SEMINAR 2
PATTERN RECOGNITION

Theory

1. Imagine that you are in a D dimensional space and using PCA of LDA you reduce the dimensionality to D'<<D determined by a basis $\{v_1, v_2, \ldots, v_{D'-1}, v_{D'},\}$:

1) How many dimensions every vector v_i will have?

2) How many dimensions does the original data have? And the projected data?

3) Once we have the data projected in the D' dimensional space, if we reproject it to the original one, will this reprojected data be the same as the original data? Why?

4) And if D'=D?

5) Which characteristics should have a dataset to guarantee that the original data and the reprojected one are the same?

Practical exercises

1. SVM vs LDA

Load the datasets "ionosphere", and compare the performance of LDA and Support Vector machines on the dataset. To do so, you will have to divide the dataset in a training set and a test set, train your models with the training set, and apply them to the test set to compute the accuracy of each model.

Load the data:

```
load ionosphere
```

Separate Training and test data:

```
p = .7     % proportion of rows to select for training
N = size(X,1)  % total number of rows
tf = false(N,1)    % create logical index vector
tf(1:round(p*N)) = true
tf = tf(randperm(N))   % randomise order
Xtrain = X(tf,:)
Ytrain = y(tf,:)
Xtest = X(~tf,:)
Ytest = y(~tf,:)
```

2. PLS vs PCR

Imagine that you are hired for an aeroline to find an easy and cheap way to determine the octane ratio of the gasoline (which determines the gasoline quality) without testing it in real planes. They give you 60 samples of gasoline at 401 wavelengths, and their octane ratings.

a) Build three models using Multivariate Linear Regression, PLS and PCR to predict the octane ratio of the gasoline given its wavelengths.

b) Make a plot of the fitted vs. observed octane ratios for your training and test samples.

c) Validate your models through cross validation using the following error measure:

$$err = \frac{1}{n} \sum_{i=0} \frac{|y_i^{pred} - y_i|}{y_i}$$

d) After that build again a PLS and a PCR but using 10 components in both cases and make a plot of the variance as a function of the number of components (from 1 to 10).

e) Finally, make a plot to show how does the error in the predictions depend on the number of components used. To do so, you will have to use cross-vaildation to determine the error produced after applying PLS and PCR with 1,2...10 components.

Load the data and visualize it:

```
load spectra
whos NIR octane

X=NIR;
y=octane;

[dummy,h] = sort(octane);
oldorder = get(gcf,'DefaultAxesColorOrder');
set(gcf,'DefaultAxesColorOrder',jet(60));
plot3(repmat(1:401,60,1)',repmat(octane(h),1,401)',NIR(h,:)');
set(gcf,'DefaultAxesColorOrder',oldorder);
xlabel('Wavelength Index'); ylabel('Octane'); axis('tight');
grid on
```