IAML (Level 10) Assignment 2

Huacheng Song

TOTAL POINTS

71 / 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 2 / 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 / 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 4 / 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 / 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 6 / 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 / 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 3 / 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 2.5 / 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 5.5 / 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`: [-3.13725490e-06, -2.26797386e-05, -1.17973856e-04, -4.07058824e-04]
The last 4 elements for the first training sample in `Xtrn_nm`: [-3.13725490e-06, -2.26797386e-05, -1.17973856e-04, -4.07058824e-04]

	1stElements	2ndElements	3rdElements	4thElements
First Training Sample	-3.14e-06	-2.27e-05	-1.18e-04	-4.07e-04
Last Training Sample	-3.14e-06	-2.27e-05	-1.18e-04	-4.07e-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.

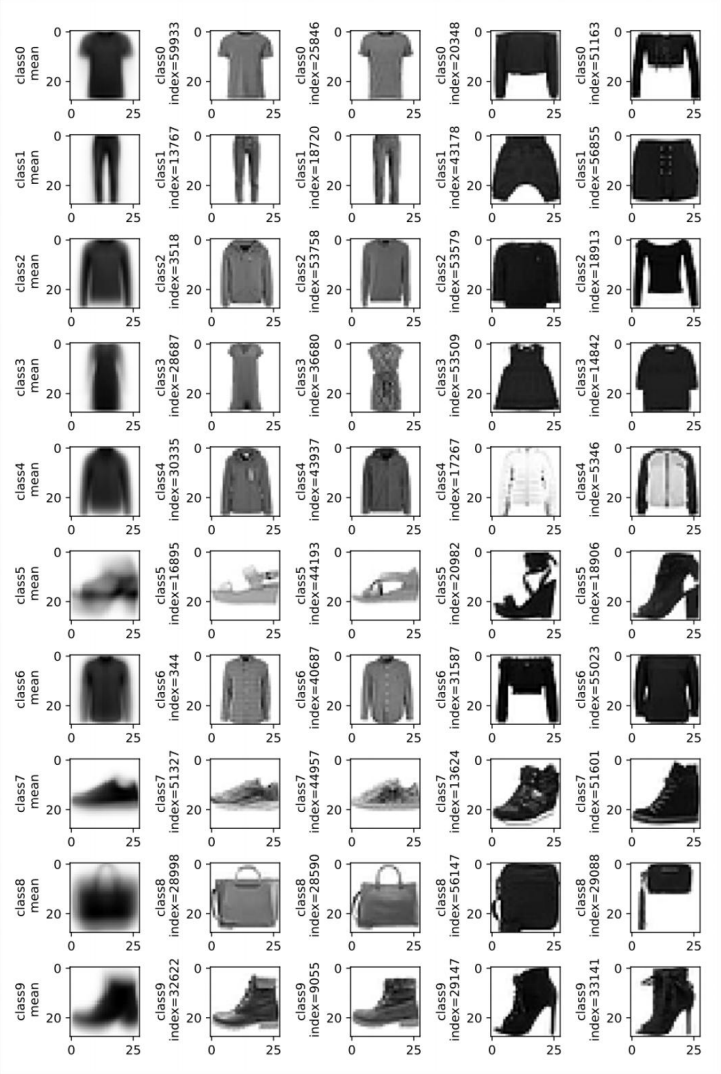
- 1.5 pts You reported values for the last training samples are incorrect. The correct values are -3.137e-06, -2.268e-05, -1.180e-04, -4.071e-04.

- 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.



Images of each mean vector is blurry and represent main and general features which are constant among all the images belong to each class. The 1st and 2nd closest images are the first two typical images of each class and they have similar shape and size to the mean image, which means they describe the main feature and represent the majority of that class). The 1st and 2nd furthest images do not have main and general features of that class and represent minority. Also in the same class there are some difference of color and orientation among different images. For example, the second furthest image in class 7 has different orientation compared with other data in the same class and this is a reason why it is the second furthest image. In class 4, last 2 images have lighter color compared with the other samples in class 4. We need to ensure that orientation, size and brightness of the picture are consistent as much as possible in order to get better results.

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`.

Principal components	1st	2nd	3rd	4th	5th
Explained Variance	19.81	12.11	4.11	3.38	2.62

(Explained Variance rounded in 2 decimals)

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

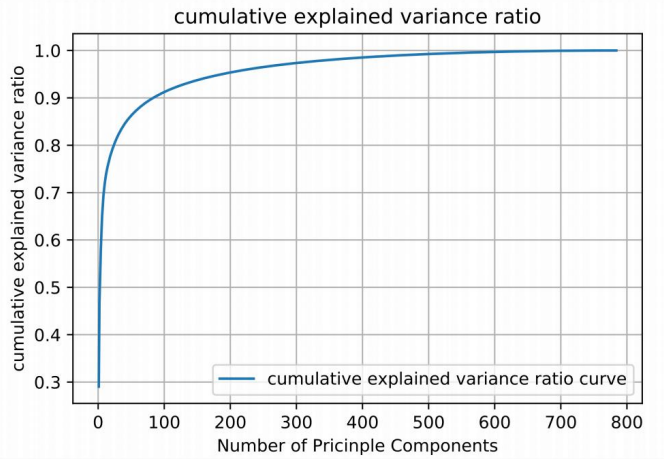
- 0.5 pts You reported results with too many/few digits after the decimal place. Less than four would have been sufficient.

- 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.



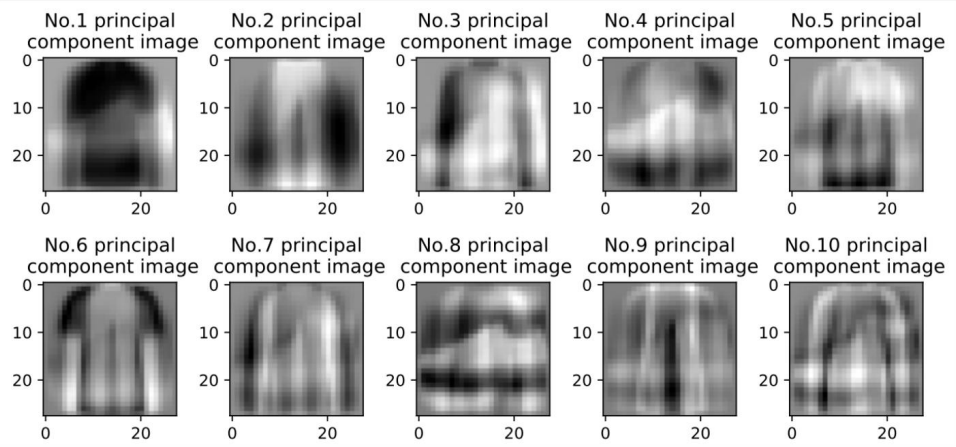
Cumulative variance ratio rises rapidly at first because principle components are sorted by the eigenvalues from big to small(means the importance of principle components from big to small) and we firstly add those more important eigenvectors. The cumulative variance ratio increases slower and slower in the rest principle components as they are less important. When $K = 83$, the cumulative explained variance ratio is over 90% and this means the principal components explain at least 90% of the variance. Remaining principle components have only make less than 10% contribution to the variance. When $K = 23$, ratio is over 80% and when $K = 8$, ratio is over 70%. This means that the cumulative variance in the first few principle components has taken up the majority of total variance.

1.4 **2 / 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.



Images of the first 10 principal components are sorted by the variances(eigenvalues) which corresponds with the importance. The first image has the highest variances and it is most important. We could clearly find there is a long sleeve(color in black) and shoe(color in white) in the first principal component image. The first few images are used to distinguish categories(Clothes patterns in those images are clear) because they have larger variance and the variances between different classes are quite large. The first few images contain dark or white features that are similar to mean images of some classes. The last few images are blurry and have lower variances, because they are less important than the first few images. Hence the patterns shown in those images are more blurry. With these eigenvectors it is possible to redraw any of the image on the data set by executing transform of the PCA object to get the eigenvectors out, and then apply inverse_transform on the eigenvectors to get all the original images.

1.5 **4 / 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.

	RMSE(K=5)	RMSE(K=20)	RMSE(K=50)	RMSE(K=200)
Class0	0.256	0.150	0.127	0.061
Class1	0.198	0.140	0.095	0.038
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	0.047
Class5	0.181	0.159	0.143	0.089
Class6	0.129	0.096	0.072	0.046
Class7	0.166	0.128	0.107	0.064
Class8	0.223	0.145	0.124	0.091
Class9	0.184	0.151	0.122	0.072

(RMSE rounded in 3 decimal places)

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

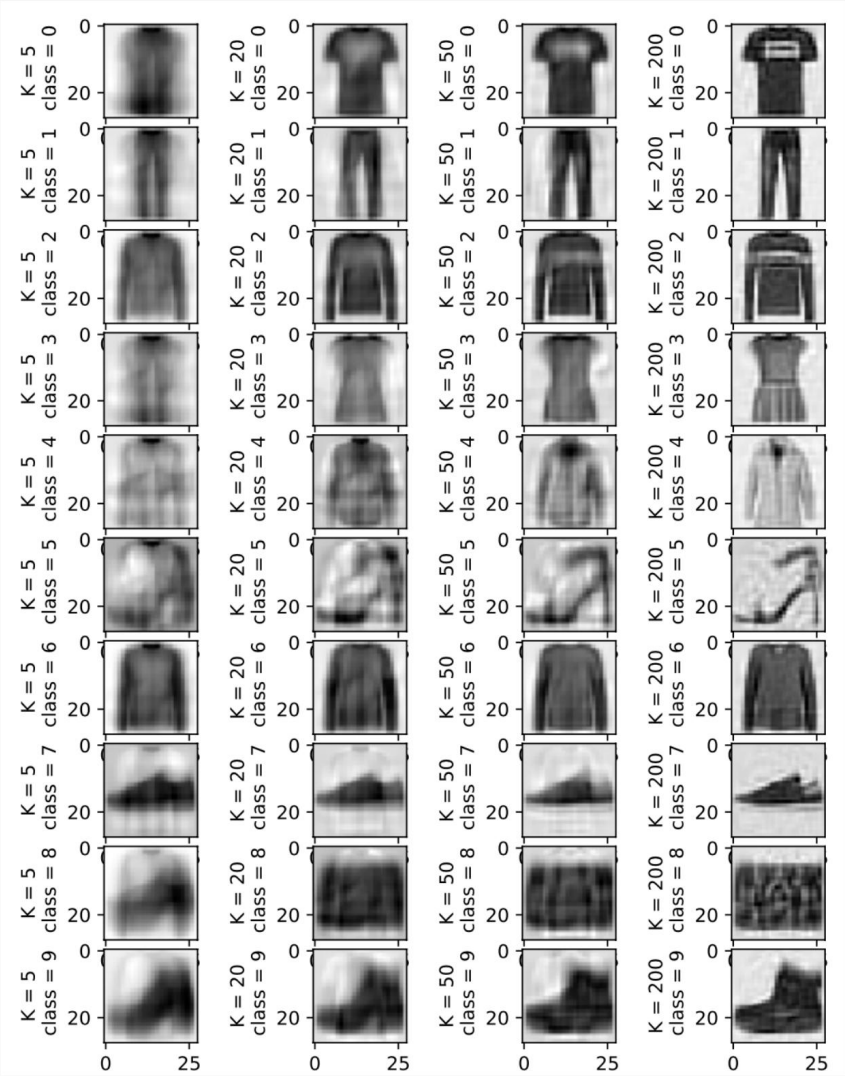
- 1 pts The answer is not presented in the specified table format

- 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 $K = 5,\ 20,\ 50,\ 200$.



When applying more number(K from 5 to 200) of principle components on dimensional reduction with PCA, reconstructed images in each class become more and more clear and we can see more and more details and the image looks like more and more similar to their original one. This is because each image is made by linear combination of principle components plus a mean vector from its class. The larger number of K we use, the larger variance will shown by the graph(bacome more like its original one). When using the first five principle components, the image is still blurry and only cover the shape of its mean vector of that class. The image lack their unique features in order to identifying itself. When using the first 200 principle components, the image is almost the same as its original picture and their unique characteristics appear clearly. Now we could identify the corresponding image in the original dataset while using the inverse_transform image.

1.7 4 / 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.

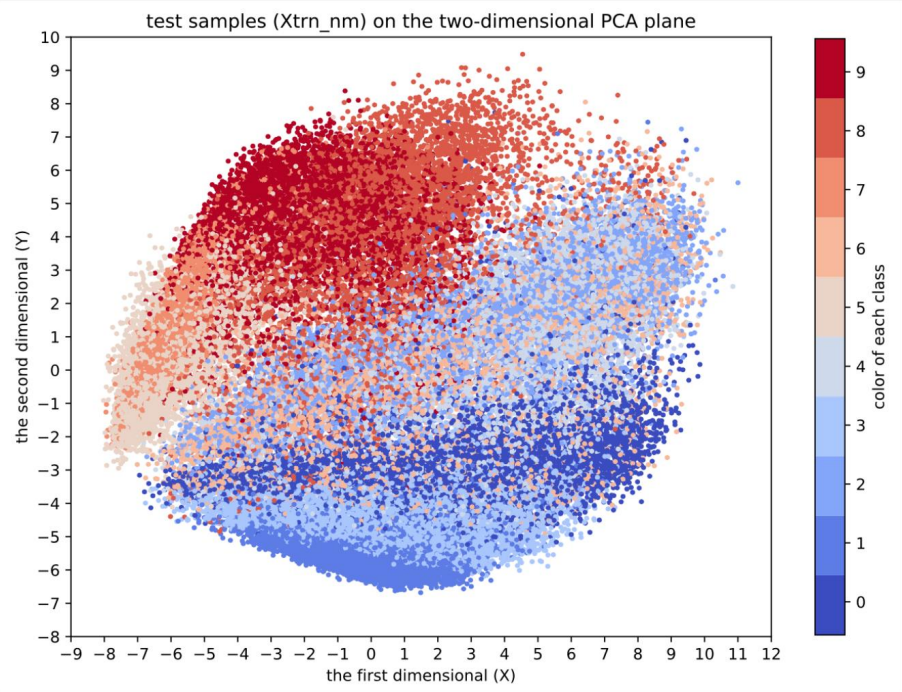
- 0.5 pts Some of your discussions do not make sense

- 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.



We can find that the same class of data are grouped together on the image, this means the general feature in samples of each class are similar to each other. We could find that the warm colors(class 5-9) and cool colors(class 0-4) are separated up and down and data of class 7,8,9 are located together, this is because they are all kinds of shoes. And data of class 0,2,4,6 are located in the middle of the graph as these classes are all kinds of clothes. Warm colors are likely to have positive values on the second dimension and cool colors are likely to have negative values on the second dimension. However, the original dimensions(number of features = 784) are too large compared with the image above which reduced to 2 dimensions by PCA and we will lost a lot of variance. That is why most of the points overlap and did not cluster perfectly on the image.

1.8 4 / 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.

Classification accuracy = 0.840(rounded to 3 decimal places)
(Original classification accuracy = 0.8401)
Confusion matrix:

```
[[819   3  15  50   7   4  89   1  12   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]]
```

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.

Classification accuracy = 0.846(rounded to 3 decimal places)
(Original classification accuracy = 0.8461)
Confusion matrix:

```
[[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]]
```

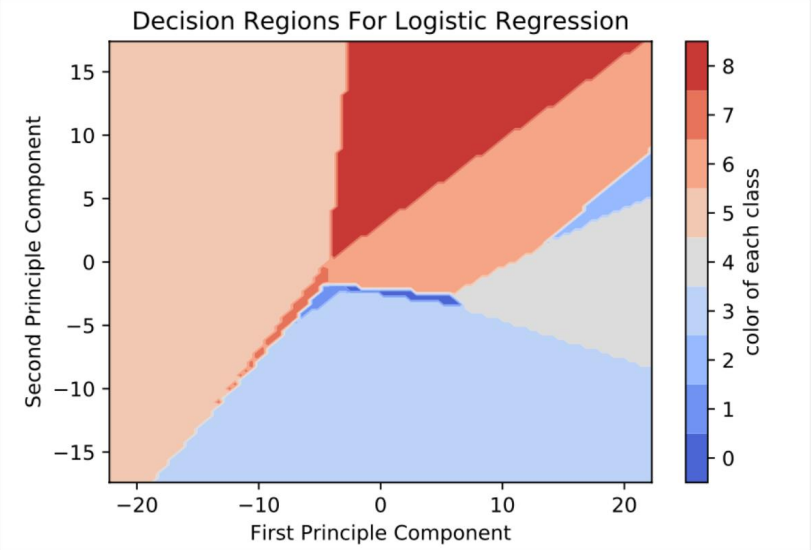
2.2 **2.5 / 3**

- **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.



Classes with cool colors are likely to have negative values on second principal components. As logistic regression classifier is based on linear classifier, the decision boundaries are all straight lines(linear). Also when applying PCA and reduced the original dimensions to 2 dimensions, data lost a lot of information. Inverse_transformation from 2D data to original dimensions will lost lots of variance and only first two components information remain, so it is hard for classifier identifying different classes. There are only appear 9 classes on the graph and there is no class 9 in this bounded area, because data in class 9 are not linearly separable on original dimensions. As the logistic regression classifier based on linear combination, we could not identify class 9 based on the information provided by the first two principle components as we lost information from original data.

2.3 6 / 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

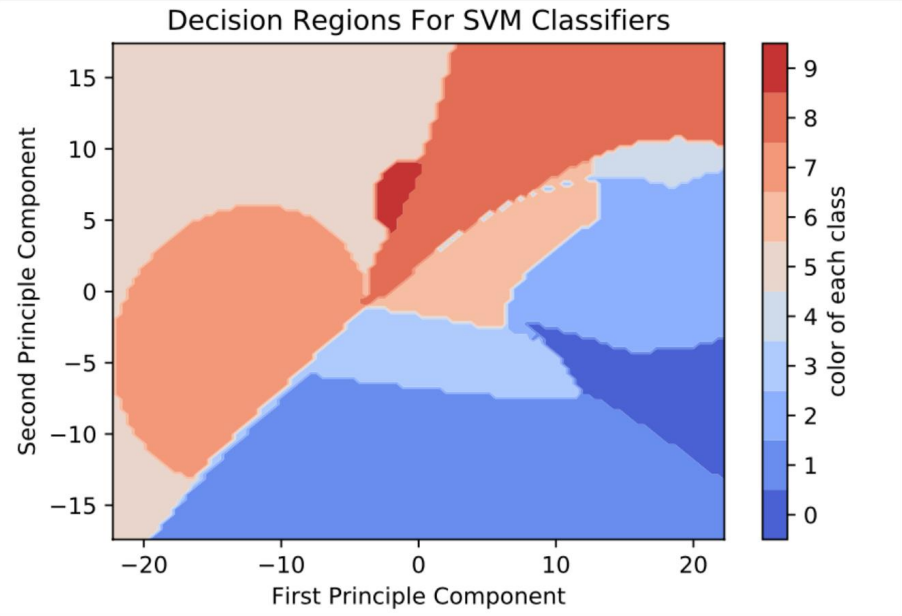
- 2 pts Your findings do not make sense

- 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 color pattern is similar to logistic regression. But the decision boundaries of SVM classifier are more smoother and not are straight lines(non-linear) compared with logistic regression. Class 9 appear on this graph but not apper in logistic regression. Because data in class 9 are not linear separable on original dimensions. SVM based on Gaussian kernel to find a hyperplane to map the low dimension to the higher dimension to make the non linearly separable data linearly separable. This will more likely to give us a better accuracy than logistic regression classifier in this task.

2.4 4 / 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.

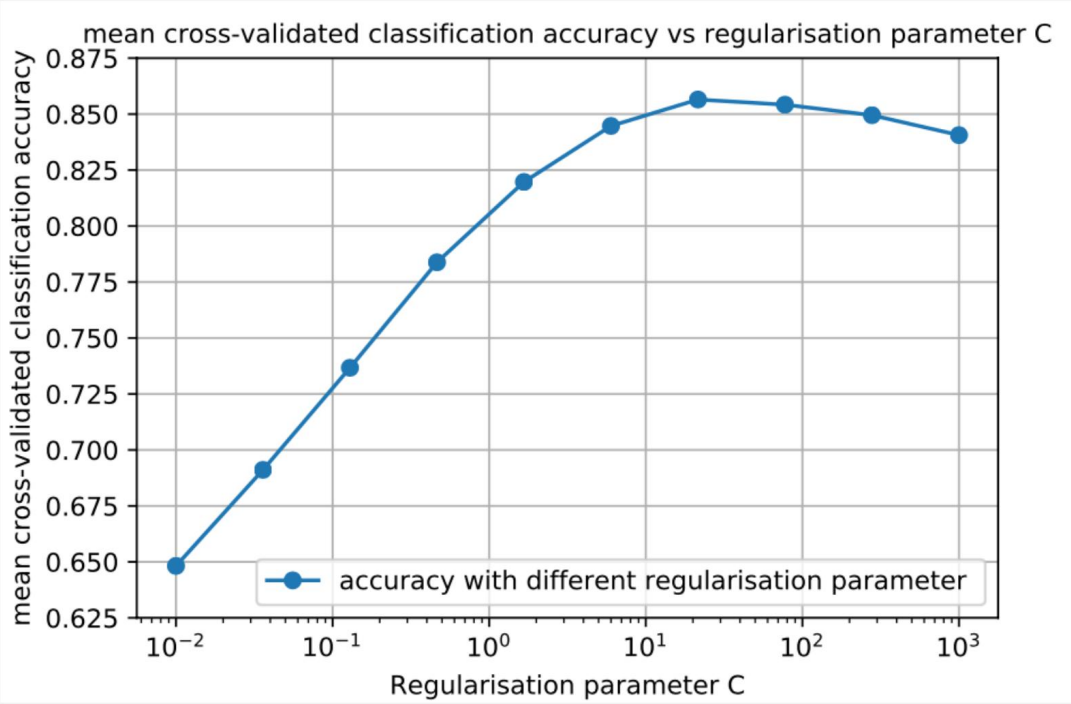
- 1 pts The plotting range is incorrect and details of decision regions are unclear

- 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`.



the highest obtained mean accuracy score and the value of C: 21.54(rounded to 2 decimal places)
Highest obtained mean accuracy score: 0.857(rounded to 3 decimal places)

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.

Training accuracy	Test accuracy
0.908	0.877

(rounded to 3 decimal places)

Original classification accuracy on the training set is 0.9084
Original classification accuracy on the test set is 0.8765

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.

Cluster centre	number of samples
class = 0	1018
class = 1	1125
class = 2	1191
class = 3	890
class = 4	1162
class = 5	1332
class = 6	839
class = 7	623
class = 8	1400
class = 9	838
class = 10	659
class = 11	1276
class = 12	121
class = 13	152
class = 14	950
class = 15	1971
class = 16	1251
class = 17	845
class = 18	896
class = 19	930
class = 20	1065
class = 21	1466

sum of squared distances of samples to their closest cluster centre = 38185.82 (rounded to 2 decimal places)

Question 3 continued on next page...

3.1 3 / 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

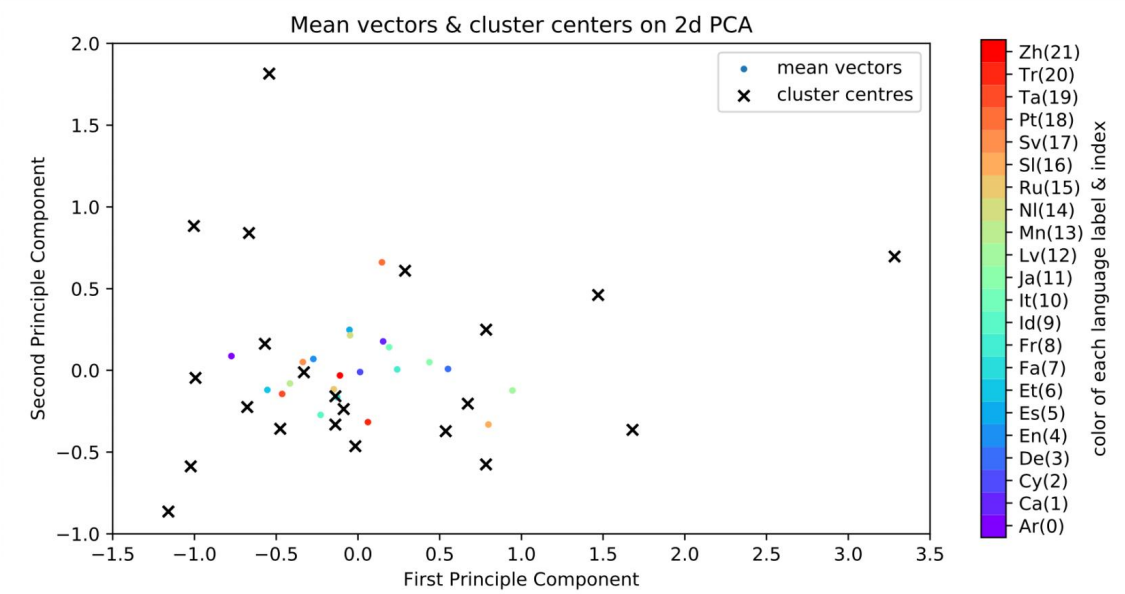
- 2 pts You failed to report the number of samples for each cluster

- 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.



Mean vectors and cluster centres are not similar. We can find that mean vectors of each class projected into 2D plane are gathered together and they are closed to each other, which means the general feature of each language are quite similar. Cluster centers are widely spread and some of them are far away from all mean vector points. Because Kmeans are sensitive to the initial cluster center, different selection methods will get different results. In addition, outliers or noisy data will have a greater impact on the mean, leading to a center shift, so we also need to detect abnormal points. Also, Kmeans is a unsupervised learning algorithm hence we could not find which cluster centre belongs to which language class centre point.

Question 3 continued on next page...

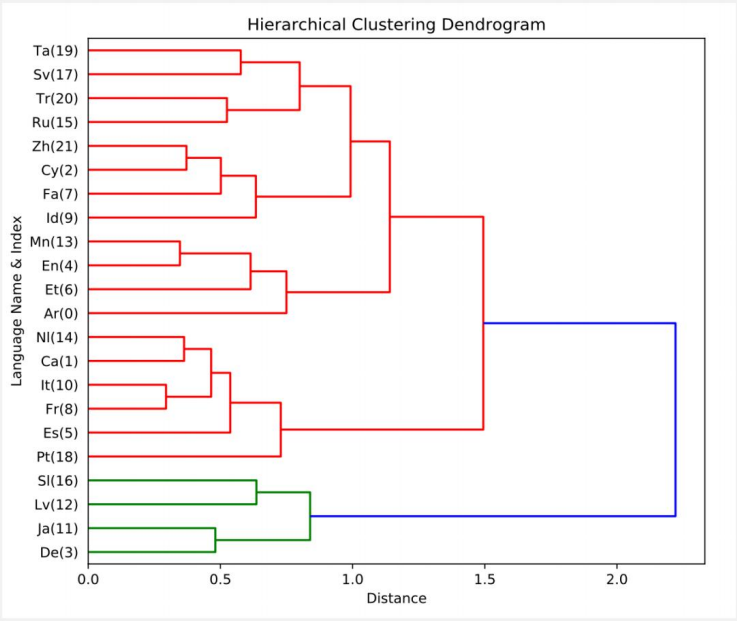
3.2 **2.5 / 3**

- **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.



Ward's minimum variance criterion minimizes the total within-cluster variance. To implement this method, at each step find the pair of clusters that leads to minimum increase in total within-cluster variance after merging. This increase is a weighted squared distance between cluster centers. Hence Ward could group the languages in the same geographical region into the same cluster such as European language cluster from Dutch to Portuguese. And we could find that Sl, Lv, Ja, De are clustered in one group(green) and other languages are clustered in the other group(red), because the two groups(green & red) have a large gap in distance thresholds. The distance threshold of Ward is not very big(around 2.0 to 2.5).

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

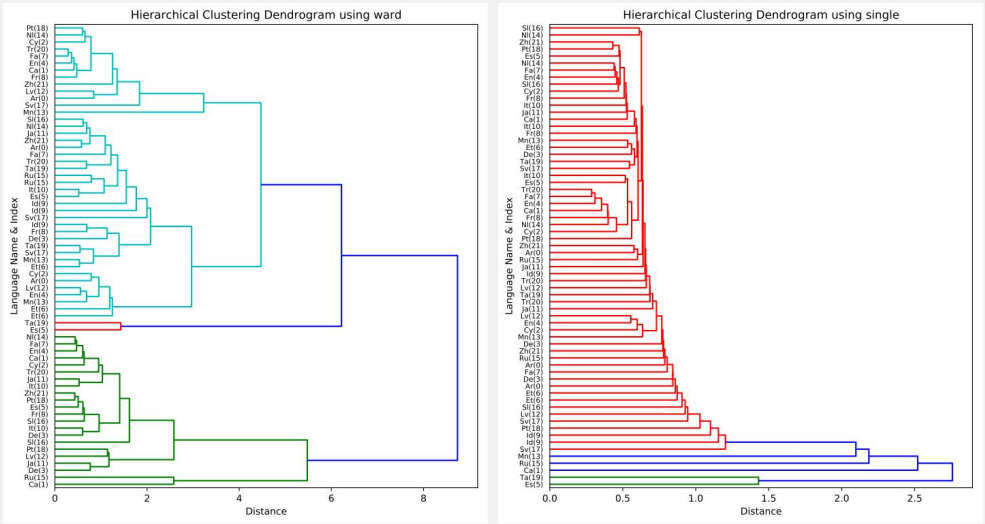
- 0.5 pts Your findings are not based on the result obtained

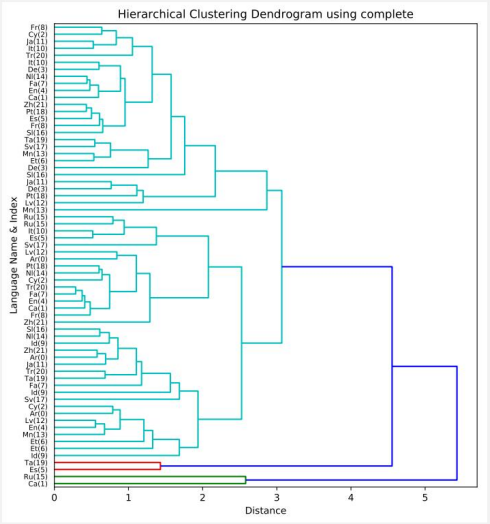
- 0.5 pts You failed to mention how your findings relate to the result in Q3.2

- 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.





As Single linkage using distance between closest elements in clusters and Complete linkage using distance between farthest elements in clusters and Wards linkage consider joining clusters. We could find that when we use different linkage methods, we will get different language clusters. Ward and Complete has a quite similar cluster pattern because they both will not act well on global data. For single linkage, although it is not robust to noisy data, it can be calculated very efficiently, so it is very useful for providing hierarchical clustering of larger data sets and also can be well represented on non-globular data. We could also find Es(Spanish) and Ta(Tamil) are always cluster together as one group as Es and Ta are two outliers. The threshold distances are different, single method is the smallest one, complete method is the middle one and ward method has the greatest threshold distance.

Question 3 continued on next page...

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

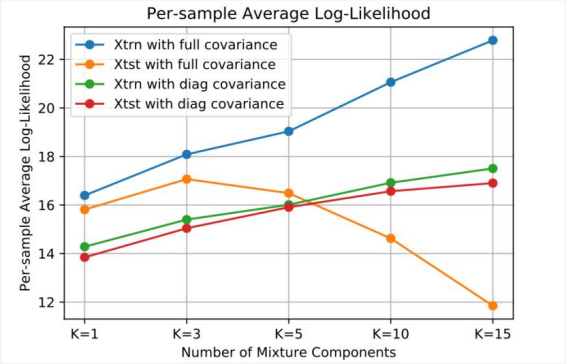
Discussions

- **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.,



	K=1	K=1	K=3	K=3	K=5	K=5	K=10	K=10	K=15	K=15
	full	diag	full	diag	full	diag	full	diag	full	diag
Xtrn	16.39	14.28	18.09	15.40	19.04	16.01	21.06	16.92	22.79	17.50
Xtst	15.81	13.84	17.07	15.04	16.49	15.91	14.62	16.57	11.85	16.90

When GMM with covariance matrix type which is full, the GMM fit training data well and the likelihood increases when K increases. But with increasing of K, the likelihood of Xtst decreases after K=3. The reason is that it added more components and this will more likely to let model overfitting. When the covariance matrix type is diagonal, the model of training data and testing data are similar, with the increasing of mixture components. This is because diagonal covariance matrix lose the correlation between different dimensions of and make different dimensions of data independent to each other, hence the following results are less affected by increasing in components.

3.5 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