Table of Contents

PCA

by Pavana Mysore Ganesh Fill in this template for the assignment. When you are done, publish this script as a PDF file.

```
% In this assignment you will analyze gene expression data available
 at:
% https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE7307
% The data contains samples from different human diseases and from
 different
% tissue types. For each sample, expression values of genes are given.
% For this assignment you are provided only with a subset of the
 diseases.
```

download the data file. Nothing for you to change here.

```
%file = bmes.downloadurl('http://sacan.biomed.drexel.edu/lib/exe/
fetch.php?media=course:ml:data:diseases_subset.xlsx');

[m,n,raw] = xlsread('diseases_subset.xlsx');
```

Load the data

Read the data from the file and convert it into a usable format.

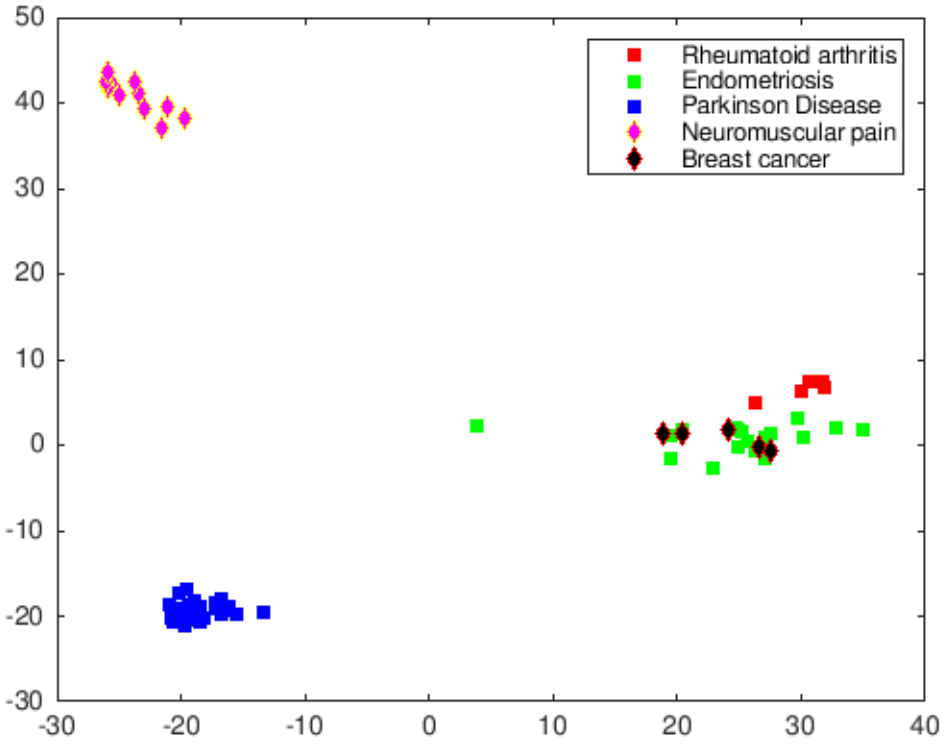
```
raw(:,1) = [];
classname = raw(1,:);
raw(1,:) = [];
raw = cell2mat(raw);
raw = raw.';
% mean shifting
std_raw = raw - repmat(mean(raw),66,1);
```

PCA Visualization

Visualize (in 2D) the provided dataset. Use a scatter-plot where each point represents a sample. Use a different color and marker for each disease. Samples from the same disease should have the same color & marker.

```
[coeff,score,latent] = pca(std_raw);
new = std_raw*coeff(:,1:2);
plot(new(1:5,1),new(1:5,2),'sr','MarkerFaceColor',[1 0 0]);
hold on
plot(new(6:24,1),new(6:24,2),'sg','MarkerFaceColor',[0 1 0]);
```

```
hold on
plot(new(25:50,1),new(25:50,2),'sb','MarkerFaceColor',[0 0 1]);
hold on
plot(new(51:61,1),new(51:61,2),'dy','MarkerFaceColor',[1 0 1]);
hold on
plot(new(62:66,1),new(62:66,2),'dr','MarkerFaceColor',[0 0 0]);
hold off
legend('Rheumatoid arthritis','Endometriosis','Parkinson
 Disease','Neuromuscular pain','Breast cancer');
```



Interpret

Answer these questions below each question, using comments and/or any supporting programming code.

Do the samples from the same diseases cluster together, i.e., do they have a similar gene expression values?

```
 disp('yes, Breast cancer and Endometriosis are clustering together');
```

```
yes, Breast cancer and Endometriosis are clustering together
```

Samples from which disease are the most tightly clustered in the 2D PCA plot?

```
disp('Samples of Neuromuscular pain,Parkinson Disease, Rheumatid
 arthritis are tightly clustered with their samples');
```

```
Samples of Neuromuscular pain,Parkinson Disease, Rheumatid arthritis
 are tightly clustered with their samples
```

Samples from which disease are the least clustered together in the 2D PCA plot?disp

```
disp('Samples from Endometriosis are least clustered together');
```

```
Samples from Endometriosis are least clustered together
```

How many dimensions should one keep (and not necessarily visualize) in order to capture a total of 70% of the variance in the original dataset ?

```
disp('To get a total of 70% of the variance in the original dataset 3
 dimensions as to be captured');
```

```
To get a total of 70% of the variance in the original dataset 3
 dimensions as to be captured
```

What is the reconstruction error if we used the first 3 principal components to represent the dataset ? Remember to compare the reconstructed data, not with the "original data", but with the "mean-shifted original data".

```
pm = score(:,1:3)*coeff(:,1:3)';
err = sqrt(mean(sum((std_raw-pm).^2,2)));
fprintf('Reconstruction error using %d components: %f \n',3,err);
```

```
Reconstruction error using 3 components: 20.996071
```

Published with MATLAB® R2018b