Q2

Dataset description

The Dataset contains data of 416 liver patient records and 167 liver patient records.Out of these 441 are male patients and 167 are female patients.The majority of dataset consists of male patients and liver patients and size is small,which might be negatively impacting the accuracy of prediction

Dataset preprocessing

1. Firstly all feature rows with missing value columns are removed.
   (4 such data points were found out of 583)

2. Then all categorical variables are encoded,the only categorical variable in this dataset being 'gender'.The value that the gender column takes is 'male' and 'female' which is encoded as 0 being for female and 1 denoting male patient.Then the gender column is appropriately replaced.

   This is done by the pandas.get_dummies() function.

3. The outliers from the dataset are removed.This is done by calculating the number of outlier features for a given data point and removing those rows which have maximum outlier features.

   Any feature value having value greater than 2*mu+sigma is considered an outlier.
   (3 data points as outliers were found out of 579 data points)
4. The initial dataset has 583 data points which after removal of outliers reduces to 580 data points.

Functions definitions and their usages:

1. **test_train_split(db):**
   This functions splits dataset randomly with ratio of 70:30 into train and test data and data
2. **remove_outlier(df_in, col_name, THRESHOLD):**
   This function removes outlier rows from data which have maximum feature as outliers and returns the filtered dataset
3. **normal_pdf(x,mean,sd):**
   This function calculate probability at a point according to normal distribution where mean and standard deviation of normal distribution is given
4. **divide_data(dataset,k):**

This divides dataset into k parts for k fold cross validation

5. **calculate_parameters(train,laplace_flag):**
   Returns the mean and standard deviation of positive and negative samples of data,the prior probability of positive and negative samples and the probability of occurrence of male and female given a positive or a negative sample[likelihood of male and female patients given is liver patient or not liver patient]
6. **predict(test_data,column_headers,MEAN_1,STD_1,MEAN_2,STD_2,train_positive_gender_prob, train_negative_gender_prob, prior_is_patient, prior_is_not_patient):**
   Predicts given a test data whether prediction matches the expected outcome and returns the boolean value.This is done by using the Bayes formula considering features are independent of each other.(Pic credits:gfg)

   $$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

   $$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

   $$P(y|x_1, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1)P(x_2)...P(x_n)}$$

   Now for each feature the probability of a particular value in test data is calculated and multiplied together to give likelihood which when multiplied with the prior gives a value proportional to the probability of an outcome given a test data All such probabilities for each outcome are compared and that with maximum probability is considered the outcome prediction for the test data.

   For continuous valued features,the probability distribution is modeled as a gaussian distribution and the density function is calculated at a point to give likelihood of data feature value given an outcome.

   For discrete valued features,the probability distribution is simply done by calculating the fraction of its occurrences given a particular outcome in the train set

   ‘Gender’ is a discrete valued feature whereas the rest are continuous.

7. **find_accuracy(test,MEAN_1,MEAN_2,STD_1,STD_2,train_positive_gender_prob,train_negative_gender_prob, prior_is_patient, prior_is_not_patient):**
   This finds accuracy of data based on test data, which is trained on the training data

8. **main():**

   The main body which carries out the computation in the following order

   1. Remove data points with missing value columns
   2. Perform encoding of categorical variables
   3. Remove outliers according to the formula
   4. Find accuracy of test data after training on train data (70:30 split)
   5. Find accuracy of test data with laplace correction.(with the same training as before)
   6. Find accuracy upon performing 5 fold cross validation splitting the entire dataset into 5 equal sized sets and print average accuracy.
   7. Find accuracy upon performing 5 fold cross validation *with laplace correction* splitting the entire dataset into 5 equal sized sets and print average accuracy.

<u>Accuracy results</u>

The accuracy obtained on 5 runs of the code(without laplace correction 70:30 split)

1. 58.139534883720934
2. 62.2093023255814
3. 59.883720930232556
4. 56.395348837209305
5. 61.04651162790697

The accuracy obtained on 5 runs of the code(with laplace correction 70:30 split)

1. 58.139534883720934
2. 62.2093023255814
3. 59.883720930232556
4. 56.395348837209305
5. 61.04651162790697

The average accuracy obtained on five fold cross validation on 5 runs of the code(without laplace correction)

1. 56.58920539730134
2. 56.599700149925035
3. 55.89205397301349
4. 56.593703148425774
5. 55.73013493253374
6.

The average accuracy obtained on five fold cross validation on 5 runs of the code(without laplace correction)

```
1. 56.58920539730135
2. 56.599700149925035
3. 55.89205397301349
4. 56.593703148425774
5. 55.73013493253374
```