# CIS 520: Problem Set #6

Due on November 10, 2017

*Lyle Ungar, Shivani Agarwal*

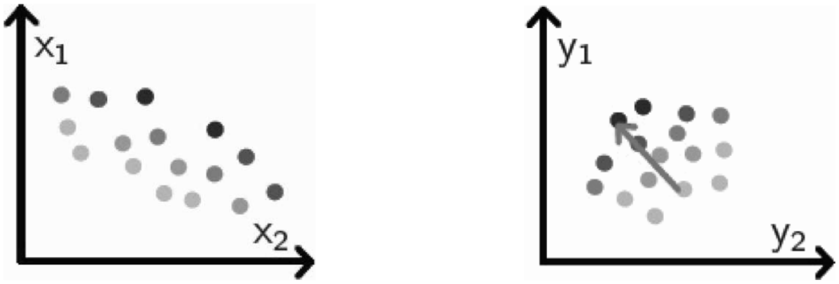
**Eric Oh**

Collaborators: Jiarui Lu

# Problem 1

CCA

## Solution



1.
2. (a)

```
load('data/breast_cancer.mat')
rng(957) ;

X = X_train ;
Y = Y_train ;
Z = (X.' * X)^(-0.5) * (X.' * Y) * (double(Y.') * double(Y))^(-0.5) ;

[U, S, V] = svds(Z) ;
corr((X * U), (Y * V)) ;
```

The correlation is given by 0.9134.

(b)

```
[PCAloadings, PCAscores, PCAvar] = pca(X) ;
betaPCR = regress(Y, PCAscores(:,1)) ;
ypred = PCAscores(:,1) * betaPCR ;

corr(ypred, Y);
```

THe correlation between $\hat{y}$ and $y$ is given by 0.9098.

# Problem 2

Sensational EM

## Solution

# Problem 3

K-Means

## Solution

1. Yes, because

Problem 3 continued on next page. . .

2. (a)

```
X = [2 1; 2 2; 3 1; 3 2;
     8 6; 7 7; 7 8; 8 9;
     12 6; 13 7; 13 8; 12 9;
     17 1; 17 2; 17 3];

rng(23) ;

c_start = [7 7; 8 9; 12 9] ;
[idx1,C1] = kmeans(X, 3, 'MaxIter', 1, 'Start', c_start) ;

figure;
plot(X(idx1==1,1),X(idx1==1,2),'b.','MarkerSize',12);
hold on;
plot(X(idx1==2,1),X(idx1==2,2),'r.','MarkerSize',12);
plot(X(idx1==3,1),X(idx1==3,2),'g.','MarkerSize',12);
plot(c_start(:,1),c_start(:,2),'kx','MarkerSize',15,'LineWidth',3);
text(c_start(1,1)+0.25,c_start(1,2)-0.25,...
['(' num2str(c_start(1,1)) ',' num2str(c_start(1,2)) ')']) ;
text(c_start(2,1)+0.25,c_start(2,2)-0.25,...
['(' num2str(c_start(2,1)) ',' num2str(c_start(2,2)) ')']) ;
text(c_start(3,1)+0.25,c_start(3,2)-0.25,...
['(' num2str(c_start(3,1)) ',' num2str(c_start(3,2)) ')']) ;
legend('Cluster␣1','Cluster␣2','Cluster␣3','Centroids',...
'Location','NorthEast');
hold off;
```

Iter 1 plot

```
[idx2,C2] = kmeans(X, 3, 'MaxIter', 1, 'Start', C1) ;

figure;
plot(X(idx2==1,1),X(idx2==1,2),'b.','MarkerSize',12);
hold on;
plot(X(idx2==2,1),X(idx2==2,2),'r.','MarkerSize',12);
plot(X(idx2==3,1),X(idx2==3,2),'g.','MarkerSize',12);
plot(C1(:,1),C1(:,2),'kx','MarkerSize',15,'LineWidth',3);
text(C1(1,1)+0.25,C1(1,2)-0.25,...
['(' num2str(C1(1,1)) ',' num2str(C1(1,2)) ')']) ;
text(C1(2,1)+0.25,C1(2,2)-0.25,...
['(' num2str(C1(2,1)) ',' num2str(C1(2,2)) ')']) ;
text(C1(3,1)+0.25,C1(3,2)-0.25,...
['(' num2str(C1(3,1)) ',' num2str(C1(3,2)) ')']) ;
legend('Cluster␣1','Cluster␣2','Cluster␣3','Centroids',...
'Location','NorthEast');
hold off;
```

Iter 2 plot

```
[idx3,C3] = kmeans(X, 3, 'MaxIter', 1, 'Start', C2) ;
```

```
figure;
plot(X(idx3==1,1),X(idx3==1,2),'b.','MarkerSize',12);
hold on;
plot(X(idx3==2,1),X(idx3==2,2),'r.','MarkerSize',12);
plot(X(idx3==3,1),X(idx3==3,2),'g.','MarkerSize',12);
plot(C2(:,1),C2(:,2),'kx','MarkerSize',15,'LineWidth',3);
text(C2(1,1)+0.25,C2(1,2)-0.25,...
['(' num2str(C2(1,1)) ',' num2str(C2(1,2)) ')']) ;
text(C2(2,1)+0.25,C2(2,2)-0.25,...
['(' num2str(C2(2,1)) ',' num2str(C2(2,2)) ')']) ;
text(C2(3,1)+0.25,C2(3,2)-0.25,...
['(' num2str(C2(3,1)) ',' num2str(C2(3,2)) ')']) ;
legend('Cluster␣1','Cluster␣2','Cluster␣3','Centroids',...
'Location','NorthEast');
hold off;
```

Iter 3 plot

```
[idx4,C4] = kmeans(X, 3, 'MaxIter', 1, 'Start', C3) ;

figure;
plot(X(idx4==1,1),X(idx4==1,2),'b.','MarkerSize',12);
hold on;
plot(X(idx4==2,1),X(idx4==2,2),'r.','MarkerSize',12);
plot(X(idx4==3,1),X(idx4==3,2),'g.','MarkerSize',12);
plot(C3(:,1),C3(:,2),'kx','MarkerSize',15,'LineWidth',3);
text(C3(1,1)+0.25,C3(1,2)-0.25,...
['(' num2str(C3(1,1)) ',' num2str(C3(1,2)) ')']) ;
text(C3(2,1)+0.25,C3(2,2)-0.25,...
['(' num2str(C3(2,1)) ',' num2str(C3(2,2)) ')']) ;
text(C3(3,1)+0.25,C3(3,2)-0.25,...
['(' num2str(C3(3,1)) ',' num2str(C3(3,2)) ')']) ;
legend('Cluster␣1','Cluster␣2','Cluster␣3','Centroids',...
'Location','NorthEast');
hold off;
```

Iter 4 plot

(b)

```
c_start = [12 6; 8 9; 12 9] ;
[idx1,C1] = kmeans(X, 3, 'MaxIter', 1, 'Start', c_start) ;

figure;
plot(X(idx1==1,1),X(idx1==1,2),'b.','MarkerSize',12);
hold on;
plot(X(idx1==2,1),X(idx1==2,2),'r.','MarkerSize',12);
plot(X(idx1==3,1),X(idx1==3,2),'g.','MarkerSize',12);
plot(c_start(:,1),c_start(:,2),'kx','MarkerSize',15,'LineWidth',3);
text(c_start(1,1)+0.25,c_start(1,2)-0.25,...
['(' num2str(c_start(1,1)) ',' num2str(c_start(1,2)) ')']) ;
text(c_start(2,1)+0.25,c_start(2,2)-0.25,...
['(' num2str(c_start(2,1)) ',' num2str(c_start(2,2)) ')']) ;
text(c_start(3,1)+0.25,c_start(3,2)-0.25,...
['(' num2str(c_start(3,1)) ',' num2str(c_start(3,2)) ')']) ;
legend('Cluster␣1','Cluster␣2','Cluster␣3','Centroids',...
'Location','NorthEast');
hold off;
```

Iter 1 plot

```
[idx2,C2] = kmeans(X, 3, 'MaxIter', 1, 'Start', C1) ;

figure;
plot(X(idx2==1,1),X(idx2==1,2),'b.','MarkerSize',12);
hold on;
plot(X(idx2==2,1),X(idx2==2,2),'r.','MarkerSize',12);
plot(X(idx2==3,1),X(idx2==3,2),'g.','MarkerSize',12);
plot(C1(:,1),C1(:,2),'kx','MarkerSize',15,'LineWidth',3);
text(C1(1,1)+0.25,C1(1,2)-0.25,...
['(' num2str(C1(1,1)) ',' num2str(C1(1,2)) ')']) ;
text(C1(2,1)+0.25,C1(2,2)-0.25,...
['(' num2str(C1(2,1)) ',' num2str(C1(2,2)) ')']) ;
text(C1(3,1)-1.3,C1(3,2)-0.25,...
['(' num2str(C1(3,1)) ',' num2str(C1(3,2)) ')']) ;
legend('Cluster␣1','Cluster␣2','Cluster␣3','Centroids',...
'Location','NorthEast');
hold off;
```

Iter 2 plot

```
[idx3,C3] = kmeans(X, 3, 'MaxIter', 1, 'Start', C2) ;

figure;
plot(X(idx3==1,1),X(idx3==1,2),'b.','MarkerSize',12);
hold on;
plot(X(idx3==2,1),X(idx3==2,2),'r.','MarkerSize',12);
plot(X(idx3==3,1),X(idx3==3,2),'g.','MarkerSize',12);
plot(C2(:,1),C2(:,2),'kx','MarkerSize',15,'LineWidth',3);
text(C2(1,1)-1,C2(1,2)-0.25,...
['(' num2str(C2(1,1)) ',' num2str(C2(1,2)) ')']) ;
text(C2(2,1)+0.25,C2(2,2)-0.25,...
['(' num2str(C2(2,1)) ',' num2str(C2(2,2)) ')']) ;
text(C2(3,1)-1.5,C2(3,2)-0.25,...
['(' num2str(C2(3,1)) ',' num2str(C2(3,2)) ')']) ;
legend('Cluster␣1','Cluster␣2','Cluster␣3','Centroids',...
'Location','NorthEast');
hold off;
```

Iter 3 plot

```
[idx4,C4] = kmeans(X, 3, 'MaxIter', 1, 'Start', C3) ;

```

Problem 3 continued on next page...

```
figure;
plot(X(idx4==1,1),X(idx4==1,2),'b.','MarkerSize',12);
hold on;
plot(X(idx4==2,1),X(idx4==2,2),'r.','MarkerSize',12);
plot(X(idx4==3,1),X(idx4==3,2),'g.','MarkerSize',12);
plot(C3(:,1),C3(:,2),'kx','MarkerSize',15,'LineWidth',3);
text(C3(1,1)+0.25,C3(1,2)-0.25,...
['(' num2str(C3(1,1)) ',' num2str(C3(1,2)) ')']) ;
text(C3(2,1)+0.25,C3(2,2)-0.25,...
['(' num2str(C3(2,1)) ',' num2str(C3(2,2)) ')']) ;
text(C3(3,1)-1,C3(3,2)-0.25,...
['(' num2str(C3(3,1)) ',' num2str(C3(3,2)) ')']) ;
legend('Cluster␣1','Cluster␣2','Cluster␣3','Centroids',...
'Location','NorthEast');
hold off;
```

Iter 4 plot

```
[idx5,C5] = kmeans(X, 3, 'MaxIter', 1, 'Start', C4) ;

figure;
plot(X(idx5==1,1),X(idx5==1,2),'b.','MarkerSize',12);
hold on;
plot(X(idx5==2,1),X(idx5==2,2),'r.','MarkerSize',12);
plot(X(idx5==3,1),X(idx5==3,2),'g.','MarkerSize',12);
plot(C4(:,1),C4(:,2),'kx','MarkerSize',15,'LineWidth',3);
text(C4(1,1)+0.25,C4(1,2)-0.25,...
['(' num2str(C4(1,1)) ',' num2str(C4(1,2)) ')']) ;
text(C4(2,1)+0.25,C4(2,2)-0.25,...
['(' num2str(C4(2,1)) ',' num2str(C4(2,2)) ')']) ;
text(C4(3,1)-1,C4(3,2)-0.25,...
['(' num2str(C4(3,1)) ',' num2str(C4(3,2)) ')']) ;
legend('Cluster␣1','Cluster␣2','Cluster␣3','Centroids',...
'Location','NorthEast');
hold off;
```

Iter 5 plot

# Problem 4

Principal Components Analysis

## Solution

1.
```
load('data/MNIST_train.mat') ;
load('data/MNIST_test.mat') ;

rng(147) ;

[PCAloadings, PCAscores, PCAvar, tsquared, explained] = pca(X_train) ;
```

```
proj1 = PCAscores(Y_train==1,:);
proj2 = PCAscores(Y_train==2,:);

figure;
plot(proj1(:,1),proj2(:,1),'ob','MarkerSize',6);
hold on;
plot(proj1(:,2),proj2(:,2),'+m','MarkerSize',6);
xlabel('PC1') ;
ylabel('PC2') ;
title('Test␣digits␣for␣the␣first␣2␣PCA␣dimensions') ;
legend('PCA␣1','PCA␣2','Location','NorthEast');
hold off;
```

plot

2.

```
mu = mean(X_train) ;
nPC = size(PCAloadings, 2);

err_mat = zeros(nPC, 2);
err_mat(:,1) = 1:nPC;

for pcnum = 1:nPC
  xhat = PCAscores(:,1:pcnum) * PCAloadings(:,1:pcnum)' ;
  xhat = bsxfun(@plus, xhat, mu) ;

  reconstruct_err = sqrt(sum(bsxfun(@minus, X_train, xhat).^2, 2)) ;

  err_mat(pcnum, 2) = mean(reconstruct_err) ;

end

figure;
plot(1:nPC, err_mat(:,2));
xlabel('Principal␣Components␣included') ;
ylabel('Average␣reconstruction␣error') ;
title({'Average␣reconstruction␣error␣as␣a␣function','of␣principal␣components␣included'
```

plot

```
PCvariation = cumsum(explained) ;
minPC = find(PCvariation >= 85, 1) ;
```

The number of principal components needed to explain 85% of the variation is.

3. For a 100 dimensions,

```
numdim = 100 ;
[idx, C] = kmeans(PCAscores(:,1:numdim), 10) ;

test_center = bsxfun(@minus, X_test, mean(X_test)) ;
```

```
project_test = test_center * PCAloadings(:,1:numdim) ;
precision = k_means(PCAscores(:,1:numdim), Y_train, project_test, Y_test, 10);
```

giving an accuracy of .
For 150 dimensions,

```
numdim = 150 ;
[idx, C] = kmeans(PCAscores(:,1:numdim), 10) ;

test_center = bsxfun(@minus, X_test, mean(X_test)) ;
project_test = test_center * PCAloadings(:,1:numdim) ;
precision = k_means(PCAscores(:,1:numdim), Y_train, project_test, Y_test, 10);
```

giving an accuracy of .
For 200 dimensions,

```
numdim = 200 ;
[idx, C] = kmeans(PCAscores(:,1:numdim), 10) ;

test_center = bsxfun(@minus, X_test, mean(X_test)) ;
project_test = test_center * PCAloadings(:,1:numdim) ;
precision = k_means(PCAscores(:,1:numdim), Y_train, project_test, Y_test, 10);
```

giving an accuracy of.

4. blah

5. We run the exact same code as in part 3 with all instances of 10 replaced by 25. For 100 dimensions,