Project 1: Linear Discriminants for Breast Cancer Detection

For this project, you will use different Discriminant Analyses for breast cancer detection.

First, read the dataset description at:

http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

For your convenience, I have reformatted the data for the MATLAB. The input matrix is called P, each row of which corresponds to different measurements of a patient's tumor cell sample. The corresponding element in the output vector T is -1 if the cell was determined to be benign, and +1 otherwise.

Load P and T into MATLAB's by using the menu "File > Import Data" option, or by simply dragging and dropping into MATLAB window.

Create a supervised classification dataset by using randomly 70% of the data for "training":

```
[trainP,valP,testP,trainInd,valInd,testInd] = dividerand(P,0.7,0,0.3);
```

This function randomly divides the input data and then returns division indices so that one can arrange the corresponding target data:

```
[trainT,valT,testT] = divideind(T,trainInd,valInd,testInd);
```

Please read MATLAB's help documentation for 'dividerand' and 'divideind' for more information. Furthermore, let's put data points from healthy patients in C1, and C2 otherwise:

```
C1=trainP(:,(find(trainT==-1)));
```

```
C2=trainP(:,(find(trainT==1)));
```

Same for test data:

```
C1t=testP(:,(find(testT==-1)));
```

```
C2t=testP(:,(find(testT==1)));
```

Now train (build) the Fisher LDA: (please see the help or the header comments for the provided 'LDA3' function. Type:

```
[w J m reg]=LDA3(C1,C2);
```

Find the 'confusion matrix' (true positive TP, true negative TN, false positive FP (false alarm), and false negative FN (miss)) for training and testing using zero as the threshold and:

```
w'*x-w' *m
```

where x is C1, C1t, C2, or C2t. Ideally, all C1's should be on the negative side, and vice versa (note: you can use MATLAB's *confusionmat* command for this purpose)

Project 1: Linear Discriminants for Breast Cancer Detection

Task 1 (required for all): For FLDA classifier, produce the (a) ROC curves and (b) TP, FN, FP, and TN, from confusion matrices. Do so both for training and testing detests. How is the generalization capability of the network, given these results for the test data?

Task 2 (required for graduate section, optional with up to 10% bonus points for undergrad section):

Change the ratios for training and testing dataset divisions to 40%-60% respectively, and repeat the above. Describe and explain the changes in your results, if any.

Deliverable (electronic upload on CANVAS): submit the requested items as a deck of slides (starting with a single slide on problem statement, a second slide on the methods, followed by the main section showing ALL the requested results: numbers, tables, ROC curves, and observations; ending with your conclusions). Also submit ALL your MATLAB programs. Compress all submission items (except the dataset) into a single file before uploading via CANVAS (as your_name.zip). Don't forget your name and section (grad/undergrad) on the title slide. Clearly label tasks 1 and 2. Failure to do any of the above steps correctly will reduce your grade proportionally.