Jason Zigelbaum

Prof. Zhang

Machine Learning

March 26th, 2012

Problem 1.

This problem requires extensive reshaping of the data. This is because performing an optimal bayes calculation requires lots of sample data in order to get meaningful data regarding each data point. In order to calculate the optimal bayes classifier, one must perform the following calculation:

Probability(class) \* Probability(attribute-1, attribute-2… attribute-n | class)

Over all possible classifications. Then one simply chooses classification with the maximum probability of occurring as the classifier.

The reason that optimal bayes is typically not used is because it requires significant amounts of data to get any meaningful probabilities. If we were to take into account our training data for example, we would have no repeating data points and therefore any optimal bayes classification is as good as random.

The trick to effectively utilize an optimal bayes classification is to re-shape our data so that:  
 1. Our continuous data becomes discrete

2. Our discrete data has a small enough state space so that some reasonable probabilistic inference can be made.

Because the testing data is so small (relative to the state space of it’s attributes) I am going to have to group them by some rather broad attributes. I performed the following reshaping of the data, using the following rules, and pertaining to the following logic.

1. The particular day of a data point is irrelevant. Snow geese do not work on weekdays, they do not go to church on Sunday and they do not have a calendar. Therefore, I am making the assumption their behavior is only subject to weather conditions.

RULE: Remove date as an attribute.

2. Temperature can be grouped together by range. This is a necessary change since we do not have nearly enough data points to achieve any probabilistic correspondence between the specific degree and behavior of the geese. I am choosing to group the weather into four classifications:

1. A : > 15 degrees

2. B : >= 10 and <= 15

3. C : > 0 and < 10

4. D : < 0

3. Humidity can, like temperature, be grouped by range. In my delineation of the ranges I am trying to keep an even spread in each grouping (within reason) so that there can be enough data points within each range to deduce a meaningful probabilistic inference for our classification (Early or Late). The classifications are as follows:

1. A :

2. B :

3. C :

4. D :

4. Again, Light can be grouped by range. I chose a delineation so that there is an even spread. The classifications are as follows:

1. A :

2. B :

3. C :

4. D :

5. Finally, Cloud is also grouped by range. The classifications are as follows:

1. A :

2. B :

3. C :

4. D :