COSC 4570/5010 Data Mining
Spring 2020
University of Wyoming

Homework #2

Due: ~~February 21, 2020, 10:00 a.m.~~ February 26, 2020, 10:00 a.m.

Submission guideline You need to submit only one *.zip* file. Please name the file as "Your Net id_Homework2.zip".

1. **Problems from the book (Introduction to Data Mining 2ⁿᵈ Edition by Tan, Steinbach et al.)**

   Solve the following:

   Chapter 3: Problems 1, 5, 7, and 10.

   Chapter 4: Problem 6

OR

Problems from the book (Introduction to Data Mining 1ˢᵗ Edition by Tan, Steinbach et al.)

   Solve the following:

   Chapter 4: Problems 1, 5, 6, and 9.

   Chapter 5: Problem 6

2. **Decision Tree Learning**

   - What does zero entropy mean?

   - What is maximum value for the entropy of a random variable that can take n values? justify.

   - What kind of real attributes create problems for entropy-based decision trees. How can we solve this problem?

- Describe pre-pruning and post-pruning techniques for dealing with decision tree overfitting.

- Is the Gini gain (Gini of the parent subtracted by the Gini of the split) always positive? What about entropy's gain? What if you use classifications error? Prove or provide counterexamples.

3. Naive Bayes Classifier

- What is the time complexity for learning a Naive Bayes Classifier?

- What is the time complexity for classifying using the Naive Bayes Classifier?