MACHINE LEARNING ASSIGNMENT 2019

1. Description of the dataset:

The skin dataset consist of (B, G, R) values where B represents blue , G represents green and R represents red. Each of the specified values takes on an integer value between 0 and 255 (RGB colour format 8 bit representation). The dataset was collected by randomly sampling B,G,R values from face images of various age groups (young, middle, and old), race groups(white, black, and asian), and genders obtained from FERET database and PAL database.

The main aim of collecting the dataset was to try and classify a given (B, G, R) value from the data as human skin or not human skin. All data points were pre-labelled, where a label of 1 represents the class human skin and a label of 2 represents the class not human skin. The dataset consists of 245057 points, out of which 50859 is the human skin sample and 194198 is the not human skin sample. The attributes of the dataset is the B, G, R values and the class labels.

Some sample of the data points:

B	G	R	Class
74	85	123	1
72	83	121	1
80	86	129	1
172	172	126	2

2. The dataset was split into training data, validation data and testing data. The training data is the randomly picked 60% of the original data (which is 147035 training data), validation data is the randomly picked 20% of the original data (which is 49011 of validation data) and the testing data is the randomly picked 20% of the original data

(which is 49011 of testing data). We normalised the data by using the following, let x be (B, G, R) values from the data. Then each component of x was transformed using

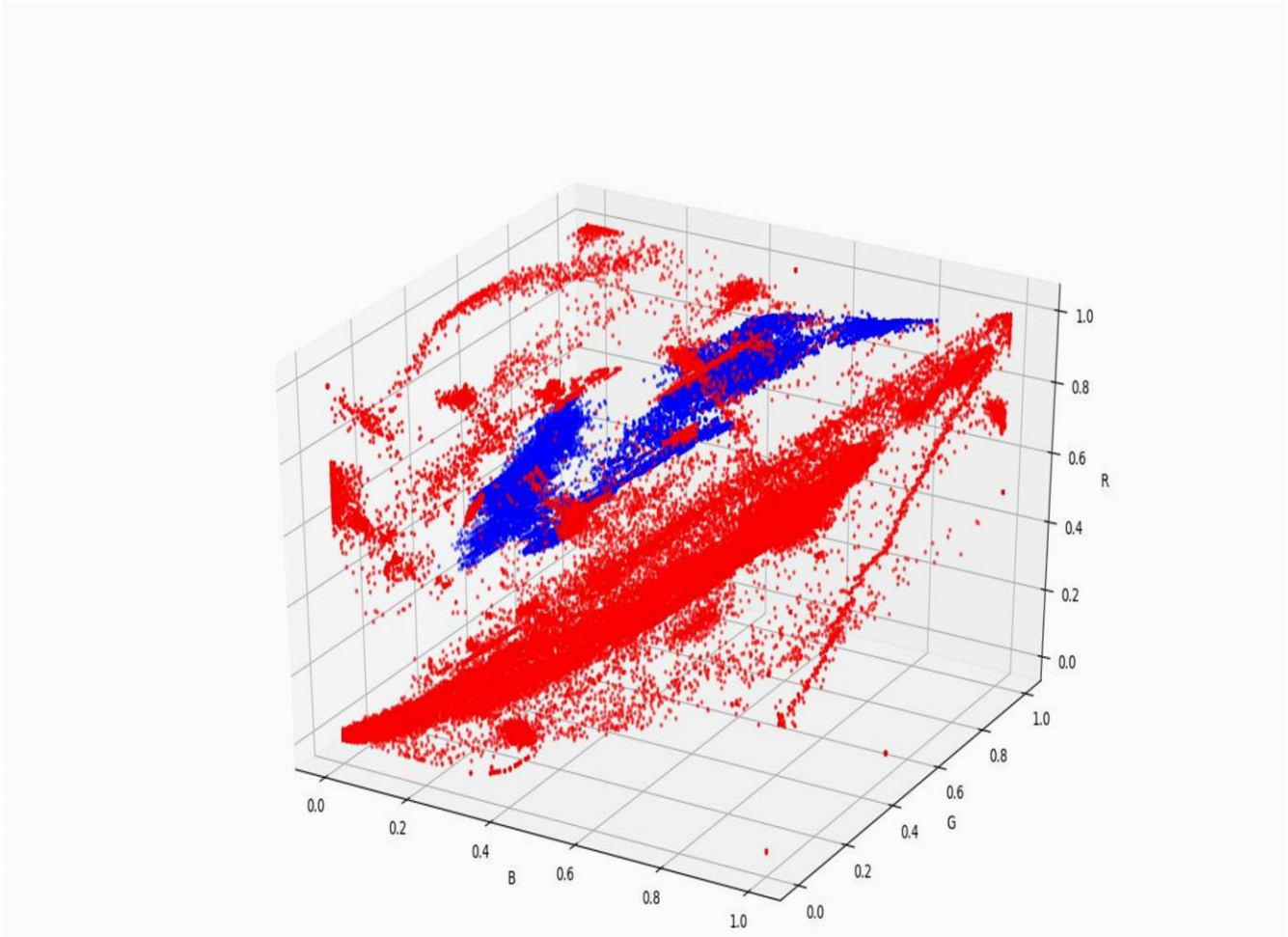
$$x_{(new,i)} = \frac{(x_i - x_{min})}{(x_{max} - x_{min})}$$

where $X_{min} = 0$ and $X_{max} = 255$, this corresponds to multiplying all the input vectors by $\frac{1}{255}$ which means that now each component of the input vector is a float value between 0 and 1. The normalization procedure greatly reduces the spacing between the different values an attribute can take, this means that the attributes can now be considered to a good approximation as continuous.

The following is a 3D graph representing the training data (Where a point is of the form (B,G,R)) :

Key: Red – represent not human skin

Blue – represent human skin



3. In trying to classify the data a naïve bayes classifier and a logistic regression classifier was used.

Naïve Bayes

The naive bayes classifier is often considered when we have labelled data in which we trying to predict discrete classes with probabilistic outputs. The dataset we have satisfies the above criteria. The naive bayes classifier is easy to program, fast to train and it can deal with uncertainty. Although the naive bayes classifier introduces the zero frequency problem, this can be tackled by using smoothing or assigning a probability density function such as a normal distribution or any other distribution suitable for the dataset.

The naïve bayes classifier implementation is as follows

We want to classify a given data value $x = (B, G, R)$ this means that we want the probability of a class c given x , this is done as follows

$$P\big(c\big|(B,G,R)\big) = \frac{P\big((B,G,R)\big|c\big)P(c)}{P((B,G,R))}$$

Where $P((B,G,R)|c) = P(B|c) \times P(G,c) \times P(R|c)$, using the Naïve Bayes Assumption (meaning the specified B, G, R values are assumed to be independent with respect to the class c), $P(c)$ is the probability of class c

And $P((B,G,R)) = P((B,G,R)|c)P(c) + P((B,G,R)|\bar{c})$

For obtaining $P(B|c)$, we assumed that the probability density function of the attribute B with respect to class c is the normal distribution which is given as

$$P(B|c) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(B-\mu)^2}{2\sigma^2}}$$

Where μ : is the mean value of the attribute B with respect to class c.

σ : is the standard deviation of attribute B with respect to class c.

Similarly the same assumption can be made for $P(G,c)$ and $P(R,c)$. This results in having six Gaussian distributions of the form

$$P\big(x_i\big|c_j\big) = \frac{1}{\sqrt{2\pi\sigma_{i,j}^2}} e^{-\frac{\left(x_i-\mu_{i,j}\right)^2}{2\sigma_{i,j}^2}} \quad , i = B,G,R \ , j = 1,2$$

The naïve bayes confusion matrix for testing data:

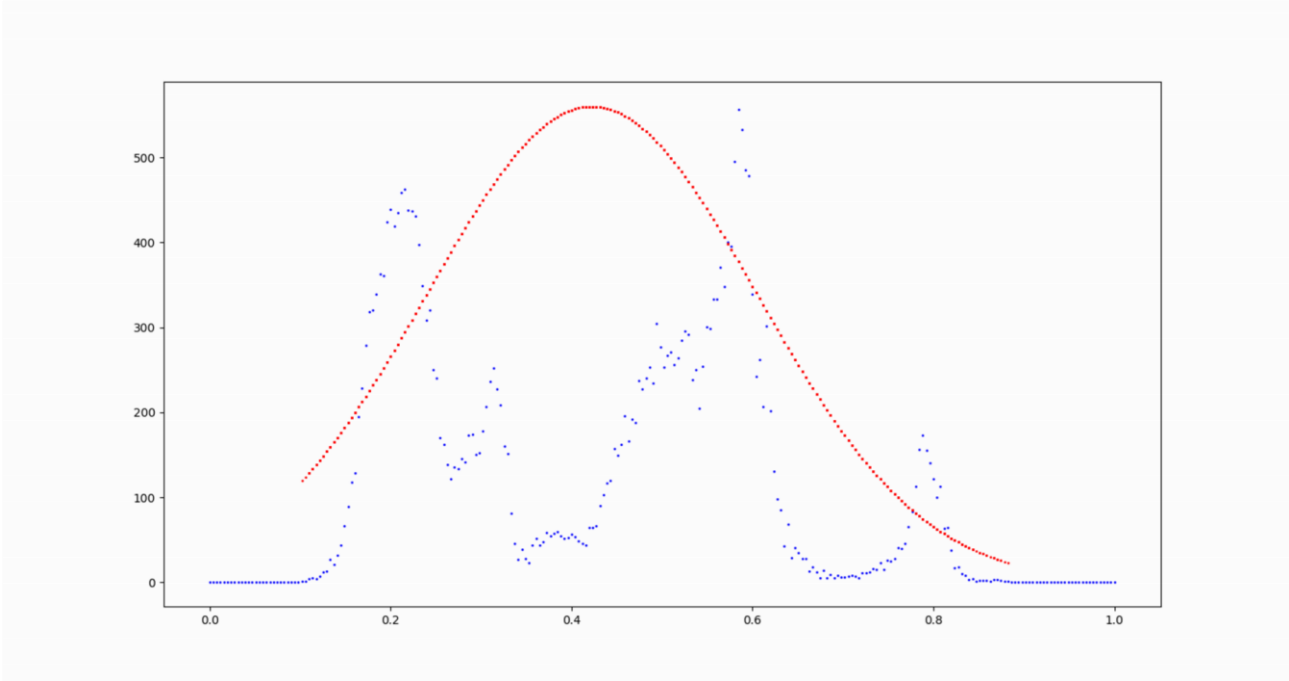
		Actual	
		Human skin	Not human skin
Predicted Class	Human skin	11262	1923
	Not human skin	389	35436

From the confusion matrix percentage accuracy of the classifier is

$$Accurracy = \frac{\sum diagonals}{\sum entries} \times 100$$

$$Accurracy = \frac{11635 + 34900}{49010} = 94.95\%$$

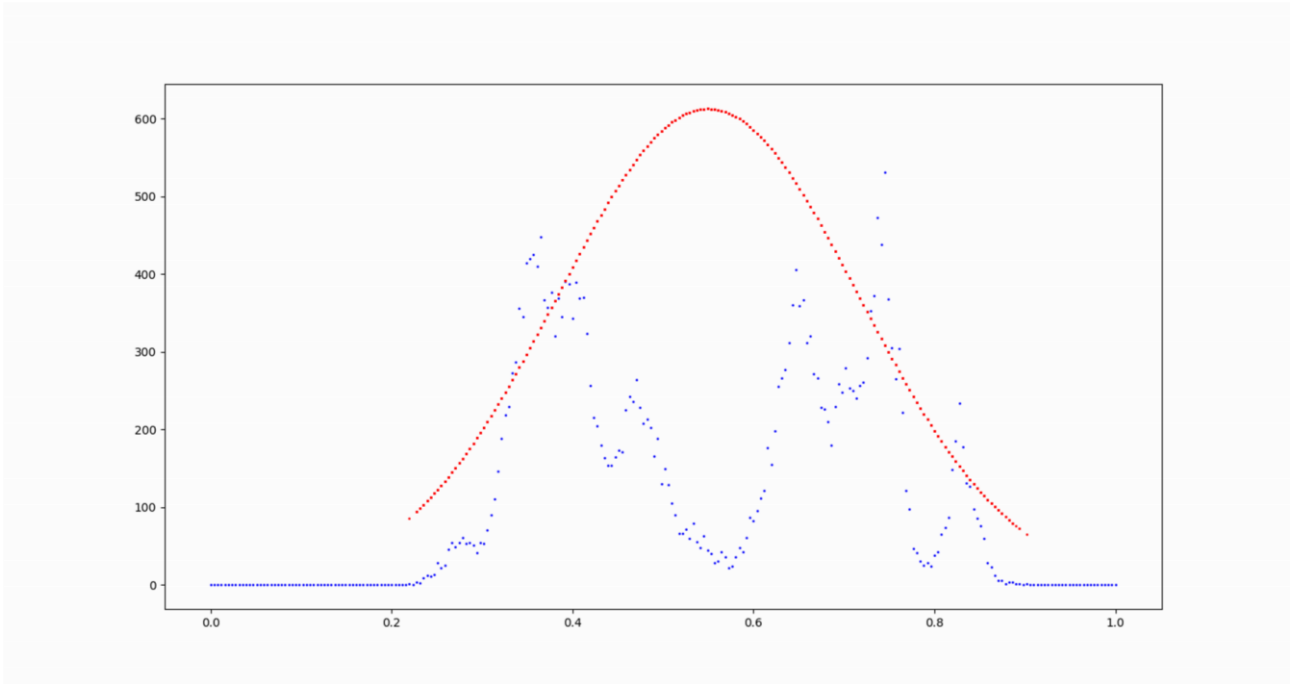
A graph representing actual distribution of feature B on class 1 and the assumed normal distribution of feature B on class 1



Key: Blue – represents the actual distribution
Red – represents the assumed normal distribution

Where x –axis represent the different values that feature B takes
y –axis represent the frequency

A graph representing actual distribution of feature G on class 1 and the assumed normal distribution of feature G on class 1



Key: Blue – represents the actual distribution
Red – represents the assumed normal distribution

Where x –axis represent the different values that feature B takes
y –axis represent the frequency

Logistic Regression:

Logistic regression is considered a good classifier when we want to predict an outcome variable that is categorical form, given predictor variables that are continuous and categorical.

The logistic regression classifier was implemented as follows. We consider the logistic function

$$h_{\theta(x)} = \frac{1}{1 + e^{(-\theta^T)x}}$$

Where $\theta = (\theta_1, \theta_2, \theta_3)$ and x is our vector of features. Since we have a three dimensional problem the decision boundary is a two dimensional hyper plane. Therefore we want the equation of the plane that separate the data points into the two classes. $\theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 = 0$

$$x_3 = -\frac{\theta_1 x_1 + \theta_2 x_2}{\theta_3}$$

The vector of parameters is initially guessed to be the vector

$\theta = (0.4,0,0)$

Then the gradient descent with regularisation was used to update the values of theta with a total number of 5000 iterations .The values used for the learning rate and strength of regularisation are tabulated below with the corresponding vector of parameters

α	λ	θ
0.01	0.001	(3.722,4.67,-8.00)
0.1	0.001	(-47.06,168.42,-204.56)
0.1	0	(-178.13,302.37,-129.22)
0.01	0	(6.26,6.96,-11.46)

The Logistic Regression confusion matrix for data:

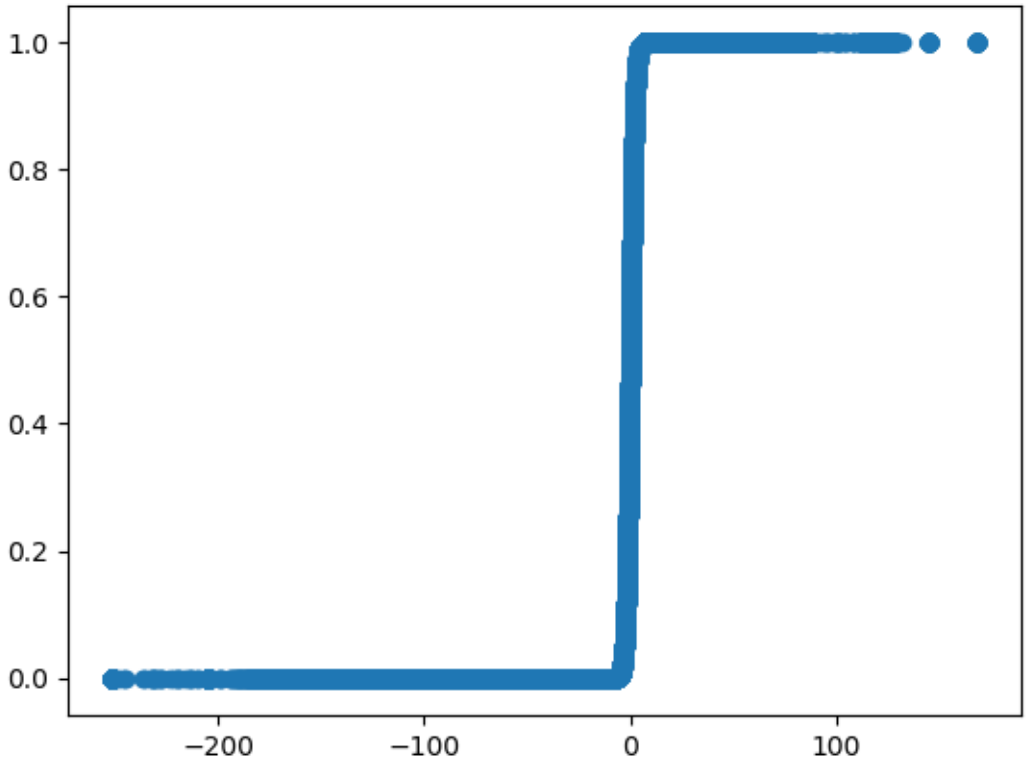
		Actual	
		Human skin	Not human skin
Predicted class	Human skin	11 651	4 459
	Not human skin	0	32 900

$Accuracy = \frac{11\,651+32\,900}{49\,010}$ = 90.90%

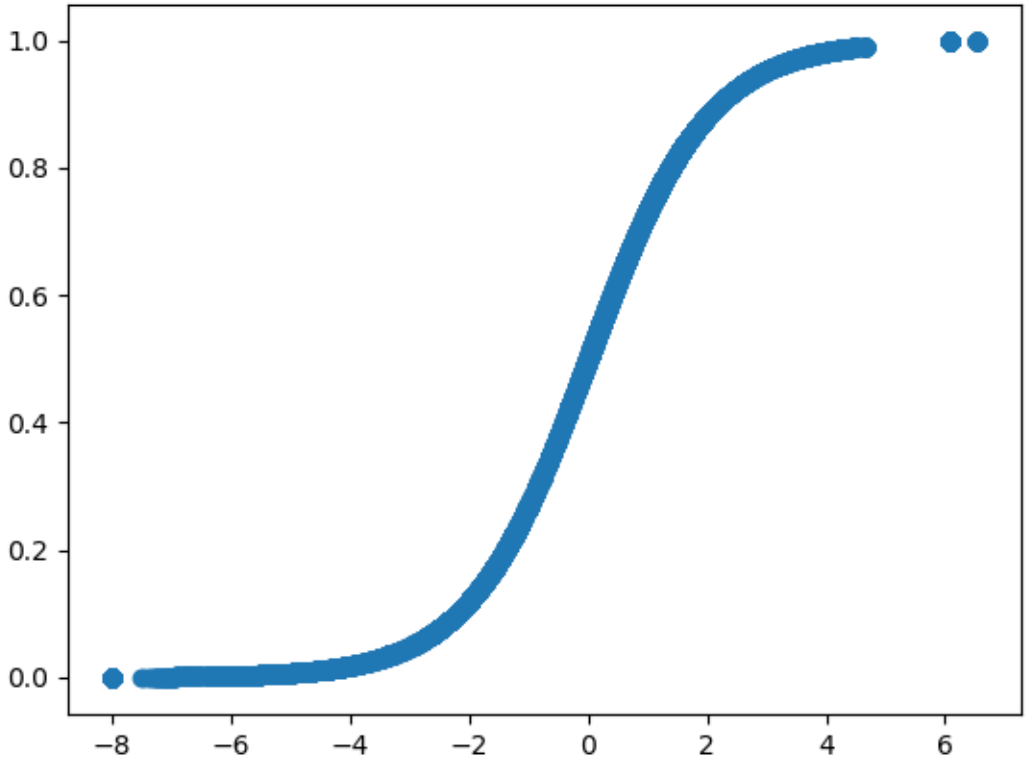
α	λ	$Performance$	θ
0.01	0.001	90.72	(3.722,4.67,-8.00)
0.1	0.001	90.45	(-47.06,168.42,-204.56)
0.1	0	85.80	(-178.13,302.37,-129.22)
0.01	0	90.90	(6.26,6.96,-11.46)

Therefore the peak performance is achieved when $\alpha = 0.01$ and when $\lambda = 0$.
When the weights are relatively high we expect to get a steep Sigmoid function.

Attached below is a plot representing a sigmoid function for $\alpha = 0.1\ and\ \lambda = 0.001$



And below is a plot representing a sigmoid function for $\alpha = 0.01\ and\ \lambda = 0.01$



4.

From the accuracy obtained by the two classifiers the naïve bayes classifier achieved an accuracy of 94.95% whereas the logistic regression classifier achieved an accuracy of 90.90%. It is clear that the naïve bayes model works better than the logistic model. Since the naïve bayes model works quite well, it means that the naïve bayes assumption that the attributes B, G, R are independent with respect to a certain class c, is a good assumption(it is not too naïve). Also the assumed normal distribution for each feature was also a fairly good distribution to choose. This can be seen from the plots of the actual distributions of the features B , G , R the assumed normal distribution for the features, although actual distributions are multi peaked, the normal distribution is a good assumption.

A key factor that affected the performance of the logistic regression model is the bias in the dataset. The dataset consist of 245057 data points of which only 20% are human skin and 80 % are not human skin. This means that when using a subset of the data that is not chosen correctly it is most likely to consist of high number of data points that are not human skin. This means that when using the gradient descent method to determine parameters, the parameters would be mostly dependent on the dominant data points which are the non-human skin data points. The logistic regression model performance is also dependant on the parameters α (the learning rate) and λ(strength of regularisation) which are adjusted to have

maximum performance. The particular model that we have had peak performance when $\alpha = 0.1$,$\lambda = 0.0001$ typically values for $0 \leq \alpha \leq 0.3$ work well and produce at least a performance of 80% whereas if $\alpha > 0.3$ the performance drops below 80% note these values were with $\lambda = 0.1$ .Other combinations of α, λ might have been overlooked and can possibly achieve better performance.

We would recommend that someone who wishes to work with this data should be very caution in choosing how to split the data as the 1:4 ratio of human skin to not human skin makes subsets of the data quite biased and models perform better in classifying data points that are not human skin as compared to data points which is human skin.

5.References

1. http://www-hsc.usc.edu/~eckel/biostat2/slides/lecture13.pdf
2. https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/
3. https://ml-cheatsheet.readthedocs.io/en/latest/gradient_descent.html
4. https://machinelearningmastery.com/naive-bayes-for-machine-learning/
5. https://towardsdatascience.com/regularization-in-machine-learning-76441ddcf99a