IAML (Level 10) Assignment 2

Yifei Zhao

TOTAL POINTS

63.5 / 75

QUESTION 1

Question 1 30 pts

1.1 3 / 3

✓ **- 0 pts Correct**

- 3 pts You did not answer the question

- 1.5 pts You reported values for the first training samples are all incorrect. The correct values are -3.14e-06, -2.27e-05, -1.18e-04, -4.07e-04.

- 1 pts You reported values for the first training samples are partially incorrect. The correct values are -3.14e-06, -2.27e-05, -1.18e-04, -4.07e-04.

- 1.5 pts You reported values for the last training samples are incorrect. The correct values are -3.137e -06, -2.268e-05, -1.180e-04, -4.071e-04.

- 1 pts You reported values for the last training samples are partially incorrect. The correct values are -3.14e-06, -2.27e-05, -1.18e-04, -4.07e-04.

- 0.5 pts You reported results with too many/few significant figures. 2 or 3 significant figures would have been sufficient.

- 1 pts Answer too long/answer box resized

1.2 3.5 / 4

- 0 pts Correct

- 4 pts You did not answer the question

- 1 pts Your displayed images of the mean vectors are not correct

- 2 pts Most of the images of samples shown are different from the correct ones

- 1 pts Some of the images of samples shown are different from the correct ones

- 1 pts The sample numbers identified are different from the correct ones

- 1 pts The plot is unclear or insufficiently labelled

- 1.5 pts You failed to report findings

- 0.5 pts You failed to provide meaningful discussions

- 1 pts Answer too long/answer box resized

✓ **- 0.5 pts Furthest samples are very different indicating that they may have been labelled wrongly**

1.3 3 / 3

✓ **- 0 pts Correct**

- 3 pts You did not answer the question

- 3 pts Your reported variances are all incorrect

- 2 pts Your reported variances are partially incorrect

- 0.5 pts You reported results with too many/few digits after the decimal place. Less than four would have been sufficient.

- 1 pts Answer too long/answer box resized

- 2 pts You were expected to report the cumulative explained variance for each component.

- 0.5 pts Wrong or insufficient labelling.

1.4 3 / 3

✓ **- 0 pts Correct**

- 3 pts You did not answer the question

- 2 pts You did not include the plot

- 2 pts The graph is not correct

- 1 pts The plot is unclear or insufficiently labelled i.e. you did not label the axes

- 1 pts You did not include discussions or your discussions do not make sense.

- 0.5 pts Your discussions are general and not specific to the data

- 1 pts Answer too long/answer box resized

1.5 3.5 / 4

- 0 pts Correct

- **4 pts** You did not answer the question
- **2 pts** You did not include the images
- **2 pts** Your displayed images do not match the correct ones
- **1 pts** Some of your displayed images do not match the correct ones
- **0.5 pts** The plot is unclear or insufficiently labelled, e.g. you did not label each image
- **2 pts** You did not include discussions
- **1 pts** Your discussions are not informative or not specific to the data
- **1 pts** Your discussions lack details, e.g. no discussions on different principal com- ponents

✓ **- 0.5 pts Images/captions are too small to read**

- **0.5 pts** Some of your discussions do not make sense
- **1 pts** Answer too long/answer box resized

1.6 5 / 5

✓ **- 0 pts Correct**

- **5 pts** You did not answer the question
- **5 pts** Your reported RMSE values are all incorrect
- **4 pts** Most of the values of RMSE are incorrect
- **3 pts** More than the half of the values of RMSE are incorrect
- **2 pts** Some of the values of RMSE are not correct
- **1 pts** The answer is not presented in the specified table format
- **1 pts** You reported results with too many digits after the decimal place. Less than four would have been sufficient.
- **1 pts** Answer too long/answer box resized
- **1 pts** There is a significant typo or error in the way (otherwise correct) values are reported.

1.7 2.5 / 4

- **0 pts** Correct
- **4 pts** You did not answer the question
- **2.5 pts** You failed to include plots
- **2.5 pts** All the images are incorrect
- **2 pts** More than the half of the images are incorrect
- **1 pts** Some of the images are incorrect

✓ **- 0.5 pts The plot is unclear or insufficiently labelled i.e. you did not label the axes**

- **1.5 pts** You failed to include discussions, or your description is too short/inaccurate.
- **1 pts** Your discussions are general and not specific to the data.

✓ **- 1 pts Your discussion does not fully relate to your results from previous questions.**

- **0.5 pts** Some of your discussions do not make sense
- **1 pts** Your discussion is too short - you should relate to your results from previous questions, and talk more specifically about the dataset.
- **1 pts** Answer too long/answer box resized

No points taken for this, but the axis ticks for each image are unnecessary here.

1.8 3.5 / 4

- **0 pts** Correct
- **4 pts** You did not answer the question
- **3 pts** You failed to include the plot
- **3 pts** Your plot is totally different from the expected plot
- **2 pts** Your plot is very different from the expected one
- **1 pts** Your plot is slightly different from the expected plot
- **1 pts** The plot is unclear or insufficiently labelled, e.g. you did not label the axes.
- **0.5 pts** You failed to comment on the separation of classes
- **0.5 pts** You failed to explain your findings
- **0.5 pts** Your findings are not specific to the data

✓ **- 0.5 pts You should have included a bit more discussion.**

- **1 pts** Answer too long/answer box resized

QUESTION 2

Question 2 25 pts

2.1 2.5 / 3

- **0 pts** Correct
- **3 pts** You did not answer the question
- **1 pts** You failed to report the classification accuracy
- **1 pts** Your reported classification accuracy is different from the expected one (0.8401)
- **0.5 pts** You reported result with too many digits after the decimal place. Less than five would have been sufficient.
- **2 pts** You failed to report the confusion matrix
- **2 pts** Your reported confusion matrix is very different from the expected one
- **1 pts** Your reported confusion matrix is slightly different from the expected one

✓ **- 0.5 pts Your confusion matrix is unclear, e.g. unclear what information each element represents**

- **1 pts** Answer too long/answer box resized

2.2 2.5 / 3

- **0 pts** Correct
- **3 pts** You did not answer the question
- **1 pts** You failed to report the classification accuracy
- **1 pts** Your reported classification accuracy is different from the expected one (0.8461)
- **0.5 pts** You reported result with too many digits after the decimal place. Less than five would have been sufficient.
- **2 pts** You failed to report the confusion matrix
- **2 pts** Your reported confusion matrix is very different from the expected one
- **1 pts** Your reported confusion matrix is slightly different from the expected one

✓ **- 0.5 pts Your confusion matrix is unclear, e.g. unclear what information each element represents**

- **1 pts** Answer too long/answer box resized

2.3 4.5 / 6

- **0 pts** Correct
- **4 pts** Your plot of decision regions is very different from the correct one
- **3 pts** Your plot of decision regions does not match the correct one very much
- **1 pts** Your plot of decision regions has a minor difference from the correct one
- **0.5 pts** The plot is unclear or insufficiently labelled, e.g. you did not label the axes
- **2 pts** You failed to mention your findings

✓ **- 1.5 pts You failed to mention that the decision boundaries are piece-wise linear, which is consistent with the fact that a logistic regression classifier is a linear classifier**

- **0.5 pts** You could have pointed out that linear decision boundaries are well explained by the fact that a logistic regression classifier is a linear classifier
- **0.5 pts** You should have made another relevant comment such as noting that not all classes are present in the plot
- **6 pts** You did not answer the question
- **1 pts** Answer too long/answer box resized
- **2 pts** The plotting range is incorrect and details of decision regions are unclear
- **2 pts** Your findings do not make sense
- **1 pts** Your findings are not specific to the data/result
- **4 pts** You did not include the plot
- **0.5 pts** The plot does not follow the specifications

2.4 2.5 / 4

- **0 pts** Correct
- **4 pts** You did not answer the question
- **2 pts** You did not include the plot
- **2 pts** Your plot of decision regions is very different from the correct one
- **1.5 pts** Your plot of decision regions does not match the correct one very much
- **0.5 pts** Your plot of decision regions has a minor difference from the correct one
- **0.5 pts** Your reported decision regions are correct, but the plot is unclear or insufficiently labelled, e.g. you did not label the axes or used the wrong heatmap (which should be coolwarm)

- **0.5 pts** Your reported decision regions are mostly correct, but the plot is unclear or insufficiently labelled, e.g. you did not label the axes or used the wrong heatmap (which should be coolwarm)
- **2 pts** You failed to mention your findings
- **2 pts** Your findings do not make sense
- **1 pts** Your findings are not specific to the data/result
- **0.5 pts** You could have pointed out that non-linear decision boundaries are well explained by the fact SVM with an RBF kernel is a non-linear classifier.
- **1 pts** Answer too long/answer box resized
- **1.5 pts** The answer does not mention non-linearity, but includes other topics based on the result, such as the appearance of all classes in the plot or any other relevant finding

✓ **- 1.5 pts You failed to mention that the decision boundaries are not linear, which is consistent with the fact that an SVM with a RBF kernel is a non-linear classifier.**

- **1 pts** The plotting range is incorrect and details of decision regions are unclear
- **1 pts** Your plot of decision regions are partially incorrect
- **0.5 pts** Your findings do not include other topics based on the result, such as the appearance of all classes in the plot or any other relevant finding
- **0.5 pts** The plot is unclear or insufficiently labelled, e.g. you did not label the axes or used the wrong heatmap (which should be coolwarm)

2.5 6 / 6

✓ **- 0 pts Correct**

- **1 pts** The value of $$C$$ you reported is different from what is expected
- **1 pts** Your highest mean accuracy is not correct
- **2 pts** Your obtained plot is slightly different from what is expected
- **4 pts** Your obtained plot is totally different from what is expected
- **1 pts** You failed to report the highest mean accuracy
- **1 pts** You failed to report the value of $$C$$
- **6 pts** You did not answer the question
- **4 pts** You failed to include the plot
- **1 pts** You did not use 10 values spaced equally log space
- **1 pts** The plot is unclear or insufficiently labelled, e.g. you did not label the axes
- **1 pts** Answer too long/answer box resized

2.6 3 / 3

✓ **- 0 pts Correct**

- **3 pts** You did not answer the question
- **1.5 pts** The classification accuracy for the training data is not correct
- **1.5 pts** The classification accuracy for the test data is not correct
- **0.5 pts** You reported results with too many digits after the decimal place. Less than four would have been sufficient
- **1 pts** Answer too long/answer box resized

QUESTION 3

Question 3 20 pts

3.1 1 / 3

- **0 pts** Correct
- **3 pts** You did not answer the question
- **1 pts** You failed to report the sum of squared distance

✓ **- 1 pts Your reported sum of squared distance is incorrect**

- **0.5 pts** You reported result with too many digits after the decimal place. Less than one would have been sufficient
- **2 pts** You failed to report the number of samples for each cluster
- **2 pts** Your reported numbers of samples for clusters are largely different from the correct ones

✓ **- 1 pts Your reported numbers of samples for clusters are slightly different from the correct ones**

- **1 pts** Answer too long/answer box resized

3.2 2 / 3

- **0 pts** Correct
- **3 pts** You did not answer the question
- **2 pts** You failed to include the plot
- **1 pts** Your plot of language mean vectors do not match the correct one
- **1 pts** Your plot of cluster centres do not match the correct one

✓ **- 0.5 pts** **The plot is unclear or insufficiently labelled, e.g. you did not label the axes or you did not show the correspondence between each language and the cor- responding mean vector**

- **1 pts** You failed to provide your findings or the findings are completely misleading
- **0.5 pts** Your findings are not based on the plot

✓ **- 0.5 pts** **Your findings include incorrect observations or you didn't provide enough findings**

- **1 pts** Answer too long/answer box resized

3.3 3 / 3

✓ **- 0 pts** **Correct**

- **3 pts** You did not answer the question
- **2 pts** You failed to include the plot
- **2 pts** Your dendrogram is very different from the expected one
- **1 pts** The plot is unclear or insufficiently labelled
- **1 pts** You failed to provide your findings
- **0.5 pts** Your findings are not based on the result obtained
- **0.5 pts** You failed to mention how your findings relate to the result in Q3.2
- **1 pts** Answer too long/answer box resized
- **0.5 pts** Need more details about findings.

3.4 4.5 / 5

- **0 pts** Correct
- **5 pts** You did not answer the question

Plots

- **3 pts** You failed to include any plots
- **1 pts** The plot for the ward linkage is incorrect
- **1 pts** The plot for the single linkage is incorrect
- **1 pts** The plot for the complete linkage is incorrect
- **1 pts** Wrong order of labels
- **1 pts** Labels cannot be read. Use a bigger font, increase the resolution or export the plot in a vectorial format
- **1 pts** The plots are unclear or insufficiently labelled
- **1.5 pts** You were not supposed to truncate the plot
- **1 pts** Wrong labels

Discussions

- **2 pts** You failed to provide discussions or it is wrong
- **1 pts** You failed to describe differences among the three plots
- **1 pts** Your discussions are not based on the result

✓ **- 0.5 pts** **Your discussions lack some theoretical aspects**

- **0.5 pts** You did not write a conclusion or it is wong
- **0.5 pts** Your discussions lack of details
- **1 pts** Answer too long/answer box resized

3.5 5 / 6

- **0 pts** Correct
- **6 pts** You did not answer the question

log-likelihoods

- **2 pts** You failed to report the log-likelihoods
- **0.5 pts** Your result for diag-cov GMM on the training data is not correct. The log-likelihood on training data monotonically increases with the number of mixture components
- **0.5 pts** Your result for diag-cov GMM on the test data is not correct. The log-likelihood on test data monotonically increases with the number of mixture components. The value is close to that of training
- **0.5 pts** Your result for full-cov GMM on the training data is not correct. The log-likelihood on training data monotonically increases with the number of mixture components

✓ **- 0.5 pts** **Your result for full-cov GMM on the test data is not correct. The log-likelihood on test data decreases for K greater than 3**

- **1.5 pts** you did not specify whether reported log-likelihoods are on training or test set. you are supposed to report on both training and test set
- **1 pts** you did not report log-likelihoods of diagonal covariance
- **1.5 pts** you only reported log-likelihood on one set. Further, you did not outline whether reported log-likelihood is on train or test set
- **1 pts** the likelihoods are not clear from the figure

Plot and table - style and format

- **1 pts** You failed to include the plot
- **1 pts** The information presented in the plot does not match the one in the table
- **0.5 pts** Bar-plot with equal spaces on x-axis is not appropriate. Line-plot should have been used instead.
- **0.5 pts** The plot is unclear or insufficiently labelled, e.g. you did not label the axes and you did not show a legend
- **1 pts** You failed to include the table

✓ **- 0.5 pts You reported results with too many digits after the decimal place. Less than two would have been sufficient**

- **0.5 pts** your reported table is not well-formatted or labelled
- **0.5 pts** The plot is unclear or insufficiently labelled, e.g. you did not label the axes
- **0.5 pts** the plot is incomplete you did not specify whether the reported performance is on training or test set
- **0.5 pts** Scatter-plot is not appropriate. Line-plot should have been used instead.
- **0.5 pts** the table is incomplete you did not specify whether the reported performance is on training or test set

Discussions

- **2 pts** You failed to provide discussions
- **1 pts** Your discussions are not based on the result
- **0.5 pts** Your failed to compare the two types of GMMs in your discussions from practical aspects
- **1 pts** you failed to discuss the overfitting that is happening for the full-cov GMM
- **0.5 pts** You failed to compare the two types of GMMs in your discussions from theoretical aspects. The full covariance model has a large number of parameters to train than diag-cov
- **0.5 pts** the test performance for full covariance starts decreasing after K=3. And the gap between train and test starts increasing. The optimal choice for full-cov is K=3
- **1 pts** Answer too long/answer box resized

Question 1 : (30 total points) Image data analysis with PCA

In this question we employ PCA to analyse image data

1.1 (3 points) Once you have applied the normalisation from Step 1 to Step 4 above, report the values of the first 4 elements for the first training sample in `Xtrn_nm`, i.e. `Xtrn_nm[0,:]` and the last training sample, i.e. `Xtrn_nm[-1,:]`.

The first 4 elements for the first training sample in `Xtrn_nm` are as follows,
-3.137e-06, -2.268e-05, -1.180e-04, -4.071e-04
That for the last training sample are as follows,
-3.137e-06, -2.268e-05, -1.180e-04, -4.071e-04

Question 1 continued on next page...

1.1 3 / 3

✓ **- 0 pts** **Correct**

- 3 pts You did not answer the question

- 1.5 pts You reported values for the first training samples are all incorrect. The correct values are -3.14e-06, -2.27e-05, -1.18e-04, -4.07e-04.

- 1 pts You reported values for the first training samples are partially incorrect. The correct values are -3.14e -06, -2.27e-05, -1.18e-04, -4.07e-04.

- 1.5 pts You reported values for the last training samples are incorrect. The correct values are -3.137e-06, -2.268e-05, -1.180e-04, -4.071e-04.

- 1 pts You reported values for the last training samples are partially incorrect. The correct values are -3.14e -06, -2.27e-05, -1.18e-04, -4.07e-04.

- 0.5 pts You reported results with too many/few significant figures. 2 or 3 significant figures would have been sufficient.

- 1 pts Answer too long/answer box resized

1.2 (4 points) Using `Xtrn` and Euclidean distance measure, for each class, find the two closest samples and two furthest samples of that class to the mean vector of the class.

The image is as follows.

The left column represents the mean of each class of data, which can reflect the category of the pictures to a certain extent. At the same time, we find that the pictures at column 2 and column 3 are the most similar to the left ones, but the right two columns of picture have the largest gaps, which is consistent with that the two nearest points on the left and the two furthest points on the right.

1.2 **3.5 / 4**

- **0 pts** Correct
- **4 pts** You did not answer the question
- **1 pts** Your displayed images of the mean vectors are not correct
- **2 pts** Most of the images of samples shown are different from the correct ones
- **1 pts** Some of the images of samples shown are different from the correct ones
- **1 pts** The sample numbers identified are different from the correct ones
- **1 pts** The plot is unclear or insufficiently labelled
- **1.5 pts** You failed to report findings
- **0.5 pts** You failed to provide meaningful discussions
- **1 pts** Answer too long/answer box resized

✓ **- 0.5 pts Furthest samples are very different indicating that they may have been labelled wrongly**

1.3 (3 points) Apply Principal Component Analysis (PCA) to the data of `Xtrn_nm` using sklearn.decomposition.PCA, and report the variances of projected data for the first five principal components in a table. Note that you should use `Xtrn_nm` instead of `Xtrn`.

The table belows shows the results,

	PC1	PC2	PC3	PC4	PC5
Variances	19.810	12.112	4.106	3.382	2.625

1.3 **3 / 3**

✓ **- 0 pts** **Correct**

- 3 pts You did not answer the question

- 3 pts Your reported variances are all incorrect

- 2 pts Your reported variances are partially incorrect

- 0.5 pts You reported results with too many/few digits after the decimal place. Less than four would have been sufficient.

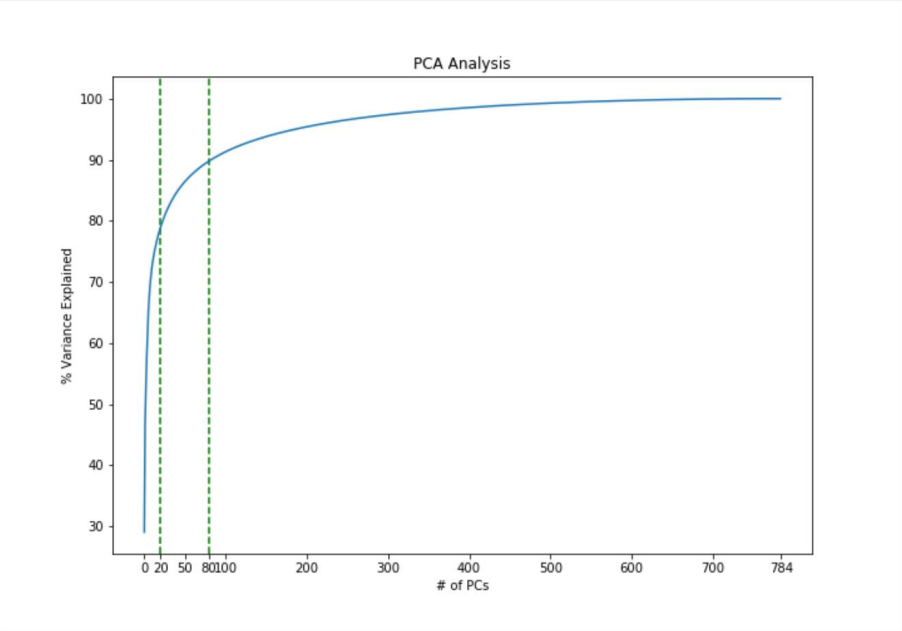
- 1 pts Answer too long/answer box resized

- 2 pts You were expected to report the cumulative explained variance for each component.

- 0.5 pts Wrong or insufficient labelling.

1.4 (3 points) Plot a graph of the cumulative explained variance ratio as a function of the number of principal components, K, where $1 \leq K \leq 784$. Discuss the result briefly.

The image is shown as below,



We can find the curve becomes very smooth when the number of principle components is greater than 100. Additionally, if we want to explain 80% or 90% of the total variance, we need to get approximately 20 and 80 principle components respectively.

1.4 3 / 3

✓ **- 0 pts** Correct

- 3 pts You did not answer the question

- 2 pts You did not include the plot

- 2 pts The graph is not correct

- 1 pts The plot is unclear or insufficiently labelled i.e. you did not label the axes

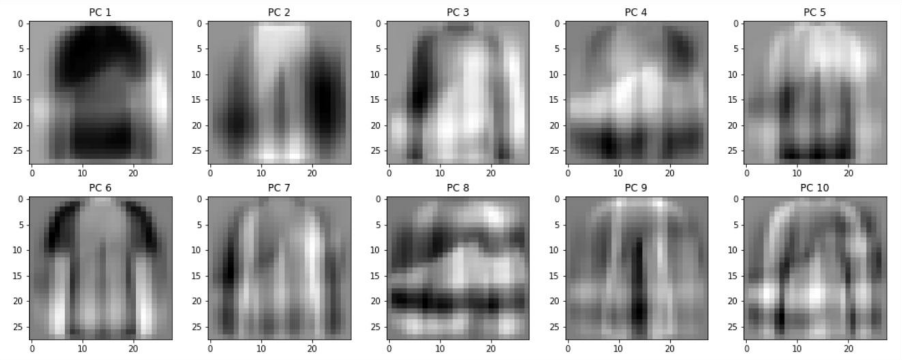
- 1 pts You did not include discussions or your discussions do not make sense.

- 0.5 pts Your discussions are general and not specific to the data

- 1 pts Answer too long/answer box resized

1.5 (4 points) Display the images of the first 10 principal components in a 2-by-5 grid, putting the image of 1st principal component on the top left corner, followed by the one of 2nd component to the right. Discuss your findings briefly.

The image is shown as below,



We find the top-left one is the image for PC1, which is much more informative than others. Additionally, we can see the features of the mixed several classes (we can find the features of clothes, pants, shoes and even bags on each image). Therefore, PCs are representative for the whole data, and if we consider the information provided by all the ten PCs together, we find we know the sorts of the whole dataset to some extent.

1.5 **3.5 / 4**

- **0 pts** Correct
- **4 pts** You did not answer the question
- **2 pts** You did not include the images
- **2 pts** Your displayed images do not match the correct ones
- **1 pts** Some of your displayed images do not match the correct ones
- **0.5 pts** The plot is unclear or insufficiently labelled, e.g. you did not label each image
- **2 pts** You did not include discussions
- **1 pts** Your discussions are not informative or not specific to the data
- **1 pts** Your discussions lack details, e.g. no discussions on different principal com- ponents

✓ **- 0.5 pts Images/captions are too small to read**

- **0.5 pts** Some of your discussions do not make sense
- **1 pts** Answer too long/answer box resized

1.6 (5 points) Using `Xtrn_nm`, for each class and for each number of principal components $K = 5, 20, 50, 200$, apply dimensionality reduction with PCA to the first sample in the class, reconstruct the sample from the dimensionality-reduced sample, and report the Root Mean Square Error (RMSE) between the original sample in `Xtrn_nm` and reconstructed one.

The table below shows the results,

	K=5	K=20	K=50	K=200
class0	0.256	0.150	0.128	0.061
class1	0.198	0.140	0.095	0.036
class2	0.199	0.146	0.124	0.080
class3	0.146	0.107	0.083	0.056
class4	0.118	0.103	0.088	0.048
class5	0.181	0.159	0.143	0.090
class6	0.129	0.096	0.072	0.047
class7	0.166	0.128	0.107	0.063
class8	0.223	0.145	0.124	0.092
class9	0.184	0.151	0.121	0.072

1.6 5 / 5

✓ **- 0 pts** Correct

- 5 pts You did not answer the question

- 5 pts Your reported RMSE values are all incorrect

- 4 pts Most of the values of RMSE are incorrect

- 3 pts More than the half of the values of RMSE are incorrect

- 2 pts Some of the values of RMSE are not correct

- 1 pts The answer is not presented in the specified table format

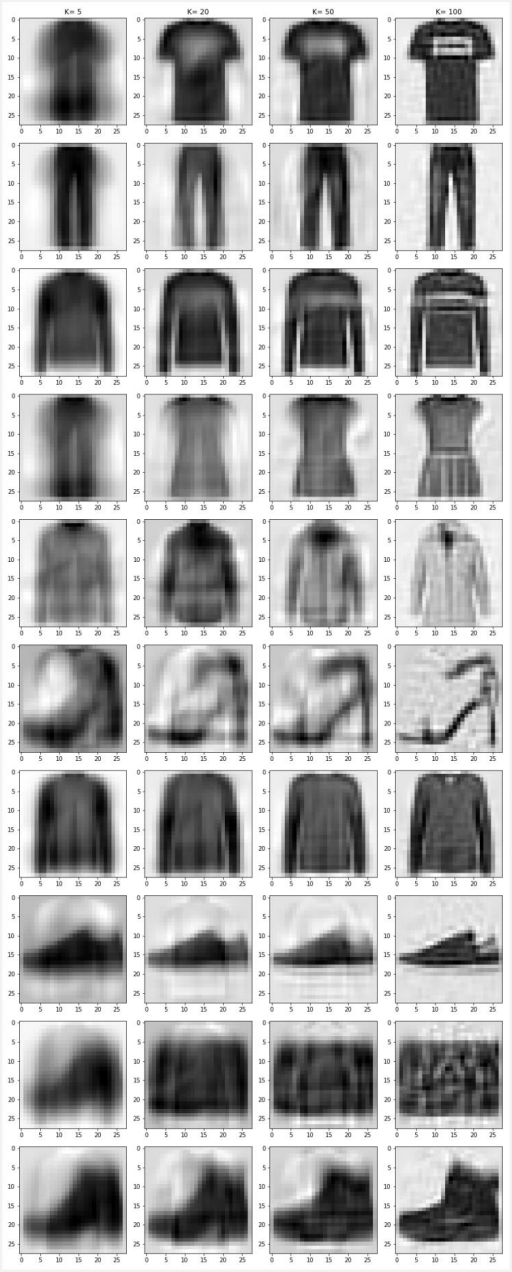
- 1 pts You reported results with too many digits after the decimal place. Less than four would have been sufficient.

- 1 pts Answer too long/answer box resized

- 1 pts There is a significant typo or error in the way (otherwise correct) values are reported.

1.7 (4 points) Display the image for each of the reconstructed samples in a 10-by-4 grid, where each row corresponds to a class and each row column corresponds to a value of $K = 5,\ 20,\ 50,\ 200$.

The image is shown as below,



We know that with the increase of the value of K, it means that the degree of dimensional reduction is weakened, so that the information of the original data can be explained more fully. From the image, we can find that when K is 5, we can only roughly understand the category of the picture. As for shoes, K=5 performs worse(with the white background like clothes), but with the increase of K, the picture on the right shows the details of the picture more and more clearly.

1.7 **2.5 / 4**

- **0 pts** Correct
- **4 pts** You did not answer the question
- **2.5 pts** You failed to include plots
- **2.5 pts** All the images are incorrect
- **2 pts** More than the half of the images are incorrect
- **1 pts** Some of the images are incorrect

✓ - **0.5 pts** **The plot is unclear or insufficiently labelled i.e. you did not label the axes**

- **1.5 pts** You failed to include discussions, or your description is too short/inaccurate.
- **1 pts** Your discussions are general and not specific to the data.

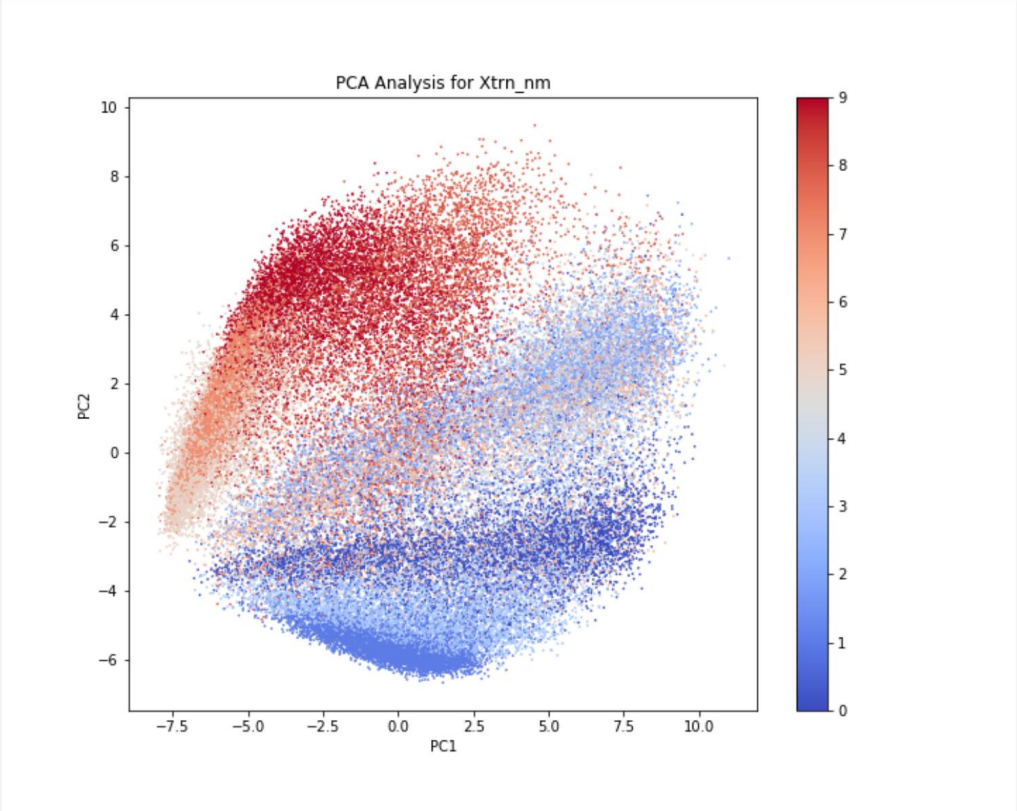
✓ - **1 pts** **Your discussion does not fully relate to your results from previous questions.**

- **0.5 pts** Some of your discussions do not make sense
- **1 pts** Your discussion is too short - you should relate to your results from previous questions, and talk more specifically about the dataset.
- **1 pts** Answer too long/answer box resized

No points taken for this, but the axis ticks for each image are unnecessary here.

1.8 (4 points) Plot all the training samples (`Xtrn_nm`) on the two-dimensional PCA plane you obtained in Question 1.3, where each sample is represented as a small point with a colour specific to the class of the sample. Use the 'coolwarm' colormap for plotting.

The image is shown as below,



As for the class separation, the scatter points of class 0, class 2, class3, class 4 and class 6 are much clustered since they are all clothes(no matter for male or female). Class 1 is at the bottom of the image and is for pants. Class 8 is at the top of the image and is for bags. The remain 3 clusters, for class 5, class 7 and class 9, are much aggregated at the left-top of the image and all for shoes.
To some extent, the results are reasonable since classes of similar sorts are easier to form larger clusters(more similarities). Overall, we think the two-dimensional PCA plane can contently explain the data.

1.8 **3.5 / 4**

- **0 pts** Correct
- **4 pts** You did not answer the question
- **3 pts** You failed to include the plot
- **3 pts** Your plot is totally different from the expected plot
- **2 pts** Your plot is very different from the expected one
- **1 pts** Your plot is slightly different from the expected plot
- **1 pts** The plot is unclear or insufficiently labelled, e.g. you did not label the axes.
- **0.5 pts** You failed to comment on the separation of classes
- **0.5 pts** You failed to explain your findings
- **0.5 pts** Your findings are not specific to the data

✓ **- 0.5 pts** **You should have included a bit more discussion.**

- **1 pts** Answer too long/answer box resized

Question 2 : (25 total points) Logistic regression and SVM

In this question we will explore classification of image data with logistic regression and support vector machines (SVM) and visualisation of decision regions.

2.1 (3 points) Carry out a classification experiment with multinomial logistic regression, and report the classification accuracy and confusion matrix (in numbers rather than in graphical representation such as heatmap) for the test set.

The classification accuracy for this logistic regression classifier is equal to 0.8401. The confustion matrix is as follows,

$$\begin{bmatrix} 819 & 3 & 15 & 50 & 7 & 4 & 90 & 1 & 11 & 0 \\ 5 & 953 & 4 & 27 & 5 & 0 & 3 & 1 & 2 & 0 \\ 27 & 4 & 731 & 11 & 133 & 0 & 82 & 2 & 9 & 1 \\ 31 & 15 & 14 & 866 & 33 & 0 & 37 & 0 & 4 & 0 \\ 0 & 3 & 115 & 38 & 760 & 2 & 72 & 0 & 10 & 0 \\ 2 & 0 & 0 & 1 & 0 & 911 & 0 & 56 & 10 & 20 \\ 147 & 3 & 128 & 46 & 108 & 0 & 539 & 0 & 28 & 1 \\ 0 & 0 & 0 & 0 & 0 & 32 & 0 & 936 & 1 & 31 \\ 7 & 1 & 6 & 11 & 3 & 7 & 15 & 5 & 945 & 0 \\ 0 & 0 & 0 & 1 & 0 & 15 & 1 & 42 & 0 & 941 \end{bmatrix}$$

2.1 **2.5 / 3**

- 0 pts Correct

- 3 pts You did not answer the question

- 1 pts You failed to report the classification accuracy

- 1 pts Your reported classification accuracy is different from the expected one (0.8401)

- 0.5 pts You reported result with too many digits after the decimal place. Less than five would have been sufficient.

- 2 pts You failed to report the confusion matrix

- 2 pts Your reported confusion matrix is very different from the expected one

- 1 pts Your reported confusion matrix is slightly different from the expected one

✓ **- 0.5 pts Your confusion matrix is unclear, e.g. unclear what information each element represents**

- 1 pts Answer too long/answer box resized

2.2 (3 points) Carry out a classification experiment with SVM classifiers, and report the mean accuracy and confusion matrix (in numbers) for the test set.

The classification accuracy for this SVM classifier is equal to 0.8461.
The confustion matrix is as follows,

$$\begin{bmatrix} 845 & 2 & 8 & 51 & 4 & 4 & 72 & 0 & 14 & 0 \\ 4 & 951 & 7 & 31 & 5 & 0 & 1 & 0 & 1 & 0 \\ 15 & 2 & 748 & 11 & 137 & 0 & 79 & 0 & 8 & 0 \\ 32 & 6 & 12 & 881 & 26 & 0 & 40 & 0 & 3 & 0 \\ 1 & 0 & 98 & 36 & 775 & 0 & 86 & 0 & 4 & 0 \\ 0 & 0 & 0 & 1 & 0 & 914 & 0 & 57 & 2 & 26 \\ 185 & 1 & 122 & 39 & 95 & 0 & 533 & 0 & 25 & 0 \\ 0 & 0 & 0 & 0 & 0 & 34 & 0 & 925 & 0 & 41 \\ 3 & 1 & 8 & 5 & 2 & 4 & 13 & 4 & 959 & 1 \\ 0 & 0 & 0 & 0 & 0 & 22 & 0 & 47 & 1 & 930 \end{bmatrix}$$

Question 2 continued on next page...

2.2 **2.5 / 3**

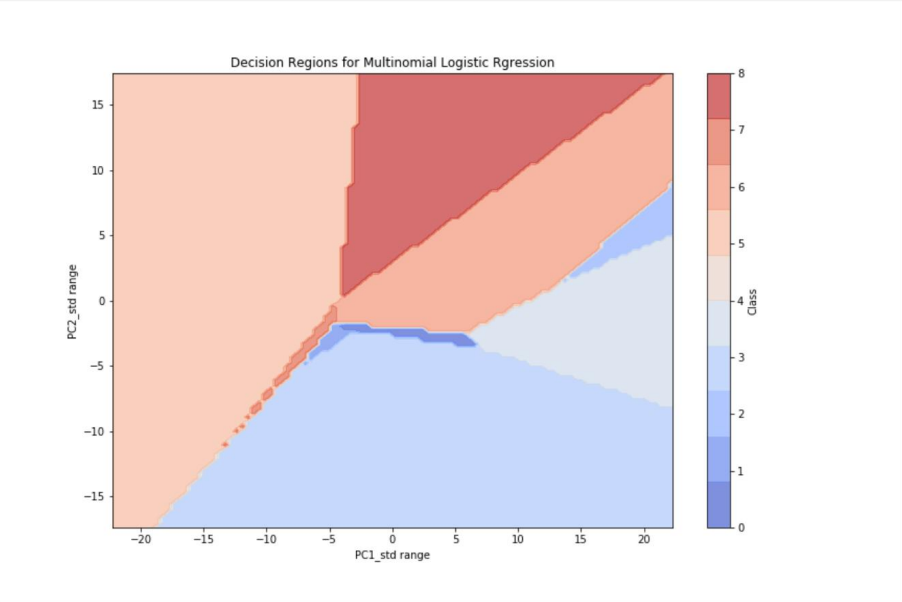
- **0 pts** Correct
- **3 pts** You did not answer the question
- **1 pts** You failed to report the classification accuracy
- **1 pts** Your reported classification accuracy is different from the expected one (0.8461)
- **0.5 pts** You reported result with too many digits after the decimal place. Less than five would have been sufficient.
- **2 pts** You failed to report the confusion matrix
- **2 pts** Your reported confusion matrix is very different from the expected one
- **1 pts** Your reported confusion matrix is slightly different from the expected one

✓ - **0.5 pts** **Your confusion matrix is unclear, e.g. unclear what information each element represents**

- **1 pts** Answer too long/answer box resized

2.3 (6 points) We now want to visualise the decision regions for the logistic regression classifier we trained in Question 2.1.

The image is shown as below,



From the image, we can see 9 discrete colors concerning each class, and the reason for not 10 colors is that we do not get predicted z for class 9 using trained model in Q2.1.
Overall, we find the regions are clear, but not accurate enough since some areas are extremely large or small, and no color for class 9.

2.3 4.5 / 6

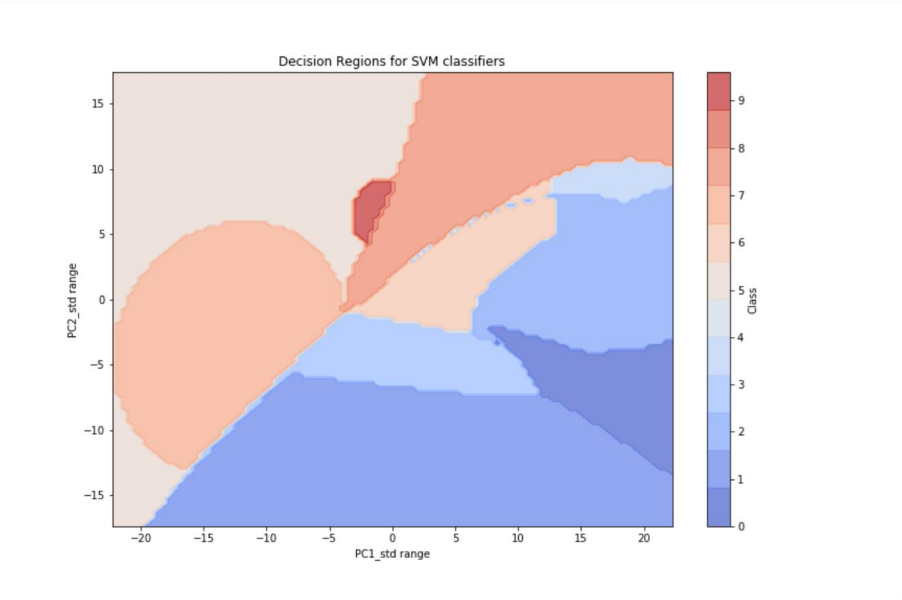
- **0 pts** Correct
- **4 pts** Your plot of decision regions is very different from the correct one
- **3 pts** Your plot of decision regions does not match the correct one very much
- **1 pts** Your plot of decision regions has a minor difference from the correct one
- **0.5 pts** The plot is unclear or insufficiently labelled, e.g. you did not label the axes
- **2 pts** You failed to mention your findings

✓ **- 1.5 pts** **You failed to mention that the decision boundaries are piece-wise linear, which is consistent with the fact that a logistic regression classifier is a linear classifier**

- **0.5 pts** You could have pointed out that linear decision boundaries are well explained by the fact that a logistic regression classifier is a linear classifier
- **0.5 pts** You should have made another relevant comment such as noting that not all classes are present in the plot
- **6 pts** You did not answer the question
- **1 pts** Answer too long/answer box resized
- **2 pts** The plotting range is incorrect and details of decision regions are unclear
- **2 pts** Your findings do not make sense
- **1 pts** Your findings are not specific to the data/result
- **4 pts** You did not include the plot
- **0.5 pts** The plot does not follow the specifications

2.4 (4 points) Using the same method as the one above, plot the decision regions for the SVM classifier you trained in Question 2.2. Comparing the result with that you obtained in Question 2.3, discuss your findings briefly.

The image is shown as below,



From the image, we can see 10 discrete colors concerning each class using the same projected point in Q2.3 under the classifier model in Q2.2.
Compared with Q2.3, the results are much more satisfying. Firstly, we have the color for class 9 which implies the projected points are better used to some extent. Secondly, the color area distribution is much more even, and we can find some extreme area in Q2.3 are modified to some extent.
Overall, in this case, SVM classifier performs better in gaining decision regions than multinomial logistic regression model.

2.4 2.5 / 4

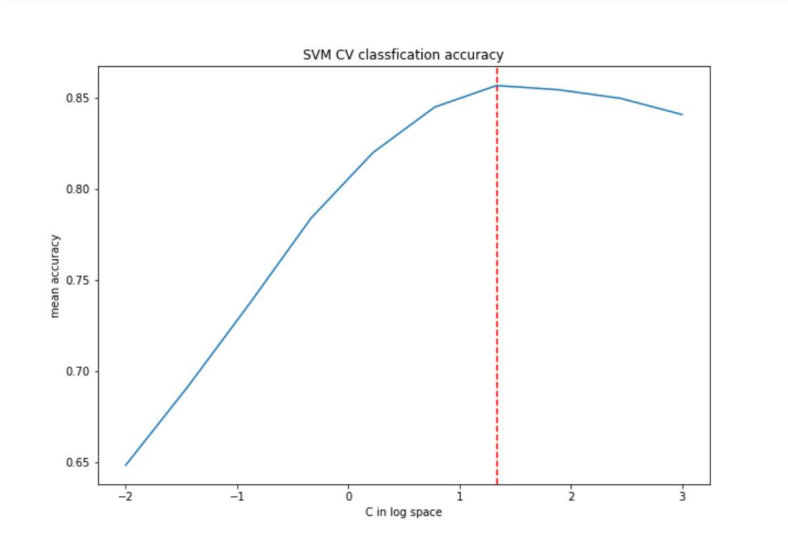
- **0 pts** Correct
- **4 pts** You did not answer the question
- **2 pts** You did not include the plot
- **2 pts** Your plot of decision regions is very different from the correct one
- **1.5 pts** Your plot of decision regions does not match the correct one very much
- **0.5 pts** Your plot of decision regions has a minor difference from the correct one
- **0.5 pts** Your reported decision regions are correct, but the plot is unclear or insufficiently labelled, e.g. you did not label the axes or used the wrong heatmap (which should be coolwarm)
- **0.5 pts** Your reported decision regions are mostly correct, but the plot is unclear or insufficiently labelled, e.g. you did not label the axes or used the wrong heatmap (which should be coolwarm)
- **2 pts** You failed to mention your findings
- **2 pts** Your findings do not make sense
- **1 pts** Your findings are not specific to the data/result
- **0.5 pts** You could have pointed out that non-linear decision boundaries are well explained by the fact SVM with an RBF kernel is a non-linear classifier.
- **1 pts** Answer too long/answer box resized
- **1.5 pts** The answer does not mention non-linearity, but includes other topics based on the result, such as the appearance of all classes in the plot or any other relevant finding

✓ **- 1.5 pts** **You failed to mention that the decision boundaries are not linear, which is consistent with the fact that an**
SVM with a RBF kernel is a non-linear classifier.

- **1 pts** The plotting range is incorrect and details of decision regions are unclear
- **1 pts** Your plot of decision regions are partially incorrect
- **0.5 pts** Your findings do not include other topics based on the result, such as the appearance of all classes in the plot or any other relevant finding
- **0.5 pts** The plot is unclear or insufficiently labelled, e.g. you did not label the axes or used the wrong heatmap (which should be coolwarm)

2.5 (6 points) We used default parameters for the SVM in Question 2.2. We now want to tune the parameters by using cross-validation. To reduce the time for experiments, you pick up the first 1000 training samples from each class to create `Xsmall`, so that `Xsmall` contains 10,000 samples in total. Accordingly, you create labels, `Ysmall`.

We use np.argmax() to gain the position of C which leads to the maximum accuracy score, and we plot it by a red line on the picture.
The image is shown as below,



We get the maximum accuracy score is equal to 0.857 with the corresponding C=$\frac{4}{3}$(in log space).

2.5 6 / 6

✓ **- 0 pts** Correct

- 1 pts The value of $$C$$ you reported is different from what is expected

- 1 pts Your highest mean accuracy is not correct

- 2 pts Your obtained plot is slightly different from what is expected

- 4 pts Your obtained plot is totally different from what is expected

- 1 pts You failed to report the highest mean accuracy

- 1 pts You failed to report the value of $$C$$

- 6 pts You did not answer the question

- 4 pts You failed to include the plot

- 1 pts You did not use 10 values spaced equally log space

- 1 pts The plot is unclear or insufficiently labelled, e.g. you did not label the axes

- 1 pts Answer too long/answer box resized

2.6 (3 points) Train the SVM classifier on the whole training set by using the optimal value of C you found in Question 2.5.

From Q2.5, we set C=$\frac{4}{3}$(in log space) for this question. After training the model, we gain the classification accuracy for the test set under this SVM classifier is equal to 0.850, and that for the training set is equal to 0.868.

2.6 3 / 3

✓ **- 0 pts** **Correct**

- 3 pts You did not answer the question

- 1.5 pts The classification accuracy for the training data is not correct

- 1.5 pts The classification accuracy for the test data is not correct

- 0.5 pts You reported results with too many digits after the decimal place. Less than four would have been sufficient

- 1 pts Answer too long/answer box resized

Question 3 : (20 total points) Clustering and Gaussian Mixture Models

In this question we will explore K-means clustering, hierarchical clustering, and GMMs.

3.1 (3 points) Apply k-means clustering on `Xtrn` for $k = 22$, where we use sklearn.cluster.KMeans with the parameters `n_clusters=22` and `random_state=1`. Report the sum of squared distances of samples to their closest cluster centre, and the number of samples for each cluster.

The sum of squared distances of samples to their closest cluster centre is equal to 38189.99.
The number of samples for each cluster is as follows,

label	0	1	2	3	4	5	6	7	8	9	10
count	1019	1131	1178	883	1156	1294	855	623	1395	840	660
label	11	12	13	14	15	16	17	18	19	20	21
count	1276	121	152	964	1976	1243	848	885	930	1091	1480

Question 3 continued on next page...

3.1 **1 / 3**

- **0 pts** Correct
- **3 pts** You did not answer the question
- **1 pts** You failed to report the sum of squared distance

✓ - **1 pts** **Your reported sum of squared distance is incorrect**

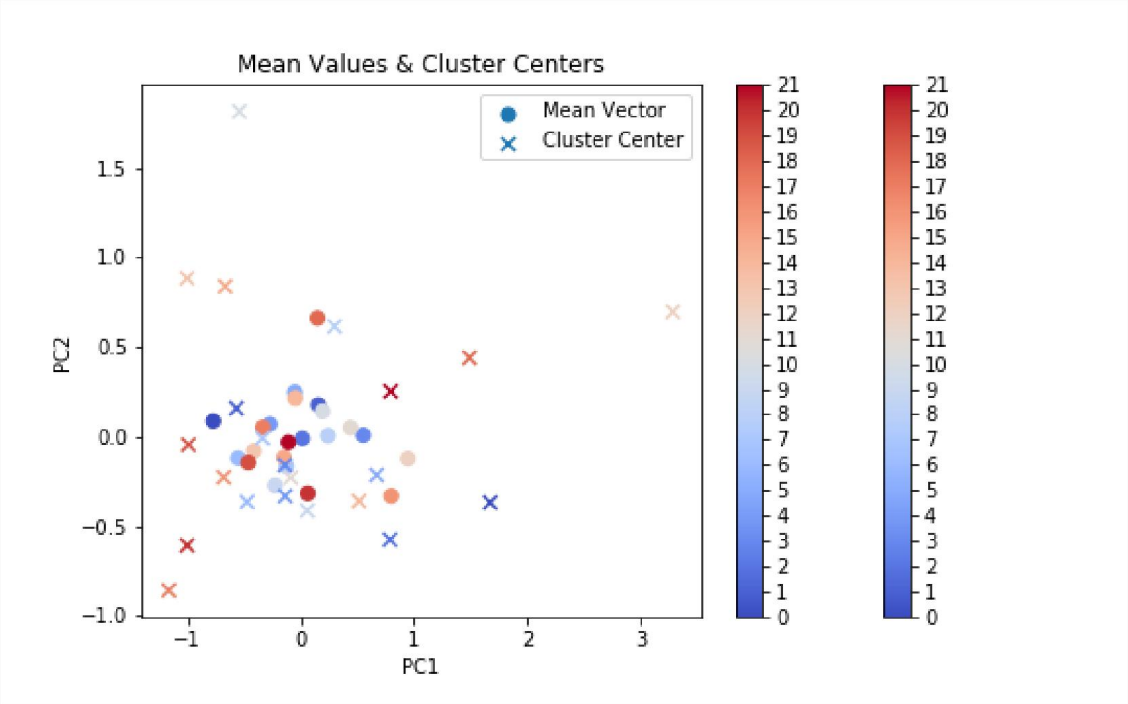
- **0.5 pts** You reported result with too many digits after the decimal place. Less than one would have been sufficient
- **2 pts** You failed to report the number of samples for each cluster
- **2 pts** Your reported numbers of samples for clusters are largely different from the correct ones

✓ - **1 pts** **Your reported numbers of samples for clusters are slightly different from the correct ones**

- **1 pts** Answer too long/answer box resized

3.2 (3 points) Using the training set only, calculate the mean vector for each language, and plot the mean vectors of all the 22 languages on a 2D-PCA plane, where you apply PCA on the set of 22 mean vectors without applying standardisation. On the same figure, plot the cluster centres obtained in Question 3.1.

The image is shown as below,



We find the mean vectors and cluster centres are not similar.
According to the image, cluster centers are much dispersed since they are trained by k-means model, and some clusters may obtain much data of different languages. Yet for the mean vectors, they are much aggregated, a possible explanation is that many languages have similarities and they can not be separated clearly.

3.2 **2 / 3**

- **0 pts** Correct
- **3 pts** You did not answer the question
- **2 pts** You failed to include the plot
- **1 pts** Your plot of language mean vectors do not match the correct one
- **1 pts** Your plot of cluster centres do not match the correct one

✓ **- 0.5 pts** **The plot is unclear or insufficiently labelled, e.g. you did not label the axes or you did not show the correspondence between each language and the cor- responding mean vector**

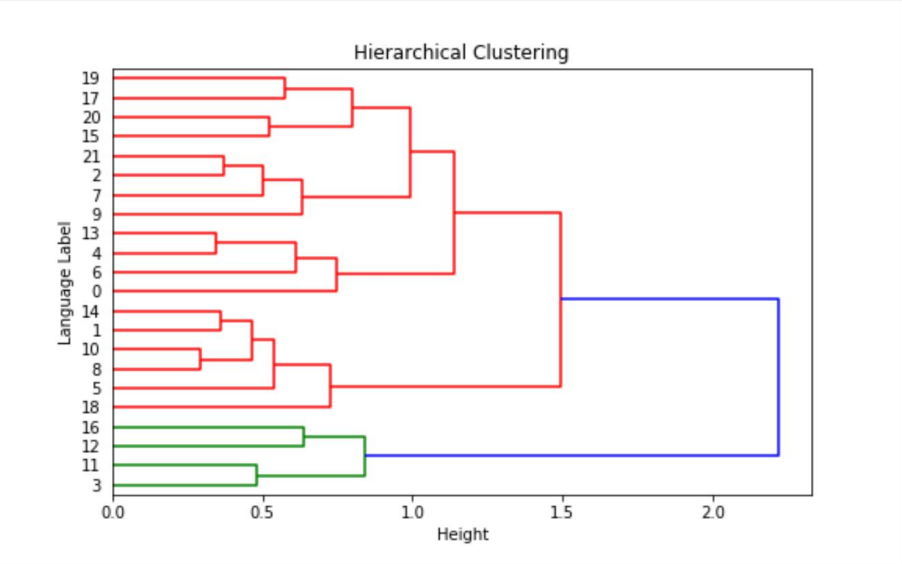
- **1 pts** You failed to provide your findings or the findings are completely misleading
- **0.5 pts** Your findings are not based on the plot

✓ **- 0.5 pts** **Your findings include incorrect observations or you didn't provide enough findings**

- **1 pts** Answer too long/answer box resized

3.3 (3 points) We now apply hierarchical clustering on the training data set to see if there are any structures in the spoken languages.

The image is shown as below,



We find the languages belonging to green branches are greatly different from that of red branches. The red sub-tree is is much more complex which implies such 18 languages have more complicated relations. Additionally, the basic combination (2 languages into 1 leaf) means such two languages are very similar, such as label 16 and label 12 (Slovenian (Sl) and Latvian (Lv)).
However, the mean vector is not accurate enough to represent the whole language data, a not satisfying case is label 19 and label 17 (Tamil (Ta) and Swedish (Sv)). Maybe these two language are not strongly related.

3.3 3 / 3

✓ **- 0 pts** **Correct**

- 3 pts You did not answer the question

- 2 pts You failed to include the plot

- 2 pts Your dendrogram is very different from the expected one

- 1 pts The plot is unclear or insufficiently labelled

- 1 pts You failed to provide your findings

- 0.5 pts Your findings are not based on the result obtained

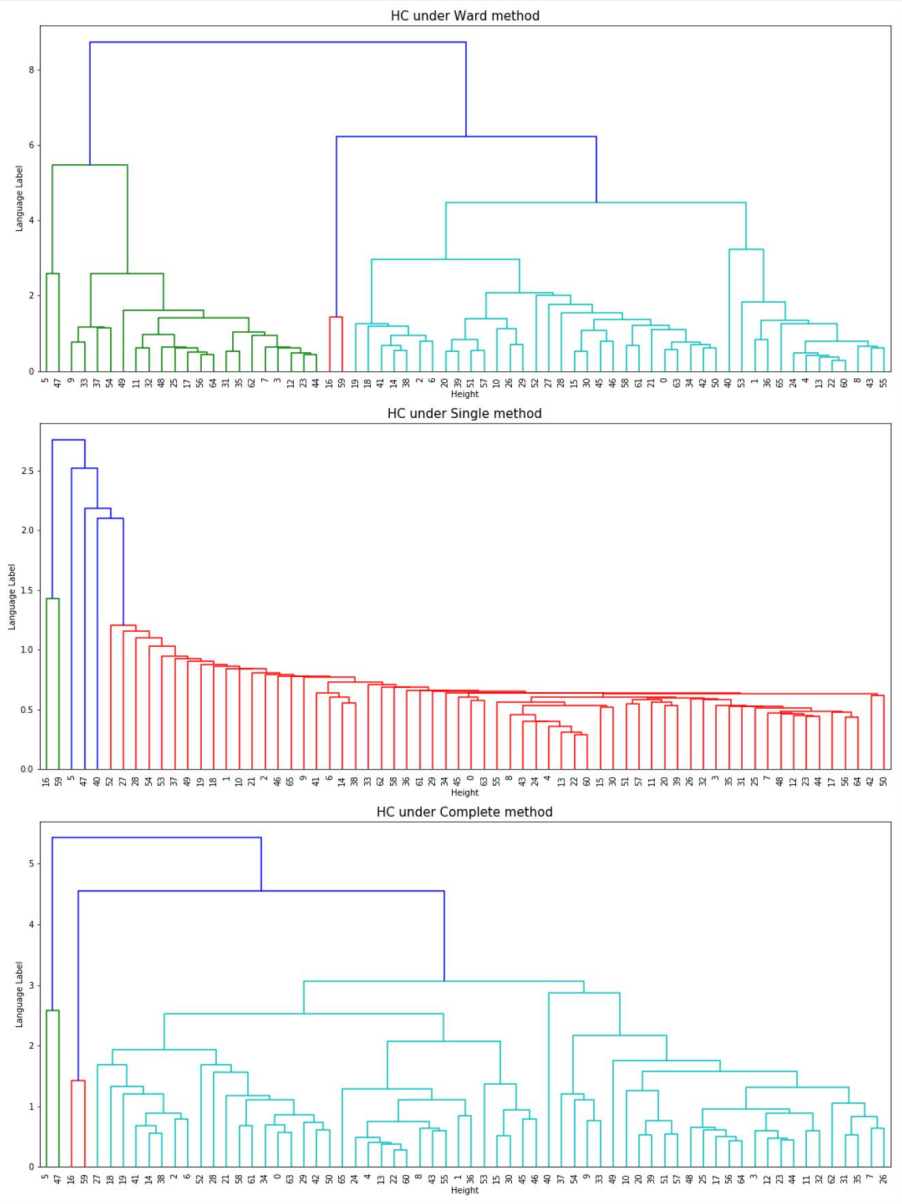
- 0.5 pts You failed to mention how your findings relate to the result in Q3.2

- 1 pts Answer too long/answer box resized

- 0.5 pts Need more details about findings.

3.4 (5 points) We here extend the hierarchical clustering done in Question 3.3 by using multiple samples from each language.

The image is shown as below,



These three images are much more informative than the image in Q3.3. To some extent, single method and complete method are not accurate enough, since the results are too extreme (color areas are extremely uneven) and not corresponding to the the real data(0,1,2 should be close, 3,4,5 should be close and so on).
Although the ward method is also not totally accurate because we apply k-means to each language, it is better than other two methods.

3.4 4.5 / 5

- **0 pts** Correct
- **5 pts** You did not answer the question

Plots

- **3 pts** You failed to include any plots
- **1 pts** The plot for the ward linkage is incorrect
- **1 pts** The plot for the single linkage is incorrect
- **1 pts** The plot for the complete linkage is incorrect
- **1 pts** Wrong order of labels
- **1 pts** Labels cannot be read. Use a bigger font, increase the resolution or export the plot in a vectorial format
- **1 pts** The plots are unclear or insufficiently labelled
- **1.5 pts** You were not supposed to truncate the plot
- **1 pts** Wrong labels

Discussions

- **2 pts** You failed to provide discussions or it is wrong
- **1 pts** You failed to describe differences among the three plots
- **1 pts** Your discussions are not based on the result

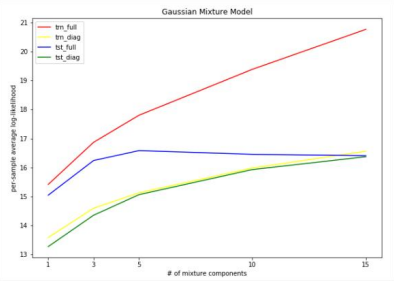
✓ **- 0.5 pts Your discussions lack some theoretical aspects**

- **0.5 pts** You did not write a conclusion or it is wong
- **0.5 pts** Your discussions lack of details

- **1 pts** Answer too long/answer box resized

3.5 (6 points) We now consider Gaussian mixture model (GMM), whose probability distribution function (pdf) is given as a linear combination of Gaussian or normal distributions, i.e.,

The image and table are shown as follows,



	trn_full	trn_diag	tst_full	tst_diag
K=1	15.418	13.588	15.047	13.275
K=3	16.870	14.598	16.242	14.356
K=5	17.803	15.131	16.583	15.065
K=10	19.387	15.984	16.456	15.931
K=15	20.763	16.561	16.409	16.373

From the results above, we find the values under 'full covariance' are mostly larger than that under 'diagonal covariance'. It is reasonable since 'full' allows each cluster to be modeled with any direction.
However, we think 'full' is not great for this GMM. The problem is kind of overfitting. According to the image, we find the yellow line and green line are smooth and close. Nevertheless, the red line and blue line are greatly different when K is greater than 3, and `tst_full` even declines as K is greater than 5.
Overall, we think the model trained by the training set under 'diagonal coveriance' is better, since it is a good trade off between complexity and performance i.e. it requires less directions for mixture.

3.5 5 / 6

- **0 pts** Correct
- **6 pts** You did not answer the question

log-likelihoods

- **2 pts** You failed to report the log-likelihoods
- **0.5 pts** Your result for diag-cov GMM on the training data is not correct. The log-likelihood on training data monotonically increases with the number of mixture components
- **0.5 pts** Your result for diag-cov GMM on the test data is not correct. The log-likelihood on test data monotonically increases with the number of mixture components. The value is close to that of training
- **0.5 pts** Your result for full-cov GMM on the training data is not correct. The log-likelihood on training data monotonically increases with the number of mixture components

✓ **- 0.5 pts Your result for full-cov GMM on the test data is not correct. The log-likelihood on test data decreases for K greater than 3**

- **1.5 pts** you did not specify whether reported log-likelihoods are on training or test set. you are supposed to report on both training and test set
- **1 pts** you did not report log-likelihoods of diagonal covariance
- **1.5 pts** you only reported log-likelihood on one set. Further, you did not outline whether reported log-likelihood is on train or test set
- **1 pts** the likelihoods are not clear from the figure

Plot and table - style and format

- **1 pts** You failed to include the plot
- **1 pts** The information presented in the plot does not match the one in the table
- **0.5 pts** Bar-plot with equal spaces on x-axis is not appropriate. Line-plot should have been used instead.
- **0.5 pts** The plot is unclear or insufficiently labelled, e.g. you did not label the axes and you did not show a legend
- **1 pts** You failed to include the table

✓ **- 0.5 pts You reported results with too many digits after the decimal place. Less than two would have been sufficient**

- **0.5 pts** your reported table is not well-formatted or labelled
- **0.5 pts** The plot is unclear or insufficiently labelled, e.g. you did not label the axes
- **0.5 pts** the plot is incomplete you did not specify whether the reported performance is on training or test set
- **0.5 pts** Scatter-plot is not appropriate. Line-plot should have been used instead.
- **0.5 pts** the table is incomplete you did not specify whether the reported performance is on training or test set

Discussions

- **2 pts** You failed to provide discussions
- **1 pts** Your discussions are not based on the result
- **0.5 pts** Your failed to compare the two types of GMMs in your discussions from practical aspects
- **1 pts** you failed to discuss the overfitting that is happening for the full-cov GMM
- **0.5 pts** You failed to compare the two types of GMMs in your discussions from theoretical aspects. The full covariance model has a large number of parameters to train than diag-cov
- **0.5 pts** the test performance for full covariance starts decreasing after K=3. And the gap between train

and test starts increasing. The optimal choice for full-cov is K=3

- 1 pts Answer too long/answer box resized