# Homework Assignment 3
# Assigned on October 13, 2014
# Due on October 20, 2014
# 11:59PM on Black Board
# 20 points

October 13, 2014

## PROBLEM STATEMENT

In this homework you will implement Logistic Regression for TWO class classification of the iris data set. The class label $y \in \{0, 1\}$.

Since the iris data set has actually three classes, you will need to successively relabel one class as class 0, and the other two as class 1.

Thus, you should run three cases as follows:

**Case 1:** setosa vs virginica and versicolor

**Case 2:** virginica vs setosa and versicolor

**Case 3:** versicolor vs setosa and virginica

### TRAINING

Implement Algorithm 8.2, page 254 to train your classifier. You will need to set up a criterion for convergence, or limit on iterations: usually converegence is signaled by very small difference between two successive values of the quantity that it is computed / updated at each iteration.

Here the weight vector $\mathbf{w}$ is updated in each iteration. So you will need to keep track of the previous and current values of $\mathbf{w}$ and use a threshold call it $\epsilon$, which you may want to set to a very small positive value.

Note that in Matlab one can use eps whose value is 2.2204e-016. But that value may be TOO small (in some / many cases). So, the convergence criterion should be combined with a bound on the number of iterations.

## TEST

Basically, for each test point you must compute

$$out = testpoint \cdot \mathbf{w}$$

(the equation above assumes that $\mathbf{w}$ is a column vector, and testpoint is a row vector).

You must convert the values in out to 0 and 1(e.g., replace all negative values by 0 and all positive by 1).

Furthermore, for each of the three classification problems consider training set size varying from $p = 0.2, \ldots, 0.9$ of the entire data set with the remaining data to form the test data.
For each classification display the confusion matrix and the roc plot. I am including below some code and the results of one run.

## DISCUSSION

We know that setosa is (linearly) separable from the each of the other two classes, and that there is overlap between virginica and vrsicolor. In the light of this overlap, it is quite normal that accuracies for these two classes will be lower than for setosa.
Include a discussion in which you discuss the overlap (you need to come ip with some measure of overlap) and dicuss the accuracies taking into accoount the overlap.

## SOME CODE

```
close all
clear all

load('fisheriris');
[r, c]=size(meas);


% Extend meas by 1 to account for the bias
col1 = ones(r,1);
emeas=[col1 meas];
```

```
%Assign numerical labels
for i=1:50, class(i)=1; end  %setosa
for i=51:100, class(i)=2; end  %versicolor
for i = 101:150, class(i) = 3; end %virginica

%Transform numerical lables into 0 / 1 labels
newclass(class==1) = 0; %setosa becomes class 0%

% the rest are in class 1
newclass(class == 2) = 1;
newclass(class == 3) = 1;


p = 0.1; % extent of training sets
randindex=randperm(r);
N = round(p*r)
train = emeas(randindex(1:N),:);
trainlabels = newclass(randindex(1:N));

test = emeas(randindex(N+1:r),:);
testlabels = newclass(randindex(N+1:r));
```

```
N =

    15
```

## Algorithm 8.2 without test for convergence

```
% initialize w
w=zeros(c+1,1);

ybar=mean(trainlabels);

w(1)=log(ybar/(1-ybar));

s=zeros(1,N);
z=zeros(N,1);
```

HERE COMES ALG. 8.2

## Test

```
ltest=length(testlabels);
```

```
% Compute the output for each test data using w
for i=1:ltest,
   out(i)=test(i,:)*w;
end

% Transform output in 0 1 labels
out1=out;
out1(out<0)=0;
out1(out>0)=1;

% compute accuracy
accuracy = 1 - sum(abs(testlabels - out1))/ltest

% plot confusion matrix
plotconfusion(testlabels, out1);

%plot ROC curves
plotroc(testlabels, out1);
```

```
accuracy =

     1
```