Data mining: Homework 2

Vishvas Vasuki

October 6, 2009

1 1

1.1 Notation

Apples, cherries, oranges are denoted by a, c, o. F denotes fruit drawn. B denotes bag selected.

1.2 Data

$Pr(F|B = 1)$ = .3, .4, .3 for F = a, c, o respectively.
$Pr(F|B = 2)$ = .5, 0, .5 for F = a, c, o respectively.
$Pr(F|B = 3)$ = .4, .3, .3 for F = a, c, o respectively.
Pr(B) = 0.2, 0.2, 0.6 for B = 1, 2, 3 respectively.

1.3 a

$Pr(F = o) = Pr(F = o|B = 1)Pr(B = 1) + Pr(F = o|B = 2)Pr(B = 2) + Pr(F = o|B = 3)Pr(B = 3) = .3 * 0.2 + .5 * .2 + .3 * .6 = 0.34.$

1.4 b

$Pr(B = 2|F = o) = (Pr(F = o|B = 2)Pr(B = 2))/Pr(F = o) = (.5 * 0.2)/.34 = 0.294.$

1.5 c

We assume that once the fruit is drawn from a bag, it is replaced into the same bag before the next drawing.

$Pr(B = 1|F = o) = (Pr(F = o|B = 1)Pr(B = 1))/Pr(F = o) = (.3 * 0.2)/.34 = 0.1765.$

$Pr(B = 3|F = o) = (Pr(F = o|B = 3)Pr(B = 3))/Pr(F = o) = (.3 * 0.6)/.34 = 0.53.$

1.5.1 Additional notation

We denote the event that the first fruit drawn is an orange by: F = o. We denote the event that the 2nd fruit drawn from the same bag is an orange by: F2 = o.

1.5.2 The solution

$Pr(F2 = o|F = o) = Pr(F2 = o|B = 1)Pr(B = 1|F = o) + Pr(F2 = o|B = 2)Pr(B = 2|F = o) + Pr(F2 = o|B = 3)Pr(B = 3|F = o) = .3 * 0.1765 + .5 * 0.294 + .3 * 0.53 = 0.359.$

2 2

2.1 a

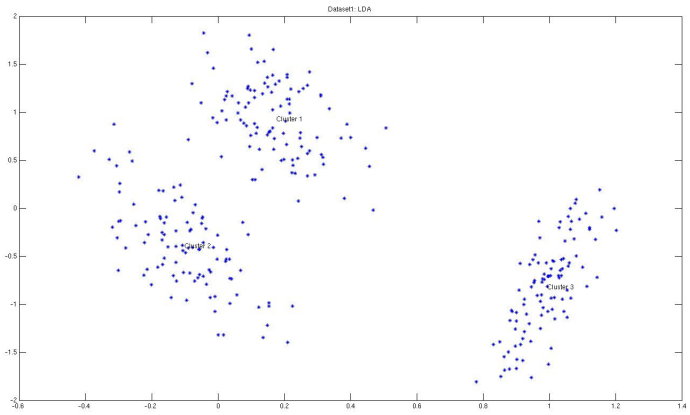
Implementation of PCA and LDA are shown below.

```
%  load /u/vvasuki/vishvas/work/statistics/hw2/hw2data/dataset1
load /u/vvasuki/vishvas/work/statistics/hw2/hw2data/dataset2

%  Common code
[numPoints,numFeatures] = size(X)

L = unique(labels);
X_1s=find(labels==1);
X_2s=find(labels==2);
X_3s=find(labels==3);

mean_1 = sum(X(X_1s,:),1)./numel(X_1s);
mean_2 = sum(X(X_2s,:),1)./numel(X_2s);
mean_3 = sum(X(X_3s,:),1)./numel(X_3s);
mean = sum(X,1)./numPoints;



%  Code for LDA
%  S_B = numel(X_1s)* (mean_1 -mean)'*(mean_1 -mean) + numel(X_2s)*
(mean_2 -mean)'*(mean_2 -mean) + numel(X_3s)* (mean_3 -mean)'*(mean_3
-mean);
%
%  S_W = zeros(numFeatures,numFeatures);
%  for i=(X_1s')
%      S_W = S_W + (X(i,:) -mean_1)'*(X(i,:) -mean_1);
%  end
%  for i=(X_2s')
```

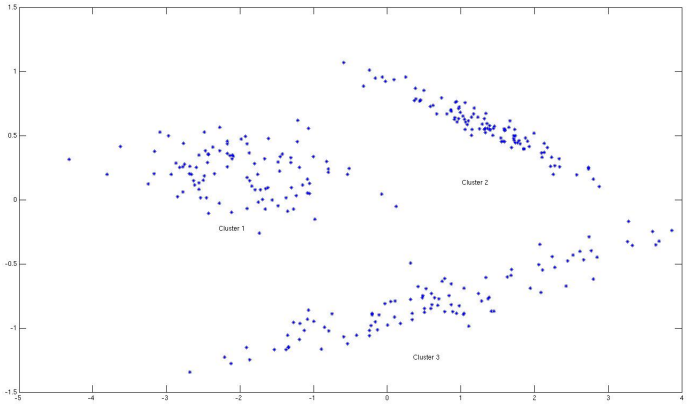
```
%      S_W = S_W + (X(i,:) -mean_2)'*(X(i,:) -mean_2);
%  end
%  for i=(X_3s')
%      S_W = S_W + (X(i,:) -mean_3)'*(X(i,:) -mean_3);
%  end
%
%  %    [U S V] = svd(inv(sqrtm(S_W))*S_B*inv(sqrtm(S_W)));
%  [V L] = eig(inv(S_W)*S_B);
%  [U R] = qr(V);
%  projectedData = X*U(:,1:2);
%  plot(projectedData(:,1),projectedData(:,2),'*')
%  projectedMean_1 = mean_1*U(:,1:2)
%  projectedMean_2 = mean_2*U(:,1:2)
%  projectedMean_3 = mean_3*U(:,1:2)
%  text(projectedMean_1(1),projectedMean_1(2),'Cluster 1')
%  text(projectedMean_2(1),projectedMean_2(2),'Cluster 2')
%  text(projectedMean_3(1),projectedMean_3(2),'Cluster 3')

%  %  Code for PCA
%
Y = X -ones(numPoints,1)*mean;
%  sum(Y,1)
covariance = Y'*Y;
[U S V] = svd(covariance);
projectedData = Y*U(:,1:2);
plot(projectedData(:,1),projectedData(:,2),'*')
projectedMean_1 = mean_1*U(:,1:2)
projectedMean_2 = mean_2*U(:,1:2)
projectedMean_3 = mean_3*U(:,1:2)
text(projectedMean_1(1),projectedMean_1(2),'Cluster 1')
text(projectedMean_2(1),projectedMean_2(2),'Cluster 2')
text(projectedMean_3(1),projectedMean_3(2),'Cluster 3')
```

2.2 b (Dataset 1)

LDA:

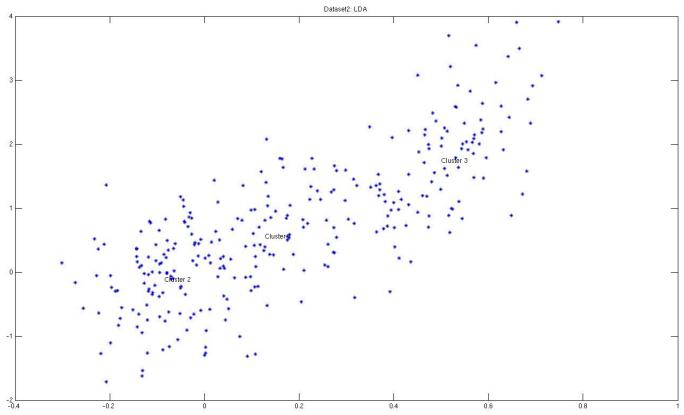


PCA:

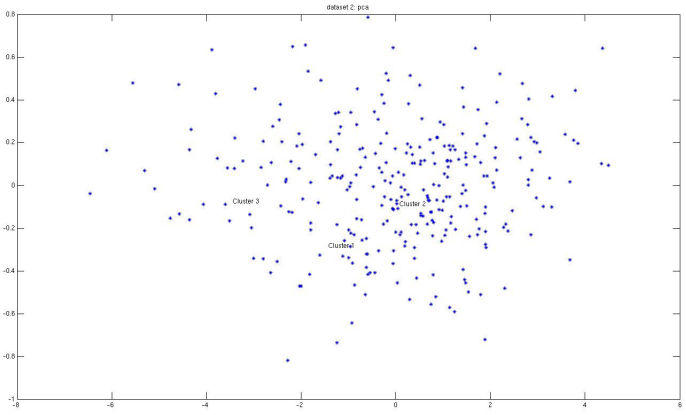


2.3 c (Dataset 2)

LDA:



PCA:



3 3

3.1 Some observations

$(m_2 - m_1)^T \Sigma^{-1} x + c = 0$. As Σ is the covariance matrix, it is real and symmetric, so is Σ^{-1} if it exists. So, $w = \Sigma^{-1}(m_2 - m_1)$. Take its SVD: $\Sigma = USU^T$.

3.2 a

Suppose that Σ is non-singular. Then, the matrix of singular values $S > 0$, $S^{-1} > 0$ exists. Then:

$$\begin{aligned} w^T(m_2 - m_1) &= (m_2 - m_1)^T \Sigma^{-1}(m_2 - m_1) && (1)\\ &= (m_2 - m_1)^T U S^{-1} U^T (m_2 - m_1) && (2)\\ &= q^T S^{-1} q \quad \texttt{(Where:} \quad q = U^T(m_2 - m_1) \neq 0\texttt{)} && (3)\\ &\neq 0 && (4)\\ & && (5) \end{aligned}$$

3.3 b

3.3.1 Rationale

Using the same notation as earlier: $w^T(m_2 - m_1) = q^T S^{-1} q \approx 0$ for $q = U^T(m_2 - m_1) \neq 0$ when $(m_2 - m_1) \approx u_1$, the first singular vector and when the first singular value s_1 is very large, causing $s_1^{-1} \approx 0$.

The singular vectors of Σ happen to correspond to the major axes of the hyper-ellipses which form the level sets of the normal distribution parametrized by Σ. As we want $s_1 >> 0$, and as this corresponds to the length of the major axes of these level sets, the spread of the data should be large. As we want $(m_2 - m_1) \approx u_1$, the distribution should be tilted towards $m_2 - m_1$. This is shown by the figure below.

3.3.2 The figure

The two ellipses represent level sets of 2 normal distributions: $N(x|m_1, \Sigma) = N(x|m_2, \Sigma) = c$ for a constant c. The arrow represents $m_1 - m_2$.

For clarity, the figure does not show the dots which are the points to be clustered, but they can be imagined to be distributed according to the aforementioned distributions. The separating hyperplane, a line in this 2 dimensional case, is almost parallel to $m_1 - m_2$, and $w^T(m_1 - m_2) \approx 0$.