CS178 Homework #5 Solution

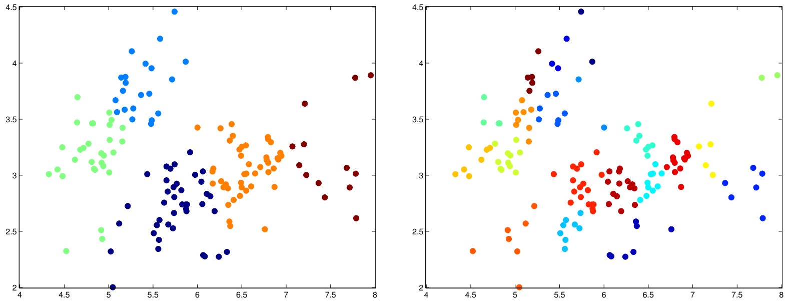
Machine Learning & Data Mining: Winter 2015

Problem 1: Clustering Basics

```
 % Load the iris data, 1st two dimensions:
 iris = load('data/iris.txt'); X=iris(:,1:2);
 % I won't bother plotting it before clustering; you know what it looks like

 % (b) K-means:
 ssd=inf;
 for it=1:10,          % find the best of 10 clusterings:
   [Zi,mui,ssdi] = kmeans(X,5,'random');  % 5 clusters
   if (ssd > ssdi) Z=Zi; mu=mui; ssd=ssdi; end;
 end;
 figure; plotClassify2D([],X,Z); ssd,
% ans = 21.0902
 print('HW5_P1b05.eps','-depsc2'); system('epstopdf HW5_P1b05.eps');

 for it=1:10,          % find the best of 10 clusterings:
   [Zi,mui,ssdi] = kmeans(X,20,'random');  % 20 clusters
   if (ssd > ssdi) Z=Zi; mu=mui; ssd=ssdi; end;
 end;
 figure; plotClassify2D([],X,Z); ssd,
% ans = 4.3728
 print('HW5_P1b20.eps','-depsc2'); system('epstopdf HW5_P1b20.eps');
```



$k = 5$ $k = 20$

Figure 1: k-means clustering results using 5 and 20 clusters. I would argue that the $k = 20$ results probably look over-clustered (k too large).

```
 % (c) Agglomerative clustering
 Z = agglomCluster(X,5,'min');  % single linkage method
 figure; plotClassify2D([],X,Z);
 print('HW5_P1c05a.eps','-depsc2'); system('epstopdf HW5_P1c05a.eps');

 Z = agglomCluster(X,5,'max');  % complete linkage method
 figure; plotClassify2D([],X,Z);
 print('HW5_P1c05b.eps','-depsc2'); system('epstopdf HW5_P1c05b.eps');

 Z = agglomCluster(X,20,'min');  % and again with more clusters
 figure; plotClassify2D([],X,Z);
```

```
print('HW5_P1c20a.eps','-depsc2'); system('epstopdf HW5_P1c20a.eps');

Z = agglomCluster(X,20,'max');  %
figure; plotClassify2D([],X,Z);
print('HW5_P1c20b.eps','-depsc2'); system('epstopdf HW5_P1c20b.eps');
```

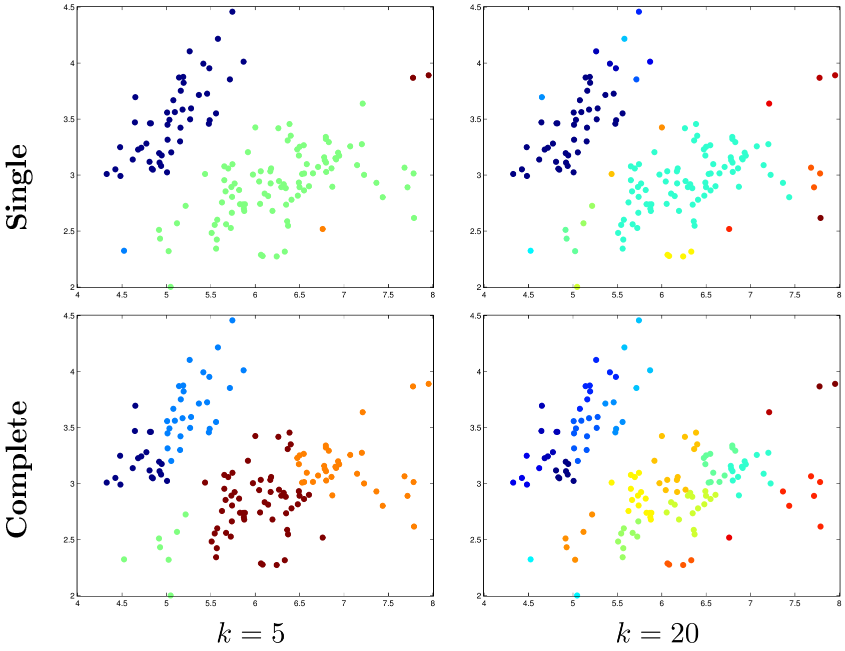


Figure 2: Agglomerative clustering results using 5 and 20 clusters. Single linkage (minimum distance) clustering gives poor results on these data, with a few large clusters and many very small ones. Complete linkage (maximum distance) looks a bit better.

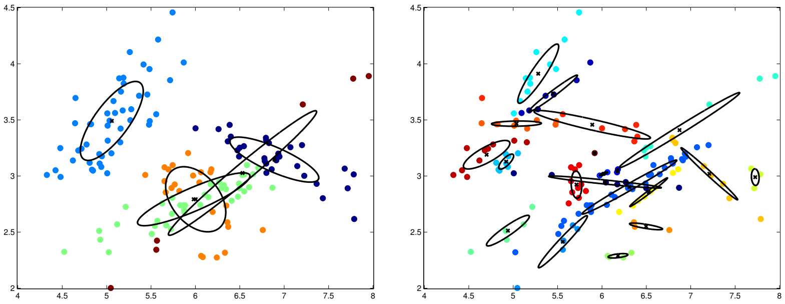
```
% (c) Expectation-Maximization for Gaussian Mixtures:
% Turn on plotting (doPlot = 1) in emCluster.m
[Zi,mui,pri,lli] = emCluster(X,5,'k++'); lli,
% ans = -205.4286
fig(1); print('HW5_P1d05.eps','-depsc2'); system('epstopdf HW5_P1d05.eps');

[Zi,mui,pri,lli] = emCluster(X,20,'k++'); lli,  % 20 clusters
% ans = -105.1284
fig(1); print('HW5_P1d20.eps','-depsc2'); system('epstopdf HW5_P1d20.eps');
```

Problem 2: K-Means on Text

```
% Read in vocabulary and data (word counts per document)
[vocab] = textread('vocab.txt','%s');
[did,wid,cnt] = textread('docword.txt','%d%d%d','headerlines',3);

X = sparse(did,wid,cnt);  % convert to a matlab sparse matrix
D = max(did);             % number of docs
W = max(wid);             % size of vocab
N = sum(cnt);             % total number of words
```



$k = 5$ $k = 20$

Figure 3: Gaussian mixture (EM) clustering results using 5 and 20 clusters. For $k = 5$ the results look reasonable; but for $k = 20$, we get many very small, very narrow clusters, suggesting significant overfitting.

```
% It is often helpful to normalize by the document length:
Xn= X ./ repmat(sum(X,2),[1,W]) ; % divide word counts by doc length
```

```
% (a), (b) : compute a few clusterings and keep the best:
ssd = inf;
for it=1:5,
  [Zi,mui,ssdi] = kmeans(Xn,20,'k++'); ssdi, % compute 20 clusters
  if (ssd > ssdi) Z=Zi; mu=mui; ssd=ssdi; end;
end;
% ans = 2.2491 , 2.1619 , 2.1150 , 2.2089 , 2.1473  : keep #3

% (c)
h = hist(Z,1:20)
% The cluster sizes are:
% 2  1  3  1  1  2  1  12  4  3  7  1  7  1  1  2  12  114  51  2
% so this is a pretty imbalanced clustering...

for i=1:20,
 [sorted,order] = sort( mu(i,:), 2, 'descend');
 fprintf('Cluster %d: ',i); fprintf('%s ',vocab{order(1:10)}); fprintf('\n');
end
% Cluster 1: bowden coach football florida field players head assistant practice ann
% Cluster 2: children including post produced programs television american art black calif
% Cluster 3: hijackers hostages pakistan burger told government indian india passengers killed
% Cluster 4: cats beijing owners police association called carry chinese eat eating
% Cluster 5: season 000 coach harrison james running yards players receiver wide
% Cluster 6: drug marijuana drugs nadelmann policy criminal director foundation group americans
% Cluster 7: plummer dictionary savannah school lady book community daughter 000 began
% Cluster 8: bradley candidates mccain campaign hampshire bush republican political voters party
% Cluster 9: algeria islamic war americans country army europe front independence called
% Cluster 10: white clinton house guests mall washington lincoln memorial president america
% Cluster 11: fireworks city millennium island midnight 000 celebration officials eve y2k
% Cluster 12: atlanta constitution journal moved fbc bowl cox fbn news warrick
% Cluster 13: tutsi hutu rwanda burundi ethnic country africa experts 1994 eck
% Cluster 14: 2000 computer problem systems city failure officials 000 100 ahead
% Cluster 15: test end houston 000 0101 0102 100 1900 1900s 1950s
% Cluster 16: buses authority diesel natural gas plan mta city york hybrid
% Cluster 17: putin yeltsin russia russian power president government political roosevelt chechnya
% Cluster 18: y2k times 2000 millennium york city square 000 century saturday
% Cluster 19: game team season games players coach league play win going
% Cluster 20: russian grozny troops rebels city fighting chechnya ministry soldiers war

% Yes, many of these are fairly interpretable (8: politics; 11: new year's; 19: sports; etc.)
```

```
% (d) Print out the actual documents (head)
Z(1), Z(15), Z(30)
% ans = 19, 18, 17

lst = find(Z==Z(1)); lst=lst(1:min(length(lst),12));
for i=lst',
  fname = sprintf('example1/20000101.%04d.txt',i);
  txt = textread(fname,'%s',2,'whitespace','\r\n'); fprintf('%s\n',txt{:}); fprintf('\n');
end;
%%%% Definitely about sports: %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% School has been out at Cal State Northridge since
% the week before Christmas, but since you can learn something every
%
% Eddie Jones has rejoined the Charlotte Hornets and
% reaffirmed his plans for a quick comeback from an elbow injury.
%
% Just four months ago, only those living in the
% small farming town tucked between the Blue Ridge Mountains and
%
% Bengals receiver Carl Pickens criticized owner and
% general manager Mike Brown, who last week decided to keep coach
%
% As the NFL heads into its last regular-season games on
% the first weekend of 2000, beat reporter Mark Schlabach takes a
%
% Carolina receiver Patrick Jeffers has been red-hot
% lately, after battling injuries through the first half of the
%
% Nobody is quite ready to do celebratory swan dives
% into the Detroit River just yet. Regular-season victories in
%
% Primeau deal goes bust
% Despite the fact Carolina's deal to send Keith Primeau, a
%
% FACEOFF: BRETT HULL
% Right winger, Dallas Stars
%
% Wile E. Coyotes
% Phoenix's decision to dangle stars Keith Tkachuk and Jeremy
%
% Ben Wallace's hairstyle has become one of the big
% sports stories in Orlando.
%
% Grant Long, whose loss to Vancouver left the Hawks
% without a rudder, finally is playing after suffering a concussion
```

```
lst = find(Z==Z(15)); lst=lst(1:min(length(lst),12));
for i=lst',
  fname = sprintf('example1/20000101.%04d.txt',i);
  txt = textread(fname,'%s',2,'whitespace','\r\n'); fprintf('%s\n',txt{:}); fprintf('\n');
end;
%%%% Mostly about the "Y2K Bug": %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% For those who believe that in the good old days _
% before calculators, before computers _ people were better at mental
%
% Airports around the world, ordinarily quiet on New Year's Eve,
% were unusually so Friday as thousands of potential travelers stayed
%
% Two thousand years after Christ's obscure birth in a dusty town
% in Judea, the world's 6 billion people _ most of them non-Christian
%
% The millennium, an idea with overtones ranging from Biblical to
% commercial, had swelled recently into a coercive miniculture as the
%
% The festivities for most of India's 1 billion people were muted
% by comparison with those in wealthier nations, but hundreds of
%
% Rain and chilly weather didn't keep thousands of
% paradegoers from camping out Friday night for the 111th Tournament
%
% Although it had seen many huge crowds over the
% years, for inaugurations, Fourth of July fireworks, protests and
%
% Half a million? A million? A zillion? The world will
% never know how many humans squeezed their way into the Designated
%
% One of the happiest aspects of the Alvin Ailey
% American Dance Theater winter season, which ends Sunday, has been
%
% Despite a few sputters and glitches, the world's computers
% appear to have survived the Year 2000 rollover without major
%
% Don't call Jim and Susan Smith survivalists.
% In fact, you don't even have to call them Jim and Susan _ those
%
% Don't call Jim and Susan Smith survivalists.
% In fact, you don't even have to call them Jim and Susan _ those
```

```
lst = find(Z==Z(30)); lst=lst(1:min(length(lst),12));
for i=lst',
  fname = sprintf('example1/20000101.%04d.txt',i);
  txt = textread(fname,'%s',2,'whitespace','\r\n'); fprintf('%s\n',txt{:}); fprintf('\n');
end;
%%%% Mostly about Russia (some repeats) %%%%%%%%%%%%%%%%%%%%%%%%
% When Boris Yeltsin unveiled his millennium surprise on New
% Year's Eve, he used the word ``power'' five times in his brief
%
% When Boris Yeltsin unveiled his millennium surprise on New
% Year's Eve, he used the word ``power'' five times in his brief
%
% On his first full day as Russia's acting president,
% Vladimir Putin left before dawn to visit with Russian troops
```

```
%
% Only Vladimir Putin knows for certain where he wants to
% take Russia as the nation's acting president. But if his visit last
%
% Such shortcomings can be overcome, he said, but only by devising
% and swiftly carrying out a plan for the nation's restructuring and
%
% Acting president Vladimir Putin moved quickly Saturday
% to establish himself as Russia's new leader by flying to the
%
% As always, Boris Yeltsin picked his moment.
% By choosing to resign on the last day of the old year, Yeltsin
%
% The New York Times said in an editorial for Sunday, Jan. 2:
% In suddenly resigning as Russian president on Friday, Boris
%
% Not content to sit pat a day after assuming power in
% Russia, the acting president, Vladimir Putin, reminded voters
%
% Not content to sit pat a day after assuming power in
% Russia, the acting president, Vladimir Putin, reminded voters
%
% Only Vladimir Putin knows for certain where he wants to
% take Russia as the nation's acting president. But if his visit last
%
% Such shortcomings can be overcome, he said, but only by devising
% and swiftly carrying out a plan for the nation's restructuring and
```

```
% (e) Re-clustering with 40 clusters breaks up some of the larger clusters
% The clustering results still look pretty good, for those documents' clusters
```

Problem 3: EigenFaces

```
% (a)--(b) : Load the data and take the SVD:
X = load('data/faces.txt');          % load face dataset
mu = mean(X);
X0 = X - repmat(mu,[size(X,1),1]);   % remove the mean

[U S V] = svds(X0,50);   % puts singular vectors in descending order
W = U*S;

% (c) Let's look at the reconstruction error as a function of the number of components:
for k=1:20,
  Xhat0 = W(:,1:k)*V(:,1:k)';
  err(k) = mean(mean( (X0 - Xhat0).^2 ));
end;
figure; plot(1:20, err, 'r-', 'linewidth', 4); set(gca,'fontsize',18);
print('HW5_P3c.eps','-depsc2'); system('epstopdf HW5_P3c.eps');
```

```
% (d) Let's look at how the faces vary along each principal component
for k=1:5,
  alpha = 2*median(abs(W(:,k)));
  im1 = reshape(mu + alpha*V(:,k)', [24 24]);
  im2 = reshape(mu - alpha*V(:,k)', [24 24]);
```

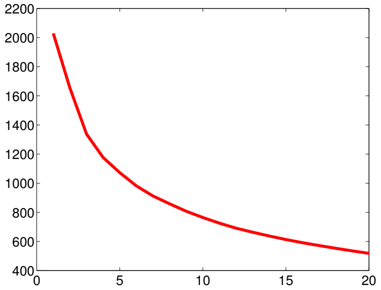
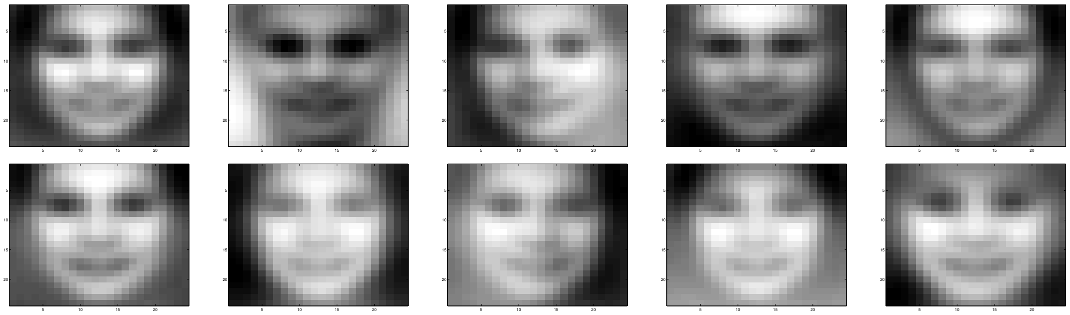


Figure 4: Residual sum-of-squares error in the data after using various numbers of principal components.

```
  fig(1); imagesc(im1); colormap gray;
  fig(2); imagesc(im2); colormap gray;
  fig(1); print(sprintf('HW5_P3d%da.eps',k),'-depsc2'); system(sprintf('epstopdf HW5_P3d%da.eps',k));
  fig(2); print(sprintf('HW5_P3d%db.eps',k),'-depsc2'); system(sprintf('epstopdf HW5_P3d%db.eps',k));
end;
```



First (±) Second (±) Third (±) Fourth (±) Fifth (±)

Figure 5: Looking at image variation in the first five principal components.

```
% (e) Let's plot some of the faces and see what the first two dimensions look like...
idx = ceil(4916*rand(1,25));      % pick some data
figure; hold on; axis ij; colormap(gray);
range = max(W(idx,1:2)) - min(W(idx,1:2)); % find range of coordinates to be plotted
scale = [200 200]./range;                  % want 24x24 to be visible but not large on new scale
for i=idx, imagesc(W(i,1)*scale(1),W(i,2)*scale(2), reshape(X(i,:),24,24)); end;
 print('HW5_P3e.eps','-depsc2'); system('epstopdf HW5_P3e.eps');
```

```
% (f) Reconstruct two faces using a few components
for i=[16 24],
  im = X(i,:);
  im = reshape(im, [24 24]); imagesc(im); colormap gray;
  str=sprintf('HW5_P3f%d.eps',i); print(str,'-depsc2'); system(['epstopdf ' str]);
  for k=[5 10 50],
    im = mu+W(i,1:k)*V(:,1:k)';
    im = reshape(im, [24 24]); imagesc(im); colormap gray;
    str=sprintf('HW5_P3f%d-%d.eps',i,k); print(str,'-depsc2'); system(['epstopdf ' str]);
  end;
end;
```

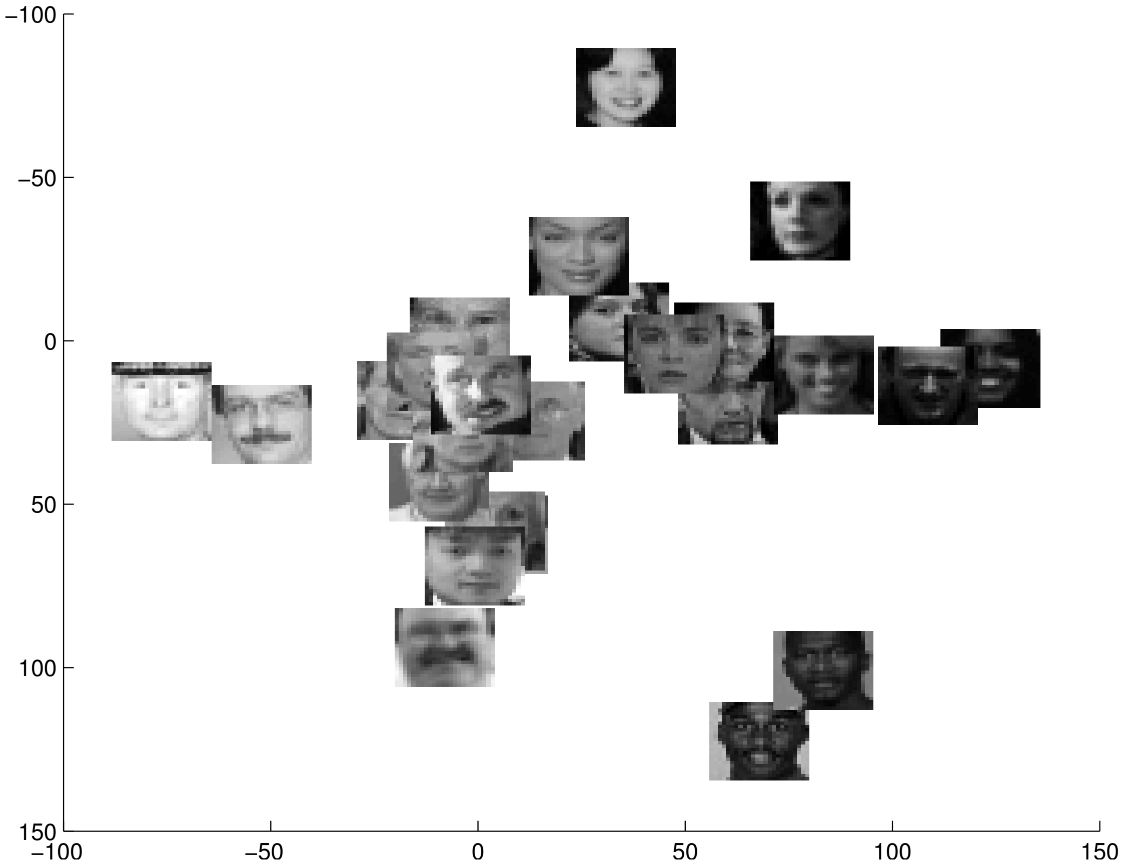
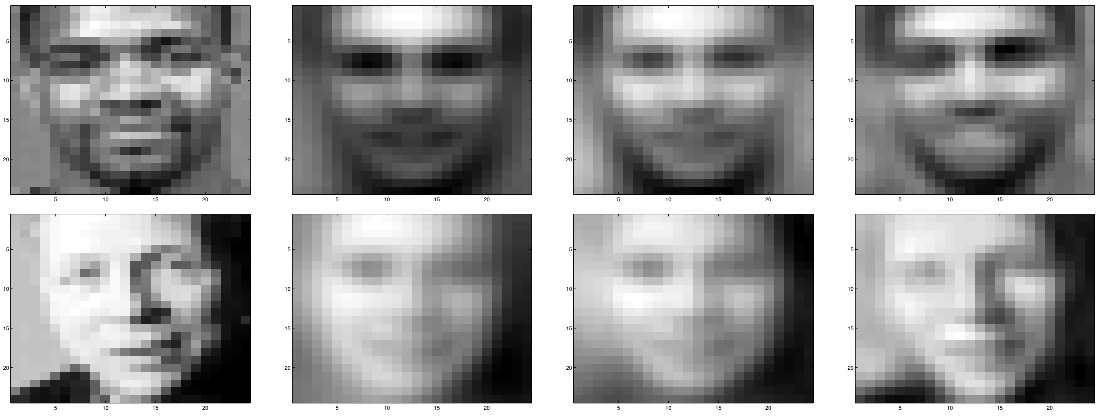


Figure 6: Scatterplot of faces along the first two principal components. The x-axis captures overall lighting, while the y-axis captures relative face vs. background shading.



Original $K = 5$ $K = 10$ $K = 50$

Figure 7: Reconstructing two faces using varying numbers of principal components.