Assignment 2

Arvind Pawar

June 3, 2019

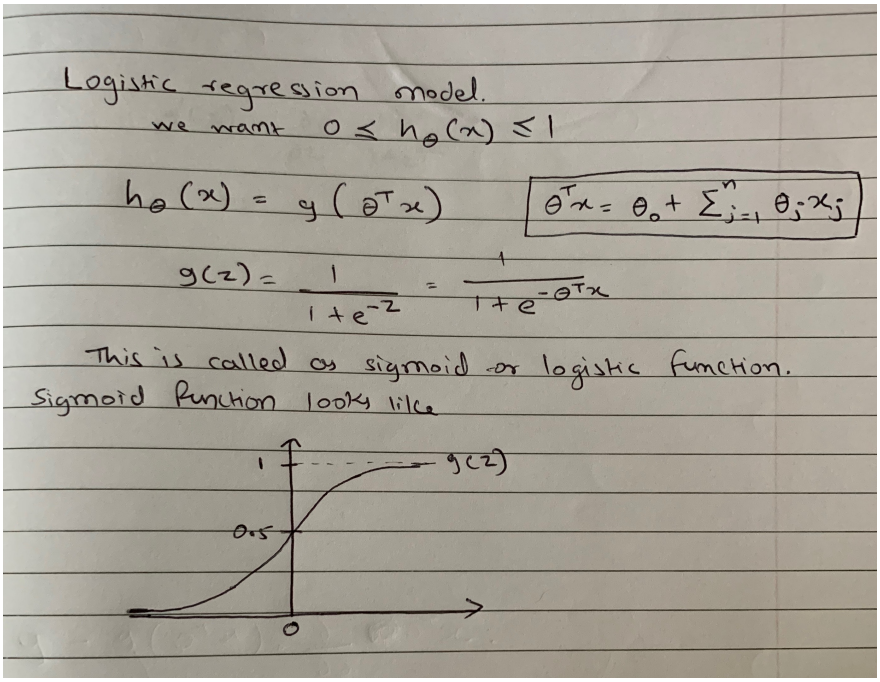
ALY6020-Predictive Analytics

Instructor: Dr. Marco Montes de Oca

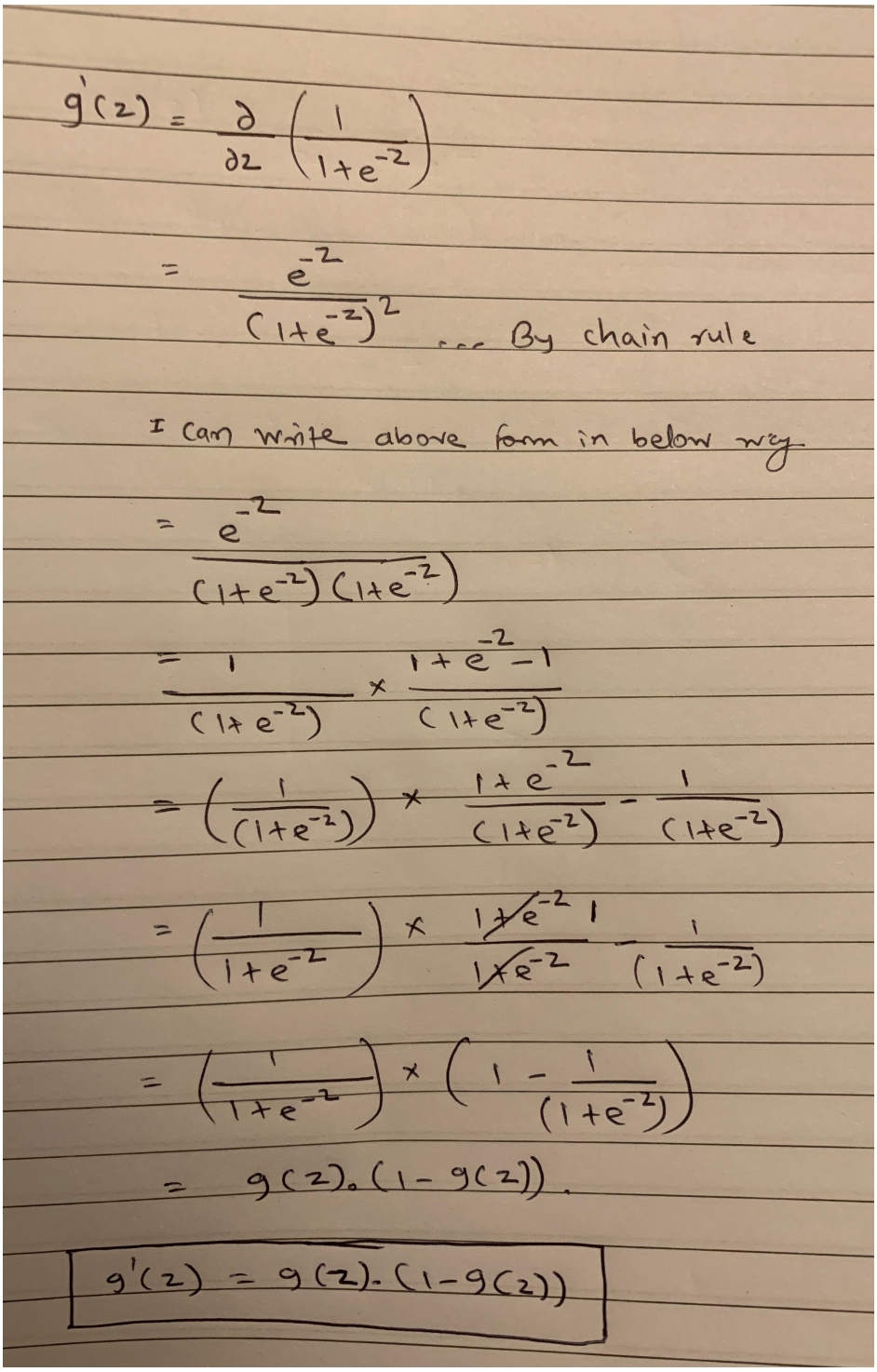
Image recognition with Logistic Regression

Image recognition is the ability of a system to identify objects, person and action in images. So, how does image recognition work? Digital images represent a matrix of numerical values which are data associated with the pixel of the image. A matrix format contains the intensity of different pixels, which averages to a single value. The intensities and the location of different pixels of the image are the input to the recognition systems, using information, a system learns to map a relation and patterns in the images. There are many machine learning algorithms can be used for image recognition, however all of them requires proper features for doing the classification. After choosing an appropriate model we can deploy it on train dataset and then system performance is validated on test data. In this assignment, we will be performing image recognition using logistic regression.

Logistic regression is one of the popular algorithms to solve a classification problem. We can identify as a classification problem when independent features are continuous, and response feature is in categorical form like true/false, yes/no, and so on. Logistic regression is a classification method that uses the logit function. It calculates the probabilities of independent variables as output values that range between 0 and 1 and based on the probabilities and threshold value; it categorizes the input. One more important fact we need to understand why linear regression is not suitable for classification problems. Linear regression gives us a straight line, and this can give us worse hypothesis, output values less than 0 and more than 1 even if our label (dependent variable) has values 0 and 1 only. Also, a single outlier disturbs the whole linear regression predictions. Therefore, we use logistic regression for classification problems. The logistic function is shown below.



When we pass more than one feature, a hypothesis estimates the probabilities y (dependent variable) is equal to 1 or 0 given X (Independent Variables) and parameterized theta. Y=1 if h??(x)>=0.5 and Y=0 if h??(x)<0.5. Let us understand the derivation of logistic function.



We have given training and testing datasets. Training dataset has 60,000 data that contains information about labels and pixels. In this assignment, we have asked to train ten random samples of size 20000 for each clothing item using logistic regression. I have written a function that will create a random sample of size 20000 and will store it in a data frame. In the same function, I have written code for relabeling the

column label of a generated sample. Later, I have called this function in "for loop" which will first generate random sample and then perform relabeling. For example, in the first iteration of for loop will call the function which will generate the 1st random sample of 20000 and will store in a variable. As I have passed the 'df$label==digit' condition, it will only relabel the selected digit(label). In this way, it will generate a total of 10 random relabeled sample for training purpose. In the same "for loop," I have applied logistic and predict function for training models and prediction respectively. After prediction, I have created a confusion matrix using observed values and predicted values.The confusion matrix is shown below.

	0	1	2	3	4	5	6	7	8	9
0	737	14	27	71	17	18	148	5	17	0
1	24	949	12	30	14	16	18	5	28	12
2	65	5	791	44	282	11	182	2	30	7
3	51	18	6	726	41	10	45	2	24	7
4	3	2	16	2	288	5	32	0	19	1
5	8	2	2	7	3	822	5	50	25	34
6	91	6	137	107	344	54	538	26	66	49
7	0	1	2	5	2	49	2	900	20	94
8	17	3	4	5	9	10	27	0	768	5
9	4	0	3	3	0	5	3	10	3	790

Analysis of obtained results:

- From the confusion matrix, we can infer that the accuracy for label 1 label 7 is more as compared to otherabels by correctly classifying 949 and 900 clothing items.
- Misclassification is more in labels 0, 2 and 6.
- Also, label 7 has no misclassification with label 4 and 8 at all and almost negligible misclassification with other labels. We can say that features of label 4 and 8 are entirely different from the same of label 7.
- We obtain the accuracy of 73.09%

Conclusion: We conclude that label 0, 2 and 6 are not well trained to distinguish clothing items accurately. There can be different reasons for misclassification. One of them is that I think different label clothing items have similar pixel weight because of that model is unable to categorize item with an accurate label. Also, we are training models with a sample of 20000; if we train the model with a bit large sample data, we can obtain more accurate results. Even if we face misclassification, we can use regularization to penalize the weights of pixel to increase the accuracy.