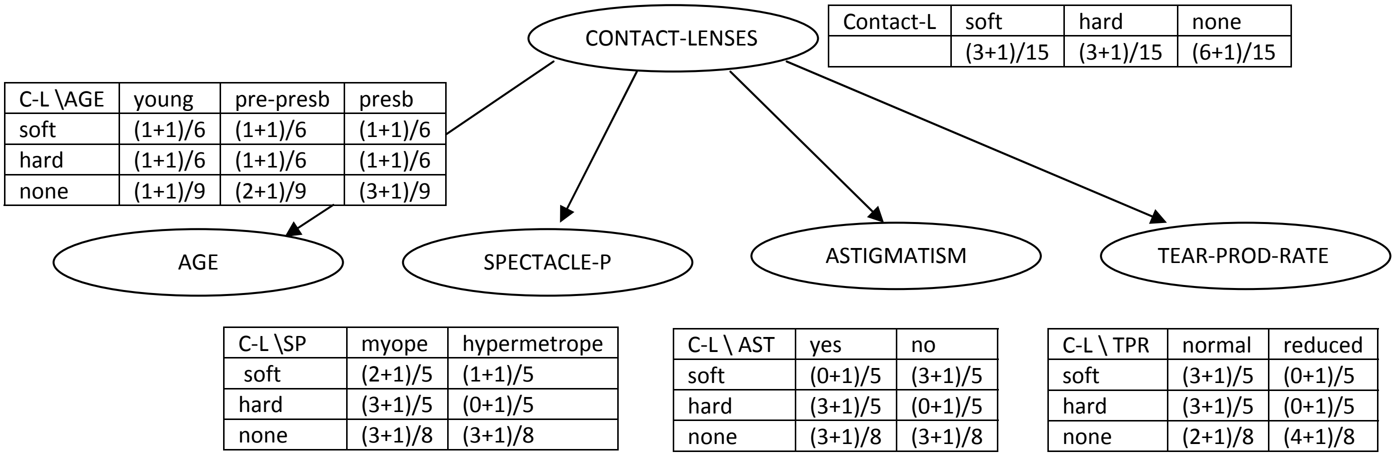
Note that this node is still heterogeneous and that the only attribute available now is SPECTACLE-PRESCRIPTION, so we use this attribute to split this node

b. **[3 points]**: What CONTACT LENSES class does the decision tree predict for the test data instance? Explain.

Since TEAR-PROD-RATE=normal, ASTIGMATISM=yes, and SPECTACLE-PRESCRIPTION=myope, the tree will predict class=hard, according to the tree branch constructed above.

4. **Naïve Bayes**:

   a. **[5 points]** Display the topology of the naïve Bayes graph for the training dataset.
   [15 points] Compute all of the Conditional Probability Tables (CPTs) in the graph. Show your work neatly.



Contact-L	soft	hard	none
	(3+1)/15	(3+1)/15	(6+1)/15

C-L \AGE	young	pre-presb	presb
soft	(1+1)/6	(1+1)/6	(1+1)/6
hard	(1+1)/6	(1+1)/6	(1+1)/6
none	(1+1)/9	(2+1)/9	(3+1)/9

C-L \SP	myope	hypermetrope
soft	(2+1)/5	(1+1)/5
hard	(3+1)/5	(0+1)/5
none	(3+1)/8	(3+1)/8

C-L \ AST	yes	no
soft	(0+1)/5	(3+1)/5
hard	(3+1)/5	(0+1)/5
none	(3+1)/8	(3+1)/8

C-L \ TPR	normal	reduced
soft	(3+1)/5	(0+1)/5
hard	(3+1)/5	(0+1)/5
none	(2+1)/8	(4+1)/8

   b. **[15 points]** What CONTACT LENSES class does the naïve Bayes model predict for the test data instance? Show your work in detail.

The naïve Bayes model predicts the class value that maximizes:

P(class=v | AGE=? & SPECT=myope & ASTIG=yes & T-P-RATE=normal). Using Bayes theorem and the naïve assumption that AGE, SPECTACLE-PRESCRIPTION, ASTIGMATISM and TEAR-PROD-RATE are independent from each other given CONTACT-LENSES, the probability above can be re-written as:

argmax P(AGE=?|class=v) P(SPECT=myope|class=v) P(ASTIG=yes|class=v) P(T-P-RATE=normal|class=v)P(class=v)
v in {soft,hard,none}

Note that P(AGE=?|class=v) = P(AGE=young|class=v) + P(AGE=pre-presb|class=v) + P(AGE=presb|class=v) = 1.

- For v = soft:
  P(SPECT=myope|class=soft)*P(ASTIG=yes|class=soft)*P(T-P-RATE=normal|class=soft)*P(class=soft)
  = (3/5)*(1/5)*(4/5)*(4/15) = 48/1875 = 0.0256
- For v = hard:
  P(SPECT=myope|class=hard)*P(ASTIG=yes|class=hard)*P(T-P-RATE=normal|class=hard)*P(class=hard)
  = (4/5)*(4/5)*(4/5)*(4/15) = 256/1875 = 0.1365
- For v = none:
  P(SPECT=myope|class=none)*P(ASTIG=yes|class=none)*P(T-P-RATE=normal|class=none)*P(class=none)
  = (4/8)*(4/8)*(3/8)*(7/15) = (1/2)*(1/2)*(1/8)*(7/15) = 7/160 = 0.04375

Hence the naïve Bayes model classifies the test data instance as class = hard.