Datasets and preprocessing

We have a class for each of the datasets given

Read in values from CSV, cast accordingly

Delete rows with missing data

Shuffle the data

For classes that need binning we use a frequency-based binning method

Next show Naïve Bayes model

Have class probabilities dictionary as well as class count

Class counts self explanatory

Class probabilities is set by the total instances of the class divided by the number of samples

We also have a multilayerd dictionary for class attribute probabilities

The structure to access a certain attribute probability, 1st key is class, 2nd key is idx, 3rd key is value

Then the final value is the probability for that feature value given the class

We have a function called total probabilities

This function multiplies together all the specific probabilities from the attribute values

It then multiplies this product by the raw class probability.

It does this for each class and returns an array of the probabilities every class

The classify function then takes the max of the probabilities returned by the total probabilities

This is then the predicted class

Next I will show our model in action. Ill use the breast cancer dataset

For this example I will test our model on one folds hold out set

To evaluate our model we will use precision, recall, and zero one loss functions