CS4445 Data Mining and Knowledge Discovery in Databases. B Term 2014

Exam 1 November 24, 2014

Prof. Carolina Ruiz
Department of Computer Science
Worcester Polytechnic Institute

NAME: ____Prof. Ruiz ____

Problem I:	**(/10 points)** Data Preprocessing
Problem II:	**(/15 points)** Model Evaluation
Problem III:	**(/30 points)** Decision Trees
Problem IV:	**(/45 points)** Bayesian Models
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 [10 points]

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

Solution: (Taken from the solutions to Exam 1 CS4445 B Term 2012)
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]** Assume that A is a nominal attribute, other than the target attribute. Consider a missing value for this attribute A.
   a. Briefly describe a possible unsupervised method to replace this missing value.

   Solution: Replace the missing value with the mode of attribute A.
   [This is an unsupervised method because it doesn't use the target attribute at all.]

   b. Briefly describe a possible supervised method to replace this missing value.

   Solution: Replace the missing value with the mode of attribute A on the data instances that have the same classification (target value) of the instance that contains the missing value.
   [This is supervised method because it uses the target attribute to modify A.]

Problem II. Model Evaluation [15 points].

1. **[10 points]** Explain how *n*-fold cross validation works (to make it easier to explain, use *n*=10). How is the accuracy reported by this evaluation method computed?

   Solution: (Taken from the solutions to Exam 1 CS4445 D term 2003. Against my own suggestion above, I will explain the procedure for a general n rather than using n=10)

   Partition the input data into n folds (i.e., mutually disjoint and collectively exhaustive parts), approximately of the same size, at random using stratification. Let's denote those folds as F1,F2,..., Fn. Now, perform the following process:

```
 For i := 1 to n do
   - construct model Mi using as training data the union of all folds except for Fi.
     That is, the union of F1, ..., F(i-1), F(1+1), ..., Fn
   - test model Mi on fold Fi, and record the accuracy (or the error) obtained.
End For
Return the average of the accuracies (or of the errors) of all the models Mi.
```

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

   Solution:
   [Although performing n-fold cross validations has several advantages, we discuss just one of them here as that's all is required by the problem statement.]

   Advantage: This systematic procedure allows each and every instance in the dataset to be part of the training set in some experiments (n-1 to be precise) and of the test set in other experiments (1 to be precise).

   Disadvantage: The process might take a long time, as n models are constructed and tested.

Problem III. Decision Trees [30 points]

An alternative metric for selecting the best attribute to split a node in a decision tree is the **Gini** metric. Below are some facts about the Gini metric.

- The formulas for the Entropy and for the Gini metrics are:

$$Entropy(t) = -\sum_{i=1}^{c} p(i|t)\log_2 p(i|t) \qquad and \qquad Gini(t) = 1 - \sum_{i=1}^{c}[p(i|t)]^2$$

  where c is the number of classes (i.e., values of the target attribute) and $p(i|t)$ is the relative frequency of class i at node t.
- As with Entropy, the Gini value of an attribute is the weighted sum of the Gini values of each of the attribute values.
- As with Entropy, the attribute with the lowest Gini value is selected to split the tree node.

Consider the following dataset of 10 data instances. Assume that **Defaulted Borrower** is the target attribute.

Home Owner (H)	**Marital Status (M)**	**Annual Income (A)**	**Defaulted Borrower (D)**
no	divorced	>85K	yes
yes	divorced	>85K	no
no	married	>85K	no
yes	married	>85K	no
no	married	≤85K	no
no	married	≤85K	no
yes	single	>85K	no
no	single	≤85K	no
no	single	≤85K	yes
no	single	>85K	yes

The Gini values of the predicting attributes for this dataset are:

Gini value of **House Owner** is 0.3428

Gini value of **Marital Status** is 0.3

Gini value of **Annual Income** is 0.4166

1. **[10 points]** Using the formula for Gini, show that the Gini value of **Annual Income** is indeed 0.4166. Show you work (please use the notation "[# of no's , # of yes'es]" to neatly summarize the counts).

Solution: The [no, yes] counts for ≤85K are [3,1] and the [no,yes] counts for >85K are [4,2].

Gini(A) = Gini([3,1],[4,2])

= (4/10)*Gini([3,1]) + (6/10)*Gini([4,2])

= (4/10)*[1 – [(3/4)^2 + (1/4)^2]] + (6/10)*[1 – [(4/6)^2 + (2/6)^2]]

= (4/10)*[1 – [(9/16) + (1/16)]] + (6/10)*[1 – [(16/36) + (4/36)]]

= (4/10)*[1 – (10/16)] + (6/10)*[1 – (20/36)] = (4/10)*(6/16) + (6/10)*(16/36) = (3/20)+(4/15)=0.4166

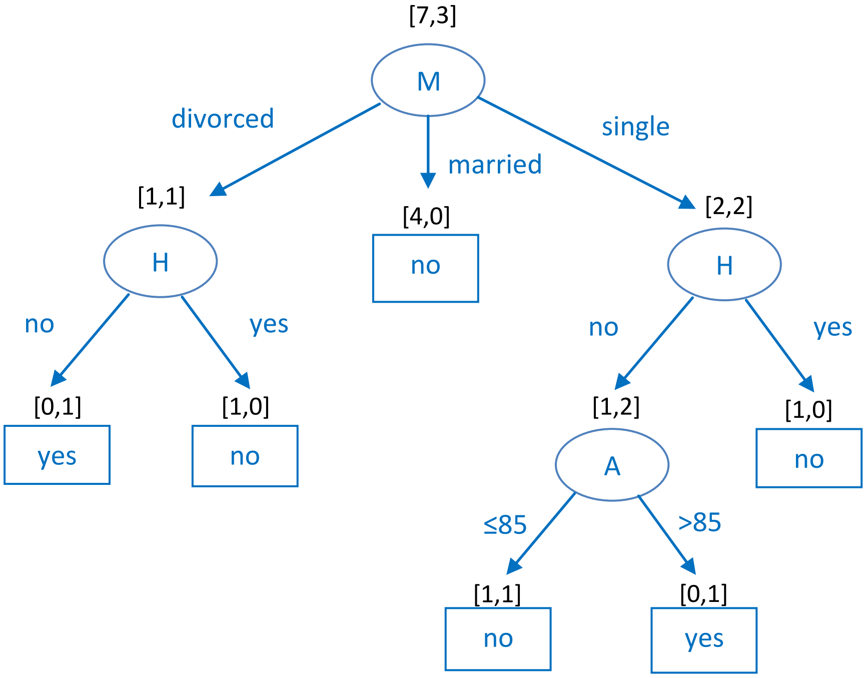
2. **[20 points]** Construct the full ID3 decision tree using Gini to rank the predicting attributes (**Home Owner, Marital Status, Annual Income**) with respect to the target/classification attribute (**Defaulted Borrower**).
   - For the root node, you can assume that the Gini value of **House Owner** is 0.3428, the Gini value of **Marital Status** is 0.3, and the Gini value of **Annual Income** is 0.4166 without calculating these values explicitly.
   - For nodes other than the root, show all the steps of your Gini calculations.

   Make sure to show your work.

Solution: Since Marital Status has the lowest Gini value, it is chosen to split the root node.

- For M=divorced (left-most child), has no/yes count [1,1]. By simple inspection, Home Owner perfectly splits this node, while Annual Income doesn't split it. Hence, we select Home Owner to split this node.
- For M=married (middle child), the node is homogenous [4,0], so it is converted into a leaf.
- For M=single (right-most child), the node is heterogeneous [2,2] and neither Home Owner nor Annual Income splits it perfectly well. So we calculate the Gini value of these two attributes for this node:
  - Gini(H) = Gini([1,2],[1,0]) = (3/4)*Gini([1,2]) + (1/4)*Gini([1,0]) = (3/4)*[1-[(1/3)^2 + (2/3)^2]]+0
    = (3/4)*[1 – (5/9)] = (3/4)*[4/9] = 1/3 = 0.33
  - Gini(A) = Gini([1,1],[1,1]) = (2/4)*Gini([1,1]) + (2/4)*Gini([1,1]) = [1-[(1/2)^2 + (1/2)^2]]
    = [1 – (1/2)] = 1/2 = 0.5

  Hence, Home Owner is chosen to split this node. The H=yes child node is homogeneous so we make it into a leaf. The H=no child node is heterogeneous, so we split it with the only remaining attribute available in that subtree, namely A. One of children of A is still heterogeneous [1,1], but since there are no more attributes available to split it, we convert it into a leaf and break the tie choosing the first class value listed on the dataset, namely "no", following Weka's convention.



Problem IV. Bayesian Models [45 points]

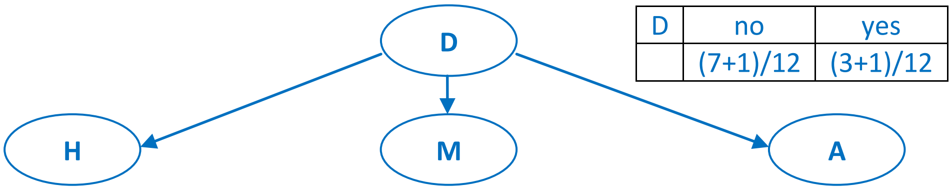
Consider the following dataset, where **Defaulted Borrower** is the target attribute:

Home Owner (H)	**Marital Status (M)**	**Annual Income (A)**	**Defaulted Borrower (D)**
yes	divorced	>85K	no
no	married	>85K	no
yes	married	>85K	no
no	married	≤85K	no
no	married	≤85K	no
yes	single	>85K	no
no	single	≤85K	no
no	divorced	>85K	yes
no	single	≤85K	yes
no	single	>85K	yes

1. **Naïve Bayes**:
   a. **[5 points]** Display the topology of the naïve Bayes graph for the training dataset.

   [10 points] Compute all of the Conditional Probability Tables (CPTs) in the graph. Show your work neatly.

Solution:



D	no	yes
	(7+1)/12	(3+1)/12

D	H	
	no	yes
no	(4+1)/9	(3+1)/9
yes	(3+1)/5	(0+1)/5

D	M		
	divorced	married	single
no	(1+1)/10	(4+1)/10	(2+1)/10
yes	(1+1)/6	(0+1)/6	(2+1)/6

D	A	
	≤85	>85
no	(3+1)/9	(4+1)/9
yes	(1+1)/5	(2+1)/5

   b. **[15 points]** Determine the **Defaulted Borrower** value that this naïve Bayes model predicts for the test data instance: **Home Owner** = yes, **Marital Status** = single and **Annual Income** ≤85K (let's abbreviate this as: **H**=yes, **M**=single and **A**≤85K). Show your work in detail.

Solution: The prediction of the Naïve Bayes model for this data instance is:

$$\underset{v}{\text{argmax}}\ \ P(\mathbf{D}=v \mid \mathbf{H}=yes\ \&\ \mathbf{M}=single\ \&\ \mathbf{A}\leq 85K) = \underset{v}{\text{argmax}}\ \ P(\mathbf{H}=yes\ \&\ \mathbf{M}=single\ \&\ \mathbf{A}\leq 85K \mid \mathbf{D}=v)\ P(\mathbf{D}=v)$$

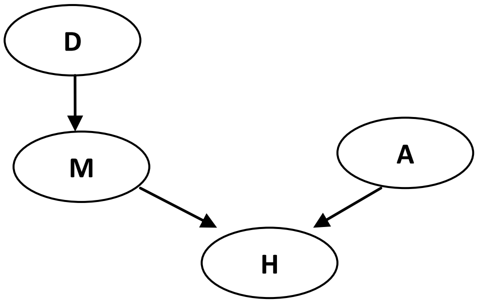
$$= \underset{v}{\text{argmax}}\ \ P(\mathbf{H}=yes \mid D=v)\ P(\mathbf{M}=single \mid D=v)\ P(\mathbf{A}\leq 85K \mid \mathbf{D}=v)\ P(\mathbf{D}=v)$$ because of the naïve assumption

For v= no: (4/9) (3/10) (4/9) (8/12) = 16/405 = 0.0395

For v= yes: (1/5) (3/6) (2/5) (4/12) = 1/75 = 0.013

Since v=no gets the highest probability, then the naïve Bayes model predicts "no".

2. Consider the following Bayesian net for the above dataset:



We want to determine the **Defaulted Borrower** value that this Bayesian net predicts for the test data instance: **H**=yes, **M**=single and **A**≤85K. One can prove (but you don't need to do so) that the prediction of this Bayesian net will be the following:

Predicted value of **D** =

$$= \underset{v}{\text{argmax}}\ \ P(\mathbf{D}=v \mid \mathbf{H}=\text{yes}\ \&\ \mathbf{M}=\text{single}\ \&\ \mathbf{A}\leq 85\text{K})$$

$$= \underset{v}{\text{argmax}}\ \ P(\mathbf{H}=\text{yes}\ \&\ \mathbf{M}=\text{single}\ \&\ \mathbf{A}\leq 85\text{K} \mid \mathbf{D}=v)\ P(\mathbf{D}=v)$$

$$= \underset{v}{\text{argmax}}\ \ P(\mathbf{H}=\text{yes} \mid \mathbf{M}=\text{single}\ \&\ \mathbf{A}\leq 85\text{K})\ P(\mathbf{M}=\text{single} \mid \mathbf{D}=v)\ P(\mathbf{A}\leq 85\text{K})\ P(\mathbf{D}=v)$$

a. **[5 points]** Assume that all the probability values above are different from 0. Simplify the last line of the derivation above as much as you can, eliminating probability expressions that don't need to be considered. Explain your answer.

Solution: Since P(**H**=yes | **M**=single & **A**≤85K) and P(**A**≤85K) don't involve D=v, they won't affect the result of the argmax. In other words, they are constant with respect to v. Hence, they can be eliminated from the last line of the derivation above without affecting the result:

$$= \underset{v}{\text{argmax}}\ \ P(\mathbf{M}=\text{single} \mid \mathbf{D}=v)\ P(\mathbf{D}=v)$$

b. **[10 points]** Using your simplified formula, determine the **Defaulted Borrower** value that this Bayesian net will predict for this test data. Calculate explicitly only the entries of the Conditional Probability Tables (CPTs) that you need in order to answer this question. Show your work.

Solution: $\underset{v}{\text{argmax}}\ \ P(\mathbf{M}=\text{single} \mid \mathbf{D}=v)\ P(\mathbf{D}=v)$

For v= no: (3/10) (8/12) = 1/5

For v= yes: (3/6) (4/12) = 1/6

[Note that the CPT tables for D and for M on this Bayesian net are identical to the ones calculated for the naïve Bayes model.]

Since v=no gets the highest probability, then this Bayesian net model predicts "no".