CISC-483-683: Assignment 4: Decision Trees and Numeric Attributes

60 points

Due Friday, Sept. 29, 2023

1. (20 points) Consider the following data set collected by DOG-ADOPT (a dog rescue center), where the Class attribute is TAKE-HOME and the predictor attributes are SHOTS, HAIR, and SIZE:

SHOTS	HAIR	SIZE	TAKE-HOME
all	long	tiny	No
all	long	tiny	Yes
all	long	tiny	Yes
all	short	tiny	Yes
none	short	huge	No
all	short	huge	No
none	short	huge	No
all	long	huge	No
all	short	huge	No
all	short	huge	Yes
all	long	medium	Yes
none	short	huge	No
none	short	huge	No
all	long	medium	Yes
all	long	medium	Yes
all	short	medium	No
all	long	medium	Yes
none	short	huge	No
none	short	huge	No
all	short	huge	No

   (a) (10 points) Consider only the predictor attributes SHOTS and HAIR, and the Class attribute TAKE-HOME

      i. Using information gain (or weighted entropy) as the split criteria, which attribute (SHOTS or HAIR) would be at the root of the decision tree? (Please show your work in detail.)

      ii. Using Gini index as the split criteria, which attribute (SHOTS or HAIR) would be at the root of the decision tree? (Please show your work in detail.)

      iii. What does this tell you about the preferences of information gain (or weighted entropy) versus the preferences of the Gini index regarding the choice of split attribute?

   (b) (10 points) Using the Gini index, consider all three predictor attributes (SHOTS, HAIR, and SIZE) along with the Class attribute TAKE-HOME. Which attribute (SHOTS, HAIR, or SIZE) would be at the root of the decision tree? (Please show your work in detail.)

2. (10 pts.) For several years, the University of Delaware has had difficulty predicting which high school students would accept Delaware's offer of admission. Alice has developed a new model for predicting whether a high school senior, who has been offered admission to the University, will actually enroll. The model is based on state of residency, student's high school GPA, the student's SAT scores, the student's major, etc. Alice wants to sell her model to the Admissions Office. She tells them that she has tested her model on 8000 high school seniors from previous years who were offered admission to the University and that her model made a correct prediction for 7000 of them. The Admissions Office would very much benefit from Alice's model if it truly gives good predictions. With 95% confidence, what can Alice tell the Admissions Office is the true range of success of her model?

3. (30 points) Suppose that you have a movie dataset with several attributes, including a numeric attribute giving the length of the movie in minutes. The Class attribute is DECISION whose value is Watch, Save, or Reject. Consider the following values for the numeric attribute LENGTH and the following class values:

LENGTH	DECISION
100	Save
109	Watch
115	Reject
121	Watch
135	Reject
140	Reject
144	Reject
150	Watch
160	Watch
163	Watch
165	Watch
169	Save
171	Watch
172	Watch
174	Watch
176	Reject
178	Watch
180	Watch
184	Watch
186	Watch
188	Watch
190	Save

(a) Suppose that you are using top-down, supervised, entropy-based discretization. Which is a better first split point, 118 or 147? Show your work.

(b) Suppose that you are using bottom-up, supervised, chi-square based discretization. Suppose that you have already constructed the following intervals: (-∞,112), (112,147), (147,175), and (175,∞).

i. Using a confidence level of .90 (ie., want 90% confidence that we can reject the null hypothesis that the intervals do not differentiate the attributes class values), do you merge the interval (112,147) with the interval (147,175)? Show your work.

ii. Using a confidence level of .90, do you merge the interval (147,175) and the interval (175,∞)? Show your work.