IT4060 – Machine Learning

Lab 5 – Logistic Regression

Ex 1: Octave Exercise for Logistic Regression

1. Open the lab5 directory. There you will find some octave scripts.

   ex2.m - Code to run logistic regression. This calls the appropriate Octave functions in other files to do logistic regression on the ex2data1.txt dataset. It contains student admission details based on the marks of two exams the student has taken. So there are two input features and you have to predict the admission status, which is a binary value. So, it's a binary classification problem.

2. Open the ex2.m file. You should get some errors. Now we'll try to modify the code to do logistic regression.

3. Add the following code under the "Your code here" section in the costFunction.m script. You can (ignore the grad calculation, which is just computing the gradient value of the cost function for the initial theta values).

```
temp1 = -1 * (y .* log(sigmoid(X * theta)));

temp2 = (1 - y) .* log(1 - sigmoid(X * theta));


J = sum(temp1 - temp2) / m;
```

4. Add the following code to the sigmoid function in sigmoid.m script.

```
denominator = 1 + exp(-1 * z);

g = 1 ./ denominator;
```

5. Add the following code to the predict.m script to predict the admission status of an unknown student.

```
result = sigmoid(X * theta);

p = round(result);
```

6. Now run ex2.m to see how the training is done and to see the prediction on the unknown student. Refer the lecture note and see whether you can understand the code.

Ex 2: Logistic Regression with regularization

7. The ex2_reg.m file runs a linear regression algorithm with regularization. It takes the input test data of two tests done on a microchip and tries to predict whether it's faulty or not. The mapFeature.m script makes the input feature set a polynomial feature set.

8. Open the costFunctionReg.m and add the following code to define the new cost function with the regularization parameter.

```
temp1 = -1 * (y .* log(sigmoid(X * theta)));

temp2 = (1 - y) .* log(1 - sigmoid(X * theta));


thetaT = theta;

thetaT(1) = 0;

correction = sum(thetaT .^ 2) * (lambda / (2 * m));


J = sum(temp1 - temp2) / m + correction;
```

9. The value of Lambda is currently set to 1 in ex2_reg.m change it to different values to see that the training accuracy changes. Note that even though the training accuracy may be high when Lambda is lower, it may lead to overfitting. How do you explain the changes of training accuracy when Lambda is increased and decreased?

Ex 3: Logistic Regression for Breast Cancer Prediction – Regularization

This example uses the following Breast Cancer dataset where it contains data of tumors along with whether a tumor is malignant (cancerous) or benign.

https://www.kaggle.com/code/dhainjeamita/breast-cancer-dataset-classification

10. Upload the attached logrbrcancer.ipynb notebook to Jupyter notebook and run it.

11. Currently the code that runs the Logistic Regression model multiple times with varying C values is commented. Uncomment it and run it again. (make sure to correct the indentation when you uncomment. Otherwise it may not compile).

12. What can you observe? What does C represent? Can you explain the behavior of the test and training accuracies when C is varied? What is the optimum value for C, according to this plot? Justify your answer.

Submission:

Upload the modified Octave code and the html files exported by Jupyter notebook, the text file with the answers for the 9th and 12th questions as a single zip file to the courseweb link. The file name should be your registration number.