PCA

Sunday, April 30, 2017 8:01 PM

Logistics

- You can work on this Lab-homework in groups of two or on your own. No teams of more than 3 students allowed.
- Step by step instructions are provided throughout the description of this assignment. Some code is included for your convenience. Additional action items or requirements you need to address are noted in **bold** font.
- Turn in report and code all in a zip file
  - Report should be a single document in docx or pdf format
  - Turn in all your code as .m files

PCA Practice

In this exercise we will use Matlab/Octave to explore PCA. We will use the Iris dataset used in Homework 2

Iris dataset: contains information about three varieties of flowers. The features (variables) used to identify flowers are petal and sepal width, and petal and sepal length.

Prepare the Data

Load the information from the iris dataset into a single N x k matrix (where N=150 is the number of observations, and k=5 is the number of features)

Each row is an observation or (feature vector describing a specific observation). Each column will be a feature, the first column representing a flower type, for example:

Type	PW	PL	SW	SL
0	2	14	33	50
1	24	56	31	67
1	23	51	31	69
0	2	10	36	46
1	20	52	30	65

The feature vector for the first row is [0 2 14 33 50]. Note that the first column is just a tag indicating the type of flower, so we won't use it for PCA

Follow the instructions in Homework 2 to randomize the order of your data. In this case, you want to keep ALL the rows and columns 1 through 5. Store the resulting dataset in a matrix named `ctotal`. We really won't use the first column but let's carry it with us for now.

Let's start preprocessing the data for PCA…

Center and scale each column

```
% Get the mean and stdv of the new total set
means = mean(ctotal);
stdv = std(ctotal);

%
% Mean-center/scale each feature, X is NORMALIZED & MEAN CENTERED  dataset
%
for i=1:nfeatures
    for j=1:nsamples
        X(j,i) = ( means(:,i) - ctotal(j,i))/stdv(:,i);
    end
end
```

Note: do not normalize the tag column! It has no statistical relevance (keep it in its original form as we will need it later.)

Interesting references

Read more about preprocessing data:
http://scikit-learn.org/stable/modules/preprocessing.html

Read more about PCA:
http://setosa.io/ev/principal-component-analysis/
http://www.dsc.ufcg.edu.br/~hmg/disciplinas/posgraduacao/rn-copin-2014.3/material/SignalProcPCA.pdf

Singular Value Decomposition

Do SVD, do not use the built-in PCA function, use the SVD function:

```
%
% X is the original dataset
% Ur will be the transformed dataset
%
[U S V] = svd(X,0);
Ur = U*S;

% Number of features to use
f_to_use = nfeatures;
feature_vector = 1:f_to_use;

r = Ur;
```

Next, compute mean and covariance for each class in the transformed space, i.e., using the transformed dataset. You will have to rearrange the `Ur` matrix based on the class tag.

Include all these results in your report. Discuss meaning and relevance.

Scree Plots

These plots make use of the information contained in the matrix `S`. Recall this is a diagonal matrix, so if we want to normalize it, we need the diagonal values add up to 1. A simple way to do this is to can divide each diagonal element by the sum of the elements along the diagonal. However, the main purpose of the normalization is to ~~generate~~ the scree plots. Let's use the following approach based on the formulas presented in the slides; this provides equivalent `S` normalizationresults and we can use it for Scree plots

```
%
% Obtain the necessary information for Scree Plots
% Obtain S^2 (and can also use to normalize S)
%
S2 = S^2;
weights2 = zeros(nfeatures,1);
sumS2 = sum(sum(S2));
weightsum2 = 0;

for i=1:nfeatures
    weights2(i) = S2(i,i)/sumS2;
    weightsum2 = weightsum2 + weights2(i);
    weight_c2(i) = weightsum2;
end
```

Plotting Scree Plots

```
figure;
plot(weights2,'x:b');
grid;
title('Scree Plot');

figure;
plot(weight_c2,'x:r');
grid;
title('Scree Plot Cummulative');
```

Generate the plots and discuss your findings. Include your plots in your report.

Loading Vectors

These come from the matrix V. Remember this is the new orthogonal basis, and also, it povides information about the effect (weights) that the original features have in the new dimensional space. Each row of V-transpose is a dimension in the new space. Each dimesion is a vector whose coefficients indicate the "weight" of the original feature in this new dimension, e.g., how much each original feature contributes to form this new dimension. We will use bar chart plots to explore each of these vectors. We will generate one bar chart for each of these row vectors, each bar is the coefficient corresponding to each dimension

We'll square all the values in V but keeping the sign. Note that this code is using V, **not** V-transpose, so instead of looking rows, the code works on each column

```
for i=1:nfeatures
    for j=1:nfeatures
        Vsquare(i,j) = V(i,j)^2;
        if V(i,j)<0
            Vsquare(i,j) = Vsquare(i,j)*-1;
        else
            Vsquare(i,j) = Vsquare(i,j)*1;
        end
    end
end
```

The following code plots a bar chart of the first loading vector (it is a column vector because we didn't transpose V or Vsquare (if you decide to transpose, then you will use Vsquare(1, :) instead).

```
    figure;
    bar(Vsquare(:,1),0.5);
    grid;
    ymin = min(Vsquare(:,1)) + (min(Vsquare(:,1))/10);
    ymax = max(Vsquare(:,1)) + (max(Vsquare(:,1))/10);
    axis([0 nfeatures ymin ymax]);
    xlabel('Feature index');
    ylabel('Importance of feature');
    [chart_title, ERRMSG] = sprintf('Loading Vector %d',1);
    title(chart_title);
```

Create a plot for each loading vector (there are four in the iris dataset). Include and discuss the plots in your report. Describe and discuss feature behaviour. Describe the correlations and discuss their relevance in the context of this problem.

Scatter Plots

`Ur` provides the data points for all the observations in the new dimensional space. Generate scatter plots using these transformed observations and compare them to the plots using the original observations.

2D

- Using Ur, generate scatter pltos like the ones you did in Homework 2 Part 0. Note that in these plots, the dimensions are no longer "Sepal Length" etc., instead they are the principal coponents, i.e., PC1, PC2, PC3, PC4
- Use the scatter plots you generated in Homework 2 Part 0 and compare them to the plot.
- Discuss your findings: do the new dimensions cluster the data any better? Why or why not?
- What happens if you plot U instead of Ur?

3D

- Similar to the 2D plots, generate 3D scatter plots of the original data points and the transformed data points.
- Discuss your findings: do the new dimensions cluster the data any better? Why or why not?
- Same as above, what happens if you plot U instead of Ur?

Example code to generate 3D scatter plots:

The class variables correspond to the observations in Ur corresponding to each of the flower types. The variables x,y,z indicate which principal components (dimensions) you want to use to plot these observations, e.g., PC1, PC2, PC3; in other words, x, y, z are the *indexes* of the columns of Ur that you want to plot on.

```
figure;
scatter3(class1(:,x), class1(:,y), class1(:,z), 'r.');
hold on;
scatter3(class2(:,x), class2(:,y), class2(:,z), 'b.');
scatter3(class3(:,x), class3(:,y), class3(:,z), 'g^');
```

Rubric

Criteria	Points	Notes
Code	2p - 0p	Code should be well commented and structured Must include all the following implementations: • Center and scale • Svd • Scree plots • Loading Vectors • Scatter: before and after transformation NOT: Must NOT use Matlab's pre-built PCA function
Analysis	Outstanding: 8pt Sufficient: 6pt Poor: 4pt Missing: 0pt	Include and discuss statistics and plots. The discussion must include interpretations for each of the following: Scree Plots: e.g., discussion of how many components are useful Loading Vectors: should have four plots (one per vector), discussion about relevant /irrelevant features, correlations, etc. Scatter: 2D plots in summary form, for data before and after svd Discuss your results, putting them in context of the results (plots and statistics) Focus on the WHYs not just the What's. Remember, your interpretation of the results is what matters the most! Points will be deducted for lousy redaction, spelling/grammar mistakes, typos -- professionalism is expected.