Introduction to Machine Learning

First midterm test sample

The test

Q1 PCA

A centered dataset with $n = 102$ observations and $p = 9$ variables was analysed to reduce its dimensionality. The following is a list of singular values of $\mathbf{X}$ in decreasing order, that is $d_1, d_2, \ldots, d_9$: 341.6762, 172.7133, 157.3526, 151.302, 143.0832, 135.1338, 132.0701, 122.3195, 109.3752.

A) Compute and write the numerical value of the eigenvalue λ_3 of $\mathbf{\Sigma}$. This eigenvalue is located in the position $(3, 3)$ of the matrix $\mathbf{\Lambda}$ and is simultaneously the sample variance of the score PC3. $\lambda_3 =$ [ ].

B) Compute and write the percentage of total variability explained by the Principal component PC3. The number you write should be between 0 and 100 and you should include decimals in your answer. The percentage is [ ].

C) A threshold of total variability explained has been set at 80%. How many principal components must you select? Write your answer (integer value). You must select [ ] components.

D) With the information given above, briefly discuss what statements can be said about a) the covariances between the variables in data $\mathbf{X}$ and b) the covariances between projected scores $\mathbf{Z}$.

Q2 Cluster

Consider the following data set with $n = 6$ observations and $p = 3$ variables. The data set is given next.

	$V1$	$V2$	$V3$
A	3.1	4.2	1.3
B	0.8	0.7	4.2
C	4.2	2.8	2.2
D	1	1	4.2
E	4.2	1.7	1.1
F	3	4.2	3.8

We also give the distance matrix using the "Manhattan'' metric. The symbol x in the matrix below is to be calculated later.

	A	B	C	D	E	F
A	0	8.7	3.4	8.2	3.8	2.6
B	8.7	0	x	0.5	7.5	6.1
C	3.4	x	0	7	2.2	4.2
D	8.2	0.5	7	0	7	5.6
E	3.8	7.5	2.2	7	0	6.4
F	2.6	6.1	4.2	5.6	6.4	0

A) In the distance matrix there is a missing distance x. Compute its value and write it. $x =$ ☐.

B) Consider two arbitrary clusters ABCDF and E Compute and write the dissimilarity between these clusters under single linkage. The dissimilarity is ☐.

C) Using the above data $\mathbf{X}$, the `R` command `KM<-kmeans(x=X,centers=3)` was run, with the following output

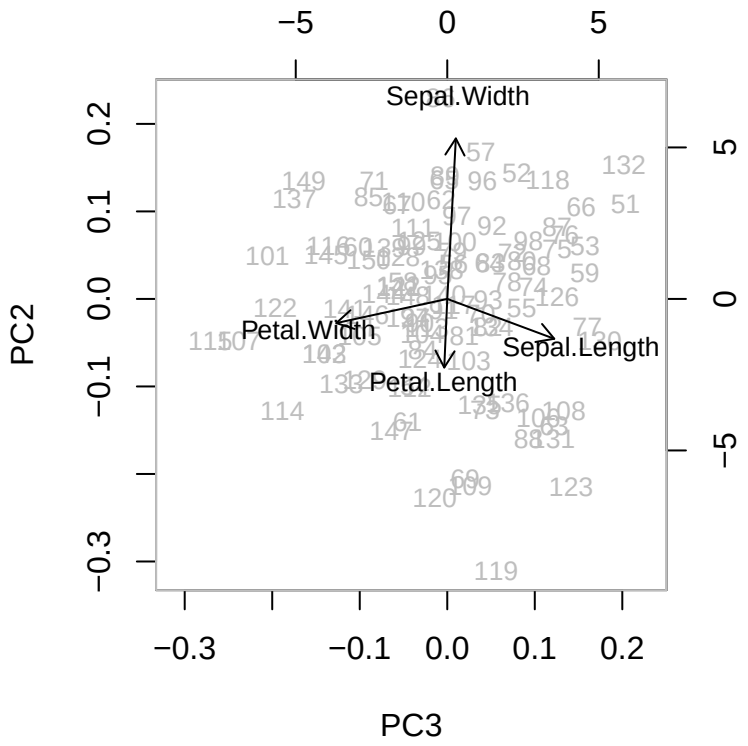
```
> KM$cluster

[1] 3, 2, 1, 2, 1, 3
```

There is interest in determining the center of the cluster identified with the label `2`. By computing this center manually or otherwise, identify which of the following is the correct centroid of this cluster:

- Centroid is (4.2, 2.25, 1.65)
- Centroid is (0.9, 0.85, 4.2)
- Centroid is (3.05, 4.2, 2.55)
- Centroid is (2.7167, 2.4333, 2.8)
- Centroid is (3.05, 2.25, 3)

Q3 Plot

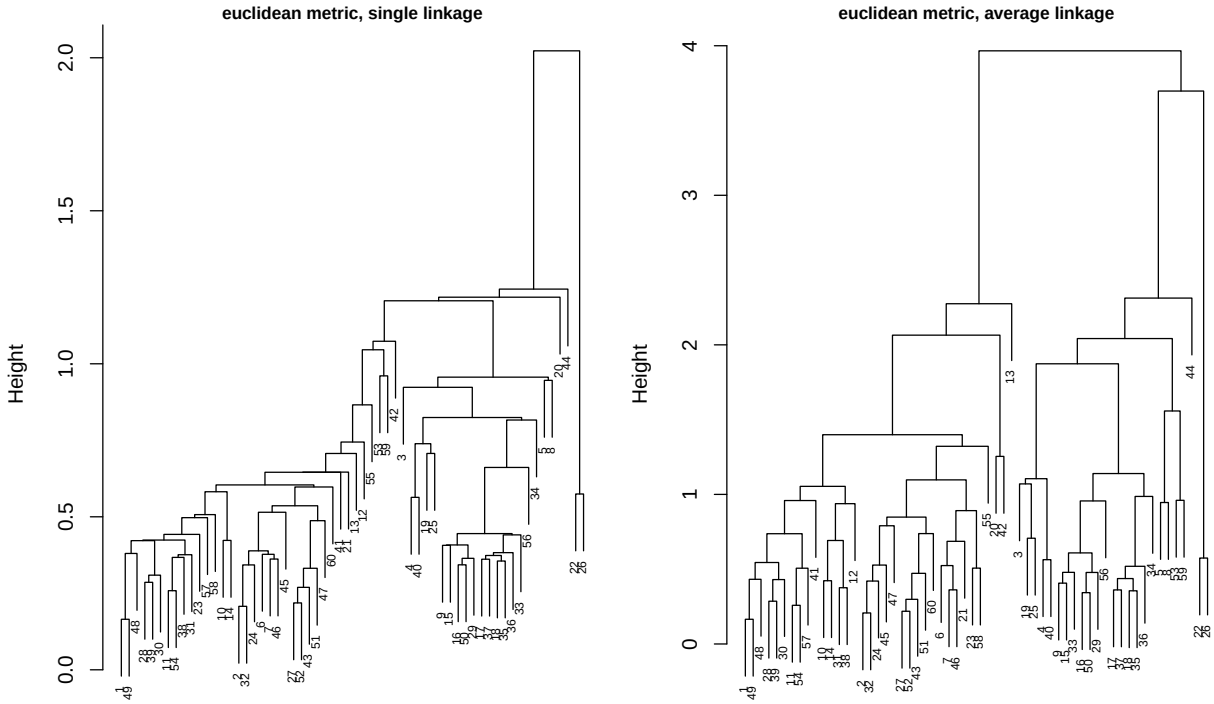
The dataset "iris" has variables `Sepal.Length, Sepal.Width, Petal.Length, Petal.Width`. A principal component analysis of this data was carried out and as part of the analysis, the following diagram was produced.



Using your own words and **no** mathematical symbols, a) briefly explain what was plotted, and b) interpret the diagram.

Q4 Plot

Bronze Age cups from Italy were measured in an archæological study. The data has $n = 60$ observations on $p = 5$ variables. To describe the data, the following dendrograms were created.



Describe the **single** linkage dendrogram and suggest a number of clusters.

N.B. You are **not** expected to describe the procedure of agglomerative clustering, but to describe the dendrogram.

The solution

Q1 PCA

A centered dataset with $n = 102$ observations and $p = 9$ variables was analysed to reduce its dimensionality. The following is a list of singular values of $\mathbf{X}$ in decreasing order, that is $d_1, d_2, \ldots, d_9$: 341.6762, 172.7133, 157.3526, 151.302, 143.0832, 135.1338, 132.0701, 122.3195, 109.3752.

A) Compute and write the numerical value of the eigenvalue λ_3 of $\mathbf{\Sigma}$. This eigenvalue is located in the position $(3, 3)$ of the matrix $\mathbf{\Lambda}$ and is simultaneously the sample variance of the score PC3. $\lambda_3 =$ $\boxed{245.1469379}$.

B) Compute and write the percentage of total variability explained by the Principal component PC3. The number you write should be between 0 and 100 and you should include decimals in your answer. The percentage is $\boxed{8.928}$.

C) A threshold of total variability explained has been set at 80%. How many principal components must you select? Write your answer (integer value). You must select $\boxed{6}$ components.

D) With the information given above, briefly discuss what statements can be said about a) the covariances between the variables in data $\mathbf{X}$ and b) the covariances between projected scores $\mathbf{Z}$.

(This item I leave it to you to answer)

Q2 Cluster

Consider the following data set with $n = 6$ observations and $p = 3$ variables. The data set is given next.

	$V1$	$V2$	$V3$
A	3.1	4.2	1.3
B	0.8	0.7	4.2
C	4.2	2.8	2.2
D	1	1	4.2
E	4.2	1.7	1.1
F	3	4.2	3.8

We also give the distance matrix using the "Manhattan'' metric. The symbol x in the matrix below is to be calculated later.

	A	B	C	D	E	F
A	0	8.7	3.4	8.2	3.8	2.6
B	8.7	0	x	0.5	7.5	6.1
C	3.4	x	0	7	2.2	4.2
D	8.2	0.5	7	0	7	5.6
E	3.8	7.5	2.2	7	0	6.4
F	2.6	6.1	4.2	5.6	6.4	0

A) In the distance matrix there is a missing distance x. Compute its value and write it. $x = \boxed{7.5}$.

B) Consider two arbitrary clusters ABCDF and E Compute and write the dissimilarity between these clusters under single linkage. The dissimilarity is $\boxed{2.2}$.

C) Using the above data $\mathbf{X}$, the `R` command `KM<-kmeans(x=X,centers=3)` was run, with the following output

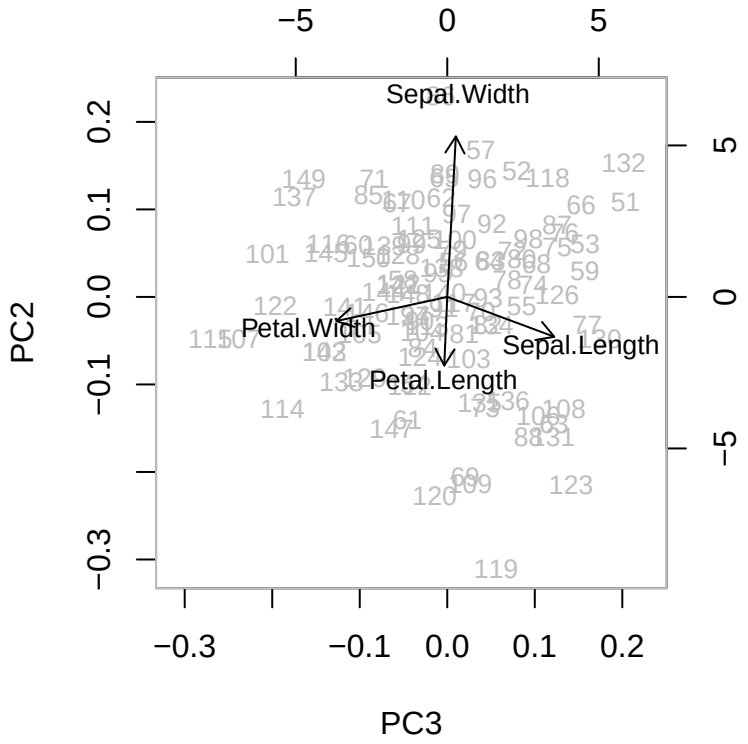
```
> KM$cluster

[1] 3, 2, 1, 2, 1, 3
```

There is interest in determining the center of the cluster identified with the label `2`. By computing this center manually or otherwise, identify which of the following is the correct centroid of this cluster:

- Centroid is (4.2, 2.25, 1.65)
- Centroid is (0.9, 0.85, 4.2) *
- Centroid is (3.05, 4.2, 2.55)
- Centroid is (2.7167, 2.4333, 2.8)
- Centroid is (3.05, 2.25, 3)

Q3 Plot

The dataset "iris" has variables `Sepal.Length, Sepal.Width, Petal.Length, Petal.Width`. A principal component analysis of this data was carried out and as part of the analysis, the following diagram was produced.



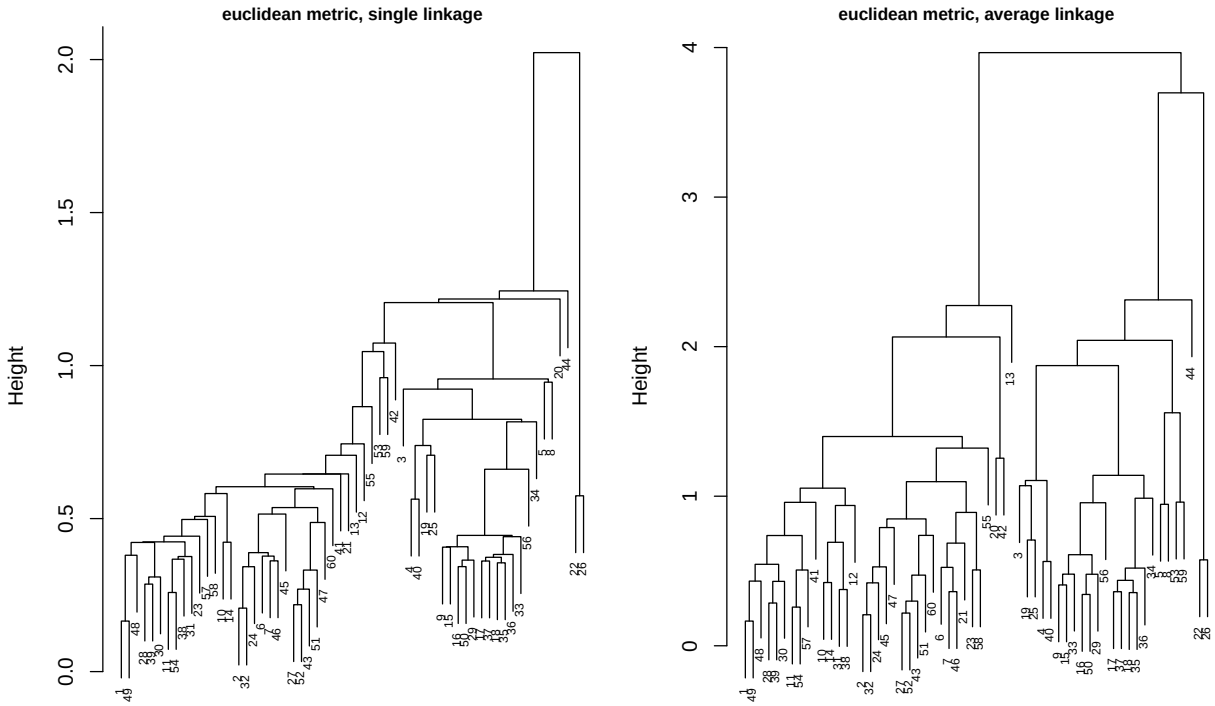
Using your own words and **no** mathematical symbols, a) briefly explain what was plotted, and b) interpret the diagram.

a) This diagram is a biplot with two sets of axes. One set of axes is given by the scores, which in this case are for the second and third principal components. With these axes, the projected data coordinates are plotted. The second set of axes are the second and third coordinates of loadings (eigenvectors of the spectral decomposition). Using this second set of axes, the arrows are the weights of variables involved, that is, the eigenvalue coordinate values.
b) The **second** principal component is a contrast between variable `Sepal.width` and a weighted average of the other variables `Petal.width`, `Sepal.length` and `Petal.length` with this last variable having slightly larger weight than the other two. The **third** principal component is a contrast between variables `Sepal.length` and `Petal.width`, with the other two variables having smaller weights.
Concerning plotted scores, it could be worthy investigating further individual 119, that appears to be relatively away from the cloud of points in the direction of `Petal.length`.
This analysis was likely done with the data scaled (I leave it to you to justify why).

Q4 Plot

Bronze Age cups from Italy were measured in an archæological study. The data has $n = 60$ observations on $p = 5$ variables. To describe the data, the following dendrograms were

created.



Describe the **single** linkage dendrogram and suggest a number of clusters.

N.B. You are **not** expected to describe the procedure of agglomerative clustering, but to describe the dendrogram.

This dendrogram is typical of single linkage, with small clusters joining at a relative large height values and a consequent cluster that is not very clear. If we cut at a height of about 1.2 we have five clusters, of which three are quite small (two with one cup and one with two cups). Cutting further down the dendrogram exacerbates this phenomenon, and if for example, we cut at a height of about 1.0, we have seven clusters (of which three have a single cup and two have two cups each).
We'd suggest five clusters cutting at about the first height of 1.2.