Project Report for K-NN classification and linear regression (ques2)

,Hem Chaitanya Reddy(T19175)
CS571P-PROGRAMMING PRACTICUM

***Abstract*—This report describes the calculation of relative frequency for each attribute of data(weight,height),mean,median calculation and plotting the each attrribute of the data.The data has been shuffeled as 70 percent and 30 percent which is used for training and testiong the linear regression and k-nn classification.The linear regression model has benn devolped with data and R^2 has been calculated.K-nn classification was implemented and confusion matrix was calculated.accuracy recall score,f-score and other parameters have been calculated.The enitre figures are plotted using seaborn and matplotlib packages.The data used in this model contains 20,000 samples and it has been taken from moodle and knn classification takes time to get the output.It takes 5 min to compile the code**

***Index Terms*—linear regression,clustering,k-nn classification**

I. Introduction

Relative frequncy one of the important parameter to get the the idead about the data.In statistics in order to calculate the relationship between different data types we will extensively use relative frequency ,mean median and other important parameters.We can plot the data types and calculate the relation ship between datatypes.But you can't predict the data behaviour by using normal mathematical models.For this we use new branch of mathematical algorithms called linear regression and K-nn classification.

A. Linear Regression and k-nn classification:

Linear Regression is machine learning algorithm .It is used to find relationship between two variables by fitting linear equation to observe the data.In this project we used simple linear regression to fit the data.Here between two data variables,one data is predictor or independent variable and other variable is response or dependent variable.From the data,The "heights" are taken as predictor and weights taken as response. This algorithm looks for statistical relationship, not deterministic relationship between variables. The relationship between two variables is said to be deterministic ,if one variable can be accurately expressed by the other.For example,relationship between temperature and Fahrenheit can be accurately predicted and this relationship is called deterministic relation ship. But in statistical relationship we can not accurately predict the relation ship between two variables.For example,the relationship between heights and weights.A linear regression has following mathematical equation

$$Y = b1X + b2 \tag{1}$$

.To calculate b1,b2 various algorthms such as least square algorithm and method of gradient desecnt algorthm have been used.The detailed explination is given in the following sections.

Another important algoritm That has been widely discussed throught the literature is k-nn calssifcation algorithm.It is simple algorithm that stores the all the avalible cases and classifes new cases based on the similarity measure using various functions such as distance functions.A case is classified by using a majority vote of its neighbors based on the k value, with the case being assigned to the class that is most common amongst its" K nearest neighbors" that is measured measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbor.The distance function here used is eucldian distance.The formule for euculedian disrtnce given by

$$dist = \sqrt{(x_i - y_i)^2} \tag{2}$$

Inorder to obseve result various plotting packages such as matplotlib and seaborn plot have been used.

II. Algorithms implemented:

For this project relative frequency of both weight,height has been calculated.Mean,median and other parameters are also calculated.The linear regression model is used between height and weight to fit linaear relationship between both of them.Also data has been classied as weiht and height by usin gk-nn clasification based on height and weight.

A. Relative frequency and other parameters:

Relative frequny gives distribution of the data in the respective range of the data.To calculate relative frequecy first we have to fix the range

- The data is read using pandas.
- The range is fixed as 5 for weight and 10 for height to calculate the relative frequency.
- for example if the values is in between $50 < wei < 55$ then 1 is added to its range.
- for this a distonary is creted and relative frequncy values have been placed in their respective range.The distonary consits of all the range values like.

  if $50 \leq wei \geq 55$
  d(50-55)+ =1
- all the relativefrequency values that have been calculated has been divided by 20,000 and multipled with 100 and we are able to get percentage for range of values.
- the mean,median are calculated using numpy and pandas.

- we are unable to calculate the mode because we have different values

B. Linear regression:

Linear regression is important stastical algorithm that is widely used in the machine learning.It is used to estimate the respose 'y' byusing input variable 'x' with the help of equation

$$y = b_0 + b_1 X \tag{3}$$

To estimate y we need to calculate b1 and b2.For this there are many algorithms.The startiong step for every algorithm is divide the data into training and testiong.

1) sorting the data::

- To divide the data for training and testing we will use csvreader package and random package.
- We will create a fuction that will divide data into traing and testing.With Csv reader we will open the csv file.This csv file is sent into the function.The entire csv file is read into lines which is turned into lists.This list is copied into data set.
- by using random.random function we will split the data into 70 percent for training and 30 percent for testing

2) coeffiecient calculation:: The coefficient for simple linear regression has to be calculted.

- first all the data that is divided into testing and training is used and converted into arrays.The mean is calulated using numpy.
- Once mean is calculate the b1 and b2 values are calculted by using following equations.

$$b1 = \frac{\sum_{i=0}^{m}((x_i - mean(x_i) * (y_i - mean(y_i)}{\sum_{i=0}^{m}((x_i - mean(x_i)} \tag{4}$$

$$b0 = mean(y) - b1 * mean(x) \tag{5}$$

- The b1 values and b2 values are further modified by using gradeint desebnt for better b1 and b2 values.

3) Gradient desecnt:: Gradient desecnt is one of the import method to get bo and b1 so that accurate value of response can be calculated.To do that initial values of b1 and b2 are to be calulated.

- Start with b1 and b2 and calculate the cost fuction.The cost function is given by

$$J(b1, b2) = \frac{1}{2m} \sum_{i=1}^{m} (y_{pred} - y_{act})^2 \tag{6}$$

- Here y_{pred} is predicted value using equation 3 and y actual value is actual value at that point.
- now the b1 and b2 values are changed to reduce the cost function in equation 6.To minimize cost function we will take values by using below equations.The b1 and b2 values are adjusted iteratively.
- The equations to calculate b1 and b2 is

$$b0 = b0 - \alpha \frac{\partial J(b0, b1)}{\partial b0} \tag{7}$$

$$b1 = b1 - \alpha \frac{\partial J(b0, b1)}{\partial b1} \tag{8}$$

- The α is learning rate.Here we have taken 0.1 values as learning rate.The learning rate is very important factor which tell how fast the minimum cost function can be reached.If α is too large gradient desent can overshoot,the equations may never converge.if alpha is too small it will take lot of time to reach the best values of b1 and b2.
- substituting equation 6 in equation 7and 8 we will get

$$b0 = b0 - \alpha \frac{1}{m} \sum_{i=1}^{m} (y_{pred} - y_i) \tag{9}$$

$$b1 = b1 - \alpha \frac{1}{m} \sum_{i=1}^{m} (y_{pred} - y_i) * x_i \tag{10}$$

- The function is created and initial b1 and b0 values are passed.
- To minimize cost function,above functions are used iteratively and best b0,b1 values calculated.

4) Plotting he data:: The b0 and b1 values are calculate and linear eqiation us calculated using equation 3. The scatter plot is plotted for all the training values and also the linar equarion is plotted.It will be explained in result section.

5) Accuracy determination:: The accuracy of linear regression can be determined by R^2 method.R-squared (R^2) is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Whereas correlation explains the strength of the relationship between an independent and dependent variable, R-squared explains to what extent the variance of one variable explains the variance of the second variable. So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

$$R^2 = 1 - \frac{unexplainedvariation}{totalvariation} \tag{11}$$

$$unexplainedvariation = \sum_{i=1}^{m} (y^i_{pred} - y^i_{act})^2 \tag{12}$$

$$toatlvariation = \sum_{i=1}^{m} (y^i_{act} - mean(y))^2 \tag{13}$$

C. k-nn classification:

The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm.K-nearest neighbors (KNN) algorithm uses 'feature similarity' to predict the values of new datapoints which further means that the new data point will be assigned a value based on how closely it matches the points in the training set. We can understand its working with the help of following steps.

- The first step of KNN algorithm is loding the dataset,he training as well as test data.
- Next, we need to choose the value of K i.e. the nearest data points. K can be any integer.

$$k = \sqrt{len(training)} \tag{14}$$

- For each point in the test data do the following.
- Calculate the distance between the query example and the current example from the data by suing eucledian equatin

$$dist = \sqrt{(x_{tr} - x_{test})^2 + (y_{tr} - y_{test})^2} \tag{15}$$

- Now, based on the distance value, sort them in ascending order.
- Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
- Pick the first K entries from the sorted collection
- Get the labels of the selected K entries
- since it is classification, return the mode of the K labels

1) Confusion matrix,accuracy, recall and f-score:: Confusion matrix is one such important tool which helps us evaluate our model's performance. As the name suggests it is a matrix of size n x n .where 'n' is the number of class labels in our problem.

In Python's implementation of confusion matrix, rows show actual values and columns indicate predicted values. Given below is the description of each cell.

Actual positives in the data, which have been correctly predicted as positive by our model. Hence it is called True Positive.(TP)

False negitives are actual Negatives in data, but our model has predicted them as Positive.(FN)

False positive is actual Negatives in data, but our model has predicted them as Positive. Hence False Positive.(FP)

Actual Positives in data, but our model has predicted them as Negative. Hence False Negative.(FN)

2) accuracy AND RECALL SCORE:: Accuracy tells the percentage of correctly predicted values out of all the data points. Often times, it may not be the accurate metric for our model performance. Specifically, when our data set is imbalanced. Let's assume I have a data set with 100 points, in which 95 are positive and 5 are negative.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

The recall is the ratio TP / (TP + FN) where tp is the number of true positives and fn the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.

$$recallscore = \frac{TP}{TP + FN} \tag{17}$$

F1 score combines precision and recall relative to a specific positive class -The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst at 0

$$f1_score = 2 * \frac{precesion * recall}{precesion + recall} \tag{18}$$

$$precesion = \frac{TP}{TP + FP} \tag{19}$$

III. Results:

A. Relativefrequency:

The below fighures gives the plot uisng matplotlib and seaborn The mean and median are calculated.

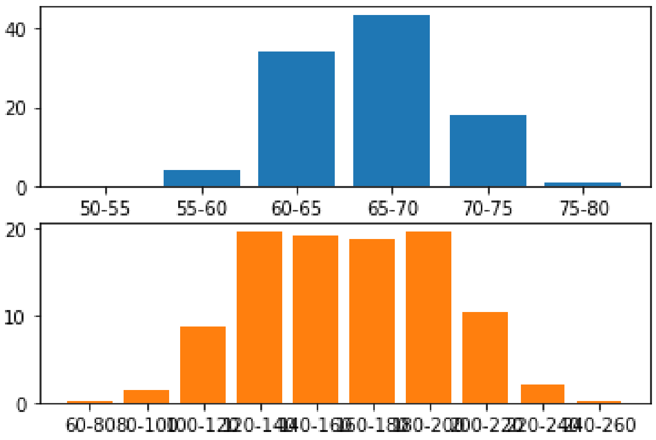


Fig. 1. Relative frequncy using matplotlib

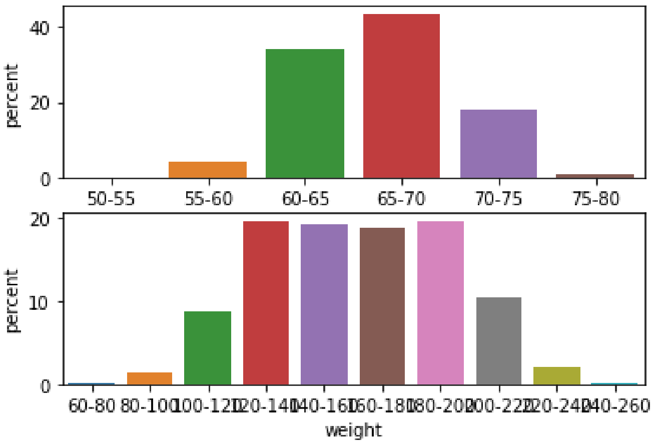


Fig. 2. Relative frequncy using seaborn

- The mean of all weights is 66.36755975482106 and median of all weights is 66.31807008178465
- the mean of all heights is 161.44035683283076 and median of all weights is 161.212927699483
- the mode cannot be calculated because every value is unique value for both heights and weights

Matplotlib is mainly deployed for basic plotting. Visualization using Matplotlib generally consists of bars, pies, lines, scatter plots and so on. Seaborn, on the other hand, provides a variety of visualization patterns. It uses fewer syntax and has easily interesting default themes. It specializes in statistics visualization and is used if one has to summarize data in visualizations and also show the distribution in the data.

B. Plottiing Height and weight and random 70 percent training data and linear regression plots:

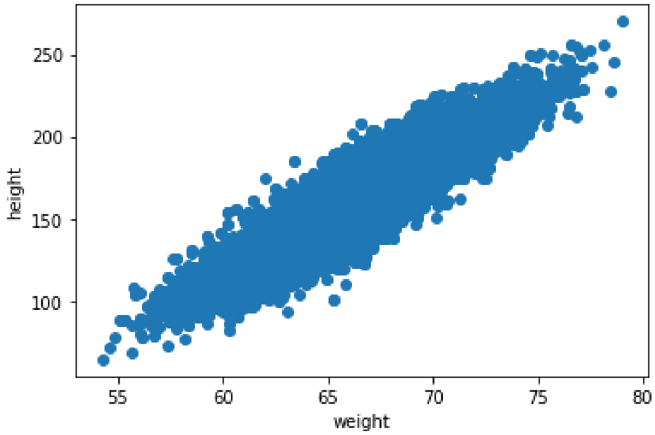


Fig. 3. Relative frequncy using seaborn

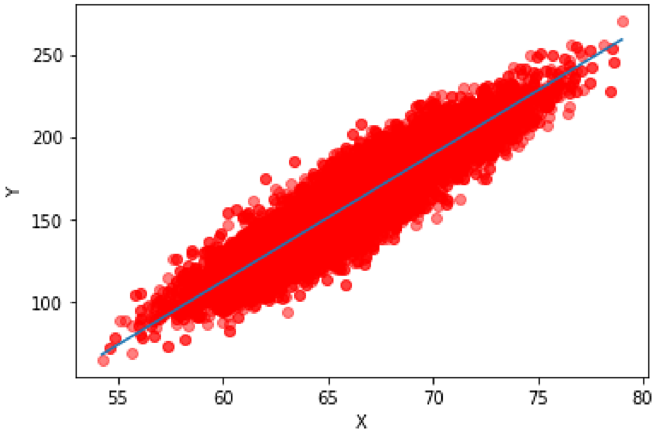


Fig. 4. Relative frequncy using seaborn

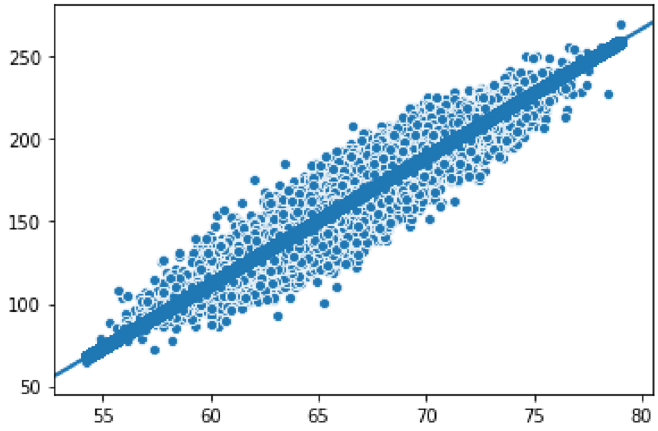


Fig. 5. Relative frequncy using seaborn

C. Explination:

From the above plot the figure 3 is normal plot between height and weiight.it looks like hyperbola.It is plooted using matplotlib package.From the the figure the relationship between height and weight canbe approximated as straight line

- The figure 4 is for linear regression plot using matplotlib package
- The figure 5 is plotted using seaborn package
- Theseaborn package is clearer and more deatiled when compared to matplotlib package.

D. K-nn classification,confusion matrix and other parametes:

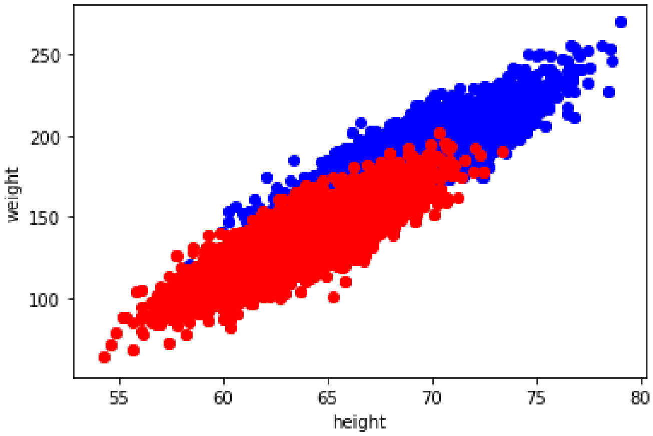
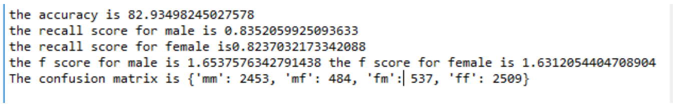


Fig. 6. Relative frequncy using seaborn



```
the accuracy is 82.93498245027578
the recall score for male is 0.8352059925093633
the recall score for female is0.8237032173342088
the f score for male is 1.6537576342791438 the f score for female is 1.6312054404708904
The confusion matrix is {'mm': 2453, 'mf': 484, 'fm':| 537, 'ff': 2509}
```

Fig. 7. Relative frequncy using seaborn

IV. CONCLUSION:

The K-nn calssification and linear regression plots are plotted.The R^2 value is calculated and found to be around 0.85. The confusion matrix,accuracy,recall score and f-score all the values are calculated.The plots are plotted using both seaborn and matplotlib.From coding point of view seaborn is some what tough.once we understand how to plot seaborn then seaborn beconmes easy.