UNIVERSITY OF VICTORIA

Department of Electrical and Computer Engineering

ECE 403/503 Optimization for Machine Learning

LABORATORY REPORT

Experiment No: 2

Title: Handwritten Digits Recognition Using PCA

Date of Experiment: 10/10/2019

Report Submitted on:10/17/2019

To: TA

Laboratory Group No.:

Name(s): Kurt Elliott & Raj Deepak S. N.

Objectives:

In this experiment, we investigate a technique for multi-category classification based on binary classifications. The technique is then applied to Fisher's 3-class datasets of Iris plants to demonstrate its effectiveness.

Introduction:

The dataset of Iris plants is used in this experiment it was created and published in 1936 by R. A. Fisher. The dataset includes features of 150 Iris plants of 3 species known as Setosa, Versicolor, and Virginica, where each sample Iris is represented by a 4-dimensional feature vector in terms of lengths and widths of the sepal and petal of the flower. Binary classification is used on the Fisher dataset to train and test a model. The training set data will include 120 samples where there are 40 samples of each species of flower. The testing data set will include 30 random samples where there are 10 of each species of flowers.

Results:

The following figures show the MATLAB script used in this lab as well as results with a brief description of how the program is implemented. The program shown in figure 1 was the code used in this lab. It consists of three primary sections. The first section of the code is used to extract the training and testing dataset for the 3 species of flowers, namely Iris setosa, Iris versicolor and Iris virginica (figure 1). The dataset consists sepal and petals' length and width, which account for the 4 features that are used to categorise the species of the Iris flower.

Figure 1 - Three species of Iris

In the training section, we also assign the y vector and assigning classes such that all data corresponding to Setosa are labeled as the P vector, while the other 2 species are labeled as the N vector. The P and N vectors are reassigned with respective to the other species for the second and third cases. Then the Linear Regression algorithm in which the following weights are computed for the three cases:

$wh_1 = (Xh_1 * Xh_1')/(xy_1)$,

$wh_2 = (Xh_2 * Xh_2')/(xy_2)$,

$wh_3 = (Xh_3 * Xh_3')/(xy_3)$,

Where wh is the weight and Xh is the altered input with respect to P and N vectors in each case.

w_s and b_s are computed from wh so as to complete the linear regression equation.

In the testing section, we use the linear regression equation $y_{te} = w_s X_{te} + b_s$, to compute the mis-classifications that arise while assigning the data to a particular species. A confusion matrix is then computed with the classification data from each species.

The following was the script used to execute the MATLAB code:

Contents


```
clear all; close all; clc;

LOAD = 1;
if LOAD == 1
    load('D_iris_tr');
    load('D_iris_te');
end
```

Training and Testing datasets:

Data sets xtr1, xtr2, and xtr3 are train data from Setosa, Versicolor, and Virginica, respectively, each contains 40 samples. Data sets xte1, xte2, and xte3 are test data from Setosa, Versicolor, and Virginica, respectively, each contains 10 samples

```
%D_iris_tr(4,:) = 1;
%D_iris_te(4,:) = 1;


xtr1 = D_iris_tr(:,(1:40));
xtr2 = D_iris_tr(:,(41:80));
xtr3 = D_iris_tr(:,(81:end));

xte1 = D_iris_te(:,(1:10));
xte2 = D_iris_te(:,(11:20));
xte3 = D_iris_te(:,(21:end));

xte = [xte1 xte2 xte3];

% Creating 3 binary classifications to produce linear models
X = [ xtr1 xtr2 xtr3 ];
Y = [ones(40,1);-ones(80,1)];

for k = 1:3
    switch k
        case 1
            P1 = X(:,1:40);
            N1 = X(:,41:end);
            X1 = [P1 N1];
            Xh1 = [X1 ;ones(1,120)];
        case 2
            P2 = X(:,41:80);
            N2 = [X(:,(1:40)) X(:,(81:end))];
            X2 = [P2 N2];
            Xh2 = [X2;ones(1,120)];
        case 3
            P3 = X(:,81:end);
            N3 = X(:,(1:80));
            X3 = [P3 N3];
```

```
            Xh3 = [X3;ones(1,120)];
    end
end

% Training
xy1 = Xh1*Y;
wh1 = (Xh1*Xh1')\xy1;

xy2 = Xh2*Y;
wh2 = (Xh2*Xh2')\xy2;

xy3 = Xh3*Y;
wh3 = (Xh3*Xh3')\xy3;

wh = [wh1';wh2';wh3'];

ws = wh(:,1:4);
bs = wh(:,5);

% Testing

mis_class = 0;
TestM = [ones(1,10) 2*ones(1,10) 3*ones(1,10)];

e1 = [1 0 0]'; e2 = [0 1 0]'; e3 = [0 0 1]';
E = zeros(3,30);

for i = 1:30
    Ytest = ws*xte(:,i) + bs;
    [~, Idx] = max(Ytest);
    E(Idx,i) = 1;
    if Idx ~= TestM(i)
        mis_class = mis_class + 1;
    end
end

%disp('testing setosa:');
E1 = E(:,1:10); c1 = sum(E1')';
[E1 c1];

%disp('testing versicolor:');
E2 = E(:,11:20); c2 = sum(E2')';
[E2 c2];

%disp('testing virginica:');
E3 = E(:,21:30); c3 = sum(E3')';
[E3 c3];

disp('Confusion matrix:')
C = [c1 c2 c3]

mis_class
```



```
Confusion matrix:
```

```
C =

    10     0     0
     0     9     0
     0     1    10


mis_class =

     1
```

Discussion:

The best classification results was one mislabeled flower which was attained by only using the fourth feature. From the confusion matrix, it has been found that the only flower that was mis-classified was Iris versicolor, which was labeled as Iris virginica. It is clear that there is quite a lot of correlation between the different flower species features, which is why the one dominant feature is essentially the only feature that is required to classify the different flower species.

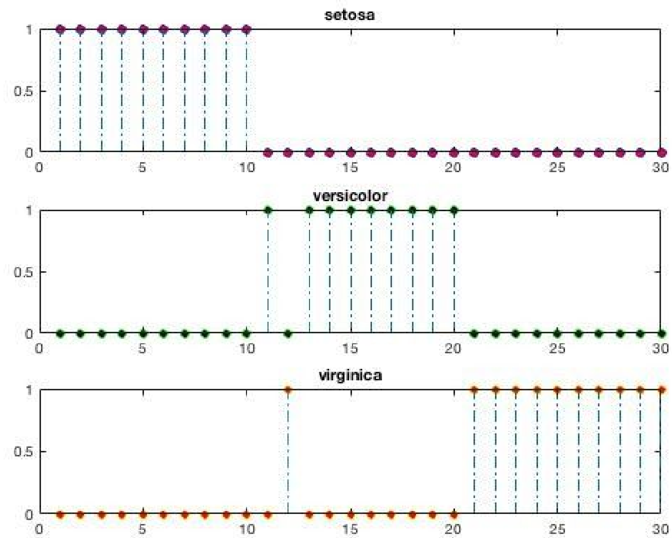

Figure 2 - Three species of Iris labelling

Conclusions:

The binary classification model was successful in classifying the different Iris flower species in the final test dataset with only one mislabelled flower. One of the disadvantages of using this linear model is the fact that it is linear and therefore can only build models with two inputs. For this experiment we used positive and negative values for inputs to distinguish the different flower species. This worked however it would be disadvantageous if working with more than three classes as you need to compare each case and a different algorithm would be better suited.