Korea Advanced Institute of Science and Technology
School of Electrical Engineering
EE331 Introduction to Machine Learning, Spring 2019

Due 29, 2019

Homework I

Bereket Eshete 20150923

1. A function that draws the estimate of the probability distribution of the input data x.

Input

$$x = \{x_1, x_2, \dots, x_n\}$$

$$[h, e] = kernelpdf(x)$$

Kernel Density Estimate

$$\hat{P}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K(\frac{x - X_i}{h})$$

Cross-validation

$$h = \frac{1}{n}\sum_{i=1}^{n}\hat{P}_{(-i)}(x_i)$$

Code:

```
% Homework Question 1
%Draws estimate of pdf
function [h,e] = kernelpdf(data)
    k= @(x) (3/4)*(1-x.^2);
    %Gaussian
    tpdf = @(x) exp(-0.5.*((x-
mean(data))/std(data)).^2)/(std(data)*sqrt(2*pi));
    h = 1;  % Bandwidth
    kernel = @(x) mean(k((x-data)/h)/h); % Kernel Density
    kpdf = @(x) arrayfun(kernel,x);
    e= fminbnd(@(x) kpdf(x),0,1);
    x = linspace(min(data),max(data),length(data));
    plot(x,kpdf(x)) %plot pdf estimate
    hold on
    plot(x,tpdf(x)) %plot gaussian pdf
end
```

2.
(i) What is the entropy of the distribution of Sport in the training data? What about Position?

Entropy is given as

$$I = -\sum p(c)\log_2 p(c)$$

Entropy for sport.

- 4 Basketball (B)
- 4 Football (F)
- Class probabilities

$$p(B) = \frac{4}{8} \qquad p(F) = \frac{4}{8}$$

- Entropy

$$-\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8} = 1\ bit$$

Entropy for position

- 4 Guards (G)
- 2 Centers (C)
- 2 Kickers (K)
- Class probabilities

$$p(G) = \frac{4}{8} \quad p(C) = \frac{2}{8} \quad p(F) = \frac{2}{8}$$

- Entropy

$$-\frac{4}{8}\log_2\frac{4}{8} - \frac{2}{8}\log_2\frac{2}{8} - \frac{2}{8}\log_2\frac{2}{8} = 1.5\ bits$$

ii) Calculate information gain for a split after partition by sport variable.

$$\boxed{Gain(A) = \ I - \ Ires\ (A)}$$

$$I_{res} = -\sum_{v} p(v) \sum_{c} p(c|v)\log_2 p(c|v)$$

$I = 1.5\text{bits}$ (Before partition)

$$I(B) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} - \frac{0}{4}\log_2\frac{0}{4} = \ 0.8113\ bits$$

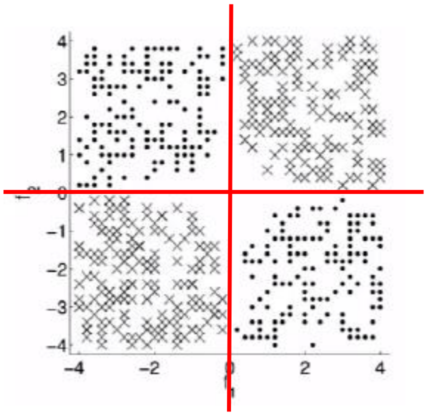
$$I(F) \ = \ -\frac{1}{4}\log_2\frac{1}{4} - \frac{2}{4}\log_2\frac{2}{4} - \frac{1}{4}\log_2\frac{1}{4} = 1.5\ bits$$

$$I_{res}(Sport) = \ p(B)I(B) + p(F)I(B) = 1.156\ bits$$
$$I_{res}(Sport) = \ \frac{4}{8}0.8113 + \frac{4}{8}1.5 = 1.156\ bits$$
$$Gain(Sport) = \ I - \ Ires\ (A) = 1.5 - 1.156 = 0.344\ bits$$

iii) Calculate the information gain for the decision stumps (one-split trees) created by first splitting on Position, Name, and College. Do any of these perfectly classify the training data? Does it make sense to use Name as a variable? Why or why not?

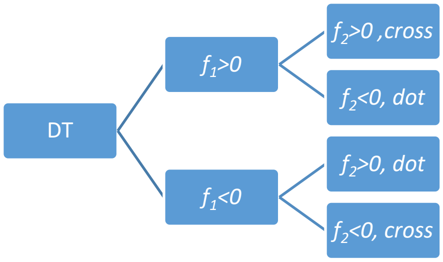
No, Name as a variable does not make sense because each player have a different name leaving no opportunity for information gain.

iv) Draw a decision tree that each correctly classify the training data, and show how their predictions vary on the test set.



Piecewise decision tree

$$f(x) = \begin{cases} \ \ Class\ dot, & (f_1 > 0, f_2 < 0) \cap (f_1 < 0, f_2 > 0) \\ Class\ cross, & (f_1 < 0, f_2 < 0) \cap (f_1 > 0, f_2 > 0) \end{cases}$$



Show how their predictions vary on the test set.

- ► It is possible to construct a decision tree that correctly classifies the training data.

v) Construct a decision tree to separate the two classes in Figure 2. Is this possible?

Yes, decision tree is shown in (iv).

vi) If you construct a decision tree all the way, such that each leaf node has a single class, are the two classes separated?

Yes.

3. Clustering

(a) Show the result of one iteration of k-means clustering assuming k=2 and the initial cluster centers are defined as c1 = (165; 48), and c2 = (178; 60) using the information in Table 1.
**Hint: find the initial clusters, new cluster centers, and obtain new clusters

	Height	Weight	D1	D2	Nearest Cluster	New Centroid
Daniel	180	67	24.20744	7.28011	2	(177.67,64.67)
Seongwu	179	63	20.51828	3.162278	2	
Sungwoon	168	55	7.615773	11.18034	1	(162.67,48)
Jaehwan	174	61	15.81139	4.123106	2	
Jieun	162	44	5	22.62742	1	
Joohyun	158	45	7.615773	25	1	

C1 = (177.67, 64.67); C2= (162.67, 48)

b) c1 = (162; 55), and c2 = (176; 72)

	Height	Weight	D1	D2	Nearest Cluster	New Centroid
Daniel	180	67	21.63331	6.403124	2	(177.67,64.67)
Seongwu	179	63	18.78829	9.486833	2	
Sungwoon	168	55	6	18.78829	1	(162.67,48)
Jaehwan	174	61	13.41641	11.18034	2	
Jieun	162	44	11	31.30495	1	
Joohyun	158	45	10.77033	32.44996	1	

C1 = (177.67, 64.67); C2= (162.67, 48)

C) Explain why Lloyd's algorithm is always guaranteed to converge (i.e. stop) in a Finite number of steps.

There are at most k^N ways to partition N data points into k clusters. This is a large but finite number. For each iteration of the algorithm, we produce a new clustering based only on the old clustering. (1) If the old clustering is the same as the new, then the next clustering will again be the same; (2) if the new clustering is different from the old then the newer one has a lower cost. Since the algorithm iterates a function whose domain is a finite set, the iteration must eventually enter a cycle. The cycle cannot have length greater than 1 because otherwise by (2) you would have some clustering which has a lower cost than itself, which is impossible. Hence, the cycle must have length exactly 1. Hence k-means converges in a finite number of iterations

D) Explain how the exact minimum of the k-means objective behaves on any data set
as we increase K from 1 to n.

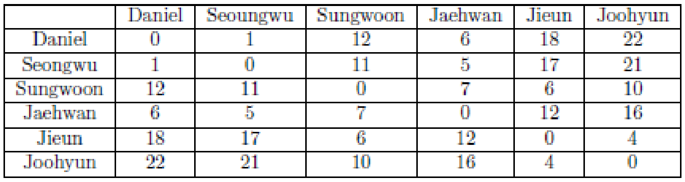
As we increase K to n, when K equals the number of samples in the data set, every point in the data set corresponds to its own cluster and the total distortion is zero.

ii) Agglomerative clustering

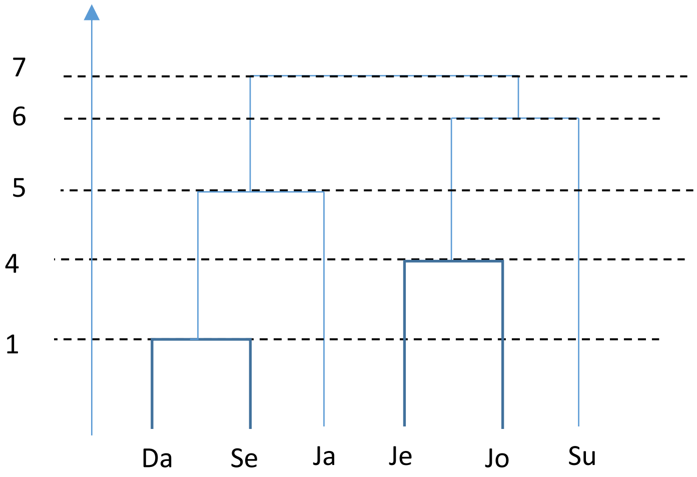
a. Show the dendrogram of hierarchical agglomerative single-link clustering using the information in Table 2.

i.e. For easier representation, we use the first two letters to represent the data variable. (eg, Daniel –Da)
- The distance between two groups is given as

$$d_{SL}(G,H) = \min_{i \in G, i' \in H} d_{i,i'}$$



	Daniel	Seoungwu	Sungwoon	Jaehwan	Jieun	Joohyun
Daniel	0	1	12	6	18	22
Seongwu	1	0	11	5	17	21
Sungwoon	12	11	0	7	6	10
Jaehwan	6	5	7	0	12	16
Jieun	18	17	6	12	0	4
Joohyun	22	21	10	16	4	0



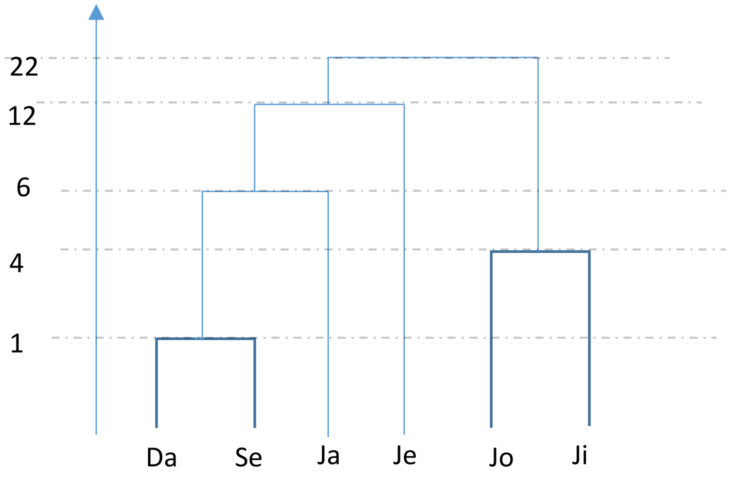
b. What clusters of singers are created if you want 2 clusters using the result of 3(ii)a?

Cluster 1= {Da, Se, Ja}
Cluster 2= {Je, Jo, Su}

c. What clusters of singers are created if you want 3 clusters using the result of 3(ii)a?

Cluster 1= {Da, Se, Ja}
Cluster 2= {Je, Jo}
Cluster 3= {Su} (i.e However in some books, one variable is considered as singleton and not cluster, in that case, 3 clusters is not possible)

d. Show the dendrogram of hierarchical agglomerative complete-link clustering using the information in Table 2.



e. What clusters of singers are created if you want 2 clusters using the result of 3(ii)d?

Cluster 1= {Da, Se, Ja, Je}
Cluster 2= {Jo, Ji}

f. What clusters of singers are created if you want 3 clusters using the result of 3(ii)d?

Cluster 1= {Da, Se, Ja}
Cluster 2= {Je}
Cluster 3= {Jo, Ji} (i.e However in some books, one variable is considered as singleton and not cluster, in that case, 3 clusters is not possible)

4. PCA

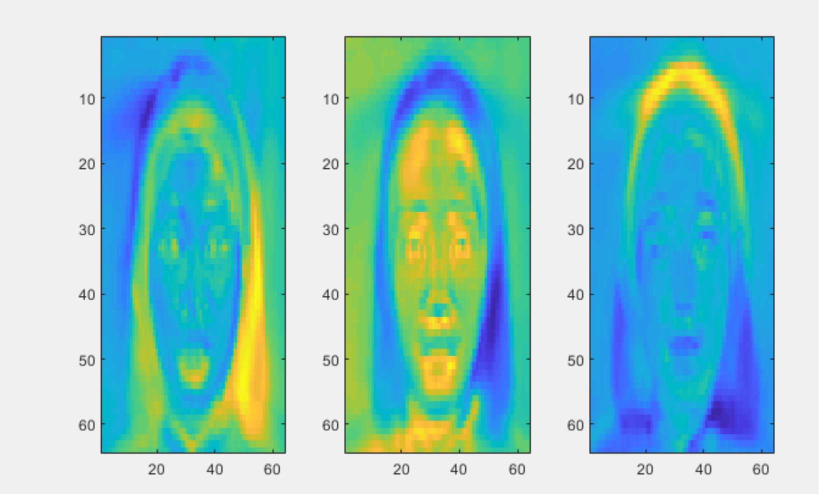
213 human-face images (163 for training, 50 for testing)

10 individuals

Each face have different expression

(i) Perform PCA on the 163 training images with different number of principal components (PC) then display the top 5 principal components. Also, reconstruct each image using 5, 50, 200 and 500 principal components. Compare the average reconstruction mean square error versus the number of PCs.

K=5 ; recon_erroe= 1.6692e+06



(ii) Use PCA representation to identify the 50 test face images. Write a Matlab code to identify input test images. Use the Euclidean distance as a measure of closeness.