

Assignment 1 (15% of final grade)

CS405/505 Data Mining

Due: Wednesday Oct 2, 2019

Consider the following dataset with 3 attributes (weight, waist, heartrate), two classes (no, yes) and 20 instances. The dataset describes 3 physiological variables from 20 individuals in the general population, and whether or not the individual can complete more than 10 chinups.

Your job for this assignment will be to employ some of the machine learning algorithms we have seen so far to this dataset, to answer basic questions about the dataset.

Marking:

Question 1: 20%

Question 2: 40%

Question 3: 40%

weight	waist	heartrate	>10 chinups
heavy	wide	slow	no
heavy	wide	slow	no
heavy	wide	fast	yes
light	thin	fast	yes
heavy	thin	slow	yes
heavy	wide	fast	no
heavy	wide	fast	no
light	thin	fast	no
light	thin	fast	yes
light	thin	fast	yes
light	thin	slow	yes
light	thin	slow	yes
light	thin	fast	yes
heavy	wide	slow	no
heavy	wide	slow	no
heavy	wide	fast	yes
light	wide	slow	no
light	thin	slow	yes
light	thin	slow	yes
light	thin	fast	no

Question 1) employ the 1R (1-rule) algorithm on the dataset to answer the following:

A) which attribute has the lowest error rate for predicting if >10 chinups = yes ?

B) what is the error rate of the attribute from (A) ?

1R Algorithm:

```
For each attribute,
  For each value of that attribute, make a rule as follows:
    count how often each class appears
    find the most frequent class
    make the rule assign that class to this attribute-value.
  Calculate the error rate of the rules.
  Choose the rules with the smallest error rate.
```

Question 2) employ Naive Bayes on the dataset to answer the following:

A) what is the probability of >10 chinups = yes, given the person is heavy, wide, and slow ?

B) which values of weight, waist, and heartrate yield the highest probability of >10 chinups = yes ?

Be sure to show the table of observed probabilities (Lecture_04.pptx, Slide 14)

Question 3) construct a Decision Tree on the dataset, using the Gini Index as the splitting criteria (instead of entropy), to answer the following:

A) which attribute offers the best (most informative) initial split? ie which attribute should be placed as the root for the decision tree?

B) list all the rules that can be read from the decision tree.

C) what is the most highly branching attribute for an initial split? (do not use gain ratio for this problem)

Be sure to show your final decision tree, including properly labeled branches/nodes (Lecture_05.pptx, Slide 11)

Gini index:

$$1 - \sum_i P_i^2$$

where 'i' indexes class value and 'P' is the fraction of instances assigned to class value 'i'