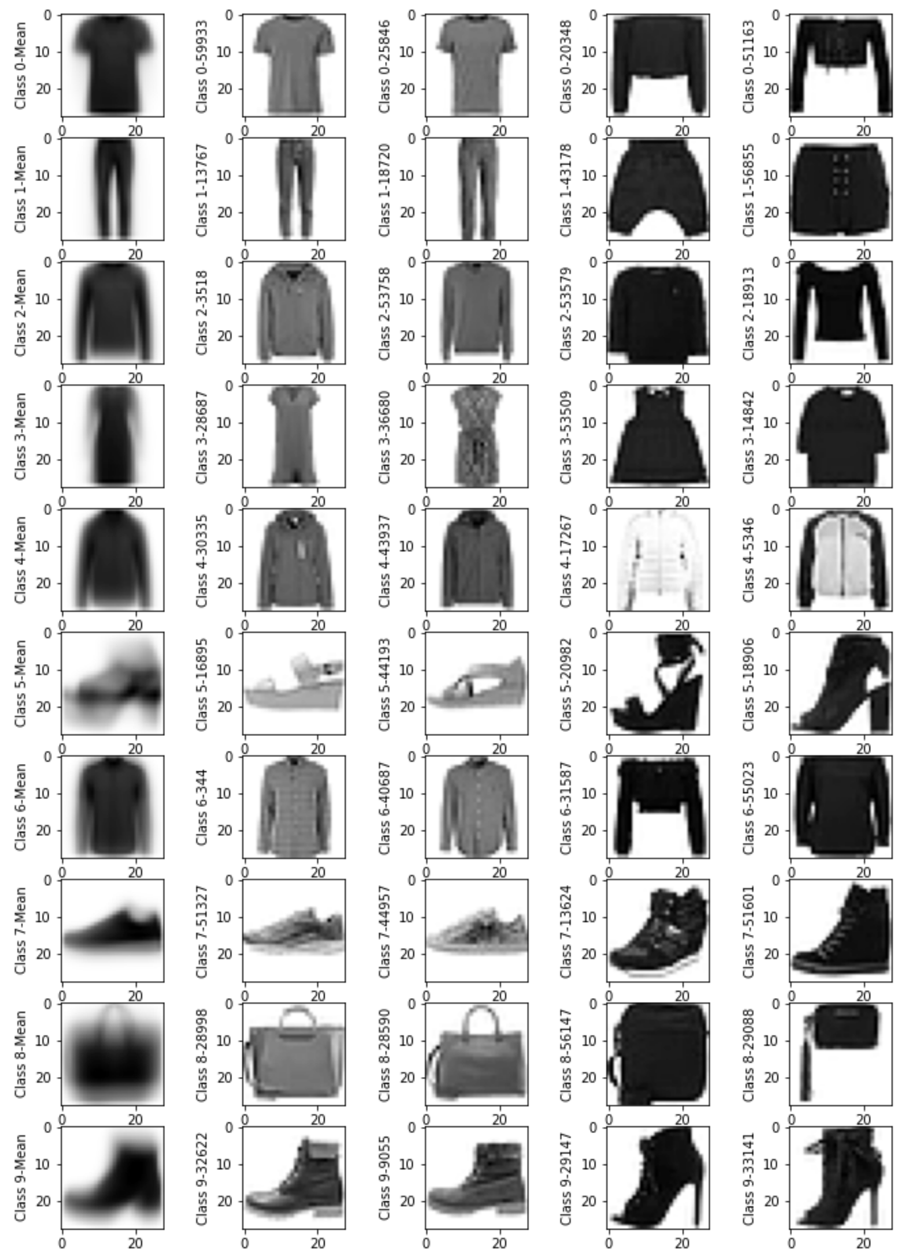
Question 1 : (30 total points) Image data analysis with PCA

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 values of the first four elements of the first training sample are $[-3.14 * 10^{-6}, -2.27 * 10^{-5}, -1.18 * 10^{-4}, -4.07 * 10^{-4}]$.
The values of the first four elements of the last training sample are also $[-3.14 * 10^{-6}, -2.27 * 10^{-5}, -1.18 * 10^{-4}, -4.07 * 10^{-4}]$.

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.

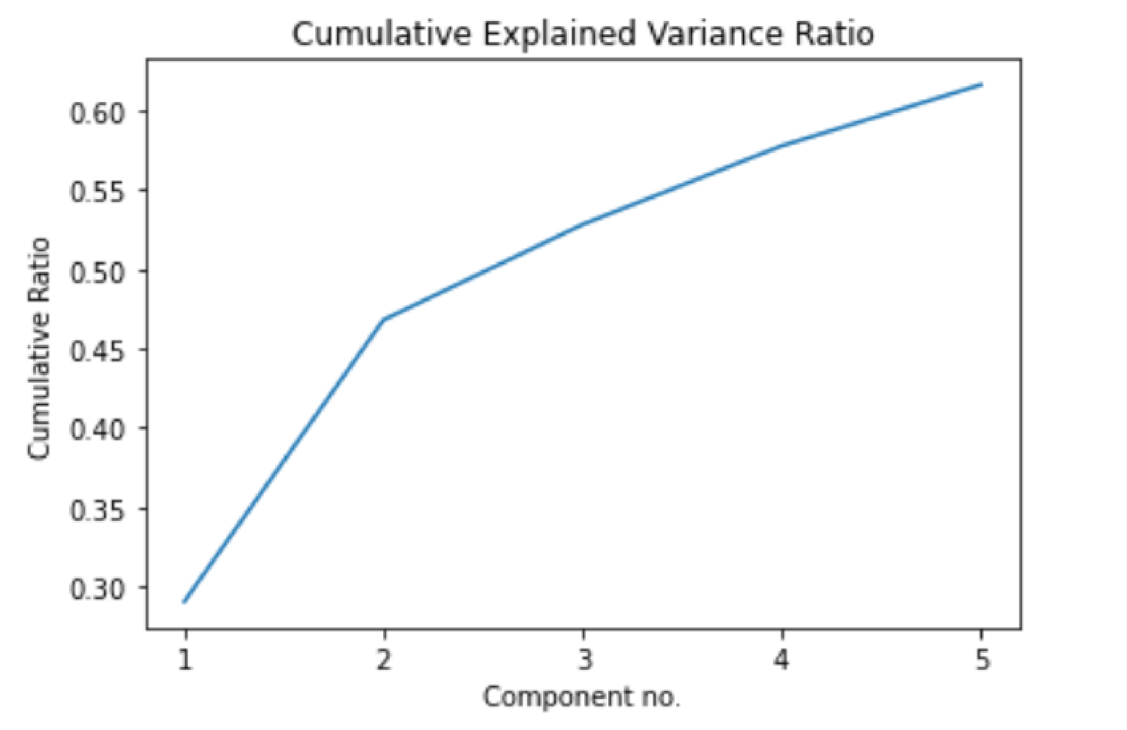


The images of the two closest samples to mean are very similar to the image of the mean vector, while the images of the two furtherest samples to mean are very different from the image of the mean vector.

1.3 (3 points) Apply Principal Component Analysis (PCA) to the data of `Xtrn_nm` using sklearn.decomposition.PCA, and find the cumulative explained variance.

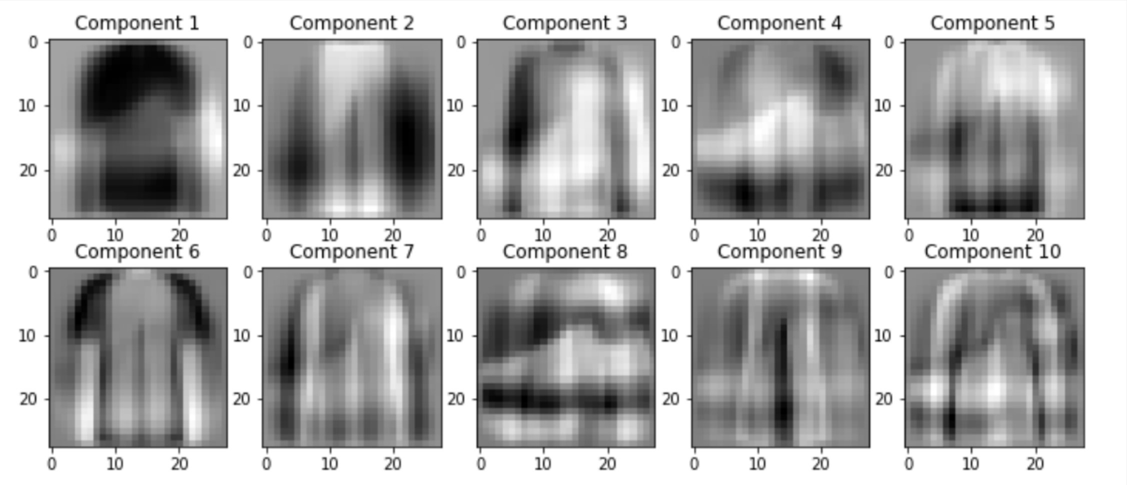
Variance	Cumulative Variance
19.81	19.81
12.11	31.92
4.11	36.03
3.38	39.41
2.62	42.03

1.4 (3 points) Plot a graph of the cumulative explained variance ratio. Discuss the result briefly.



The first two components contribute to 46.8% ratio, which is almost half of the total ratio. The contribution of the third, fourth and the fifth components gradually becomes insignificant. The first five principal components contribute to 61.6% ratio.

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.



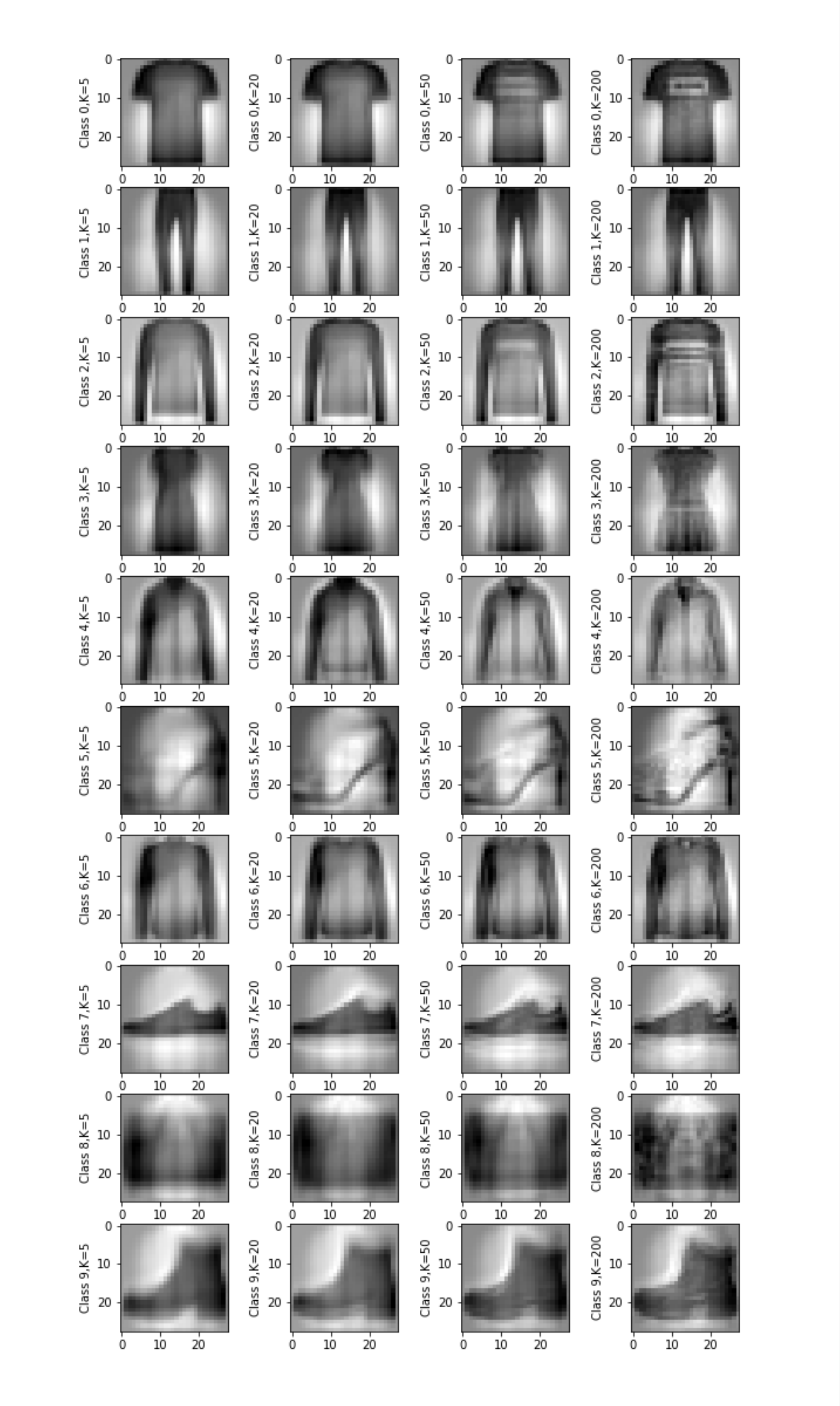
The images get more and more ambiguious because the ratio of variance explained gets smaller from the first component to the tenth component, yet they still show the overall pattern of each class.

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 RMSE for $K = 5, 20, 50, 200$ are given by

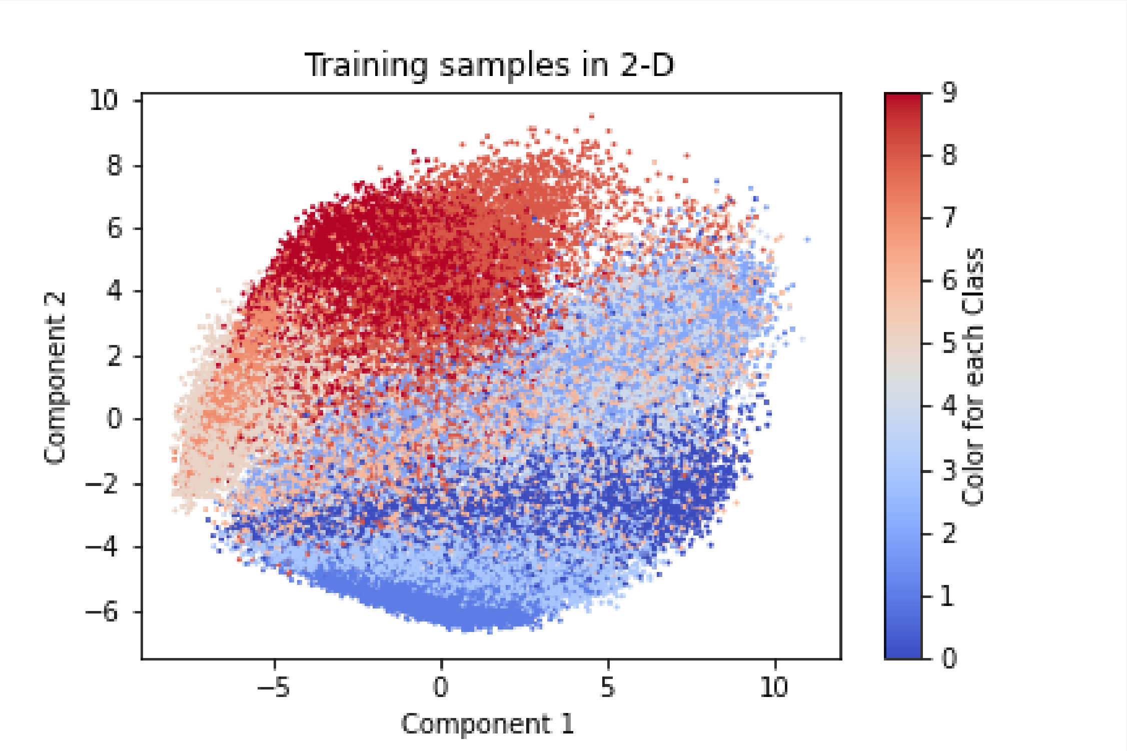
Class	K=5	K=20	K=50	K=200
0	0.219	0.202	0.183	0.169
1	0.226	0.215	0.213	0.215
2	0.202	0.200	0.189	0.170
3	0.193	0.180	0.180	0.172
4	0.189	0.191	0.180	0.174
5	0.225	0.219	0.222	0.218
6	0.139	0.120	0.116	0.106
7	0.165	0.198	0.201	0.216
8	0.199	0.198	0.193	0.179
9	0.247	0.222	0.219	0.217

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 images get clearer and more detailed as the number of components K gets larger, because more information of the original data is captured.

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.



Class 0,2,4,6 are relatively close to each other, as they all belong to the clothes' category; Class 5,7,9 are relatively close to each other, as they all belong to the shoes' category; Class 3 lies between Class 0,2,4,6 and Class 1, as Class 1 belongs to the pants' category and Class 3 belongs to the dress‘s category, which includes features of both clothes and pants.

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 accuracy is 84.01%. The confusion matrix is

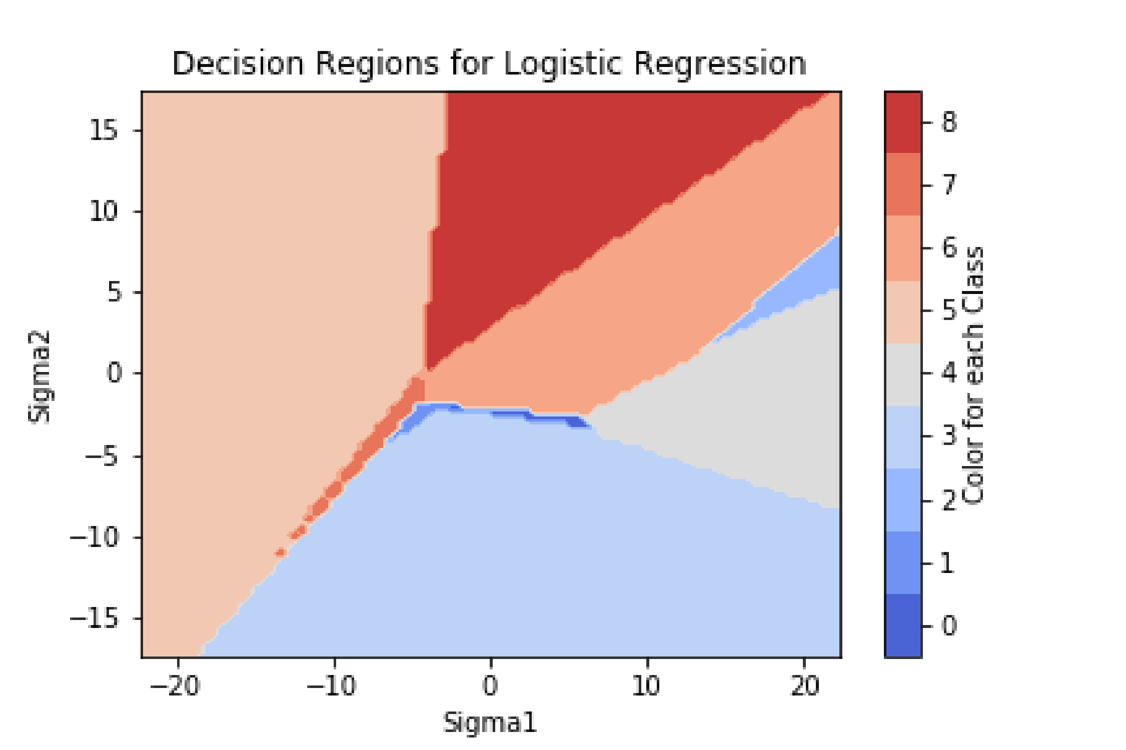
$$\begin{pmatrix}
819 & 3 & 15 & 50 & 7 & 4 & 89 & 1 & 12 & 0 \\
5 & 953 & 4 & 27 & 5 & 0 & 3 & 1 & 2 & 0 \\
27 & 4 & 731 & 11 & 133 & 0 & 82 & 2 & 9 & 1 \\
31 & 15 & 14 & 866 & 33 & 0 & 37 & 0 & 4 & 0 \\
0 & 3 & 115 & 38 & 760 & 2 & 72 & 0 & 10 & 0 \\
2 & 0 & 0 & 1 & 0 & 911 & 0 & 56 & 10 & 20 \\
147 & 3 & 128 & 46 & 108 & 0 & 539 & 0 & 28 & 1 \\
0 & 0 & 0 & 0 & 0 & 32 & 0 & 936 & 1 & 31 \\
7 & 1 & 6 & 1 & 11 & 3 & 7 & 15 & 945 & 0 \\
0 & 0 & 0 & 1 & 0 & 15 & 1 & 42 & 0 & 941
\end{pmatrix}$$

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 accuracy is 84.61%. The confusion matrix is given by

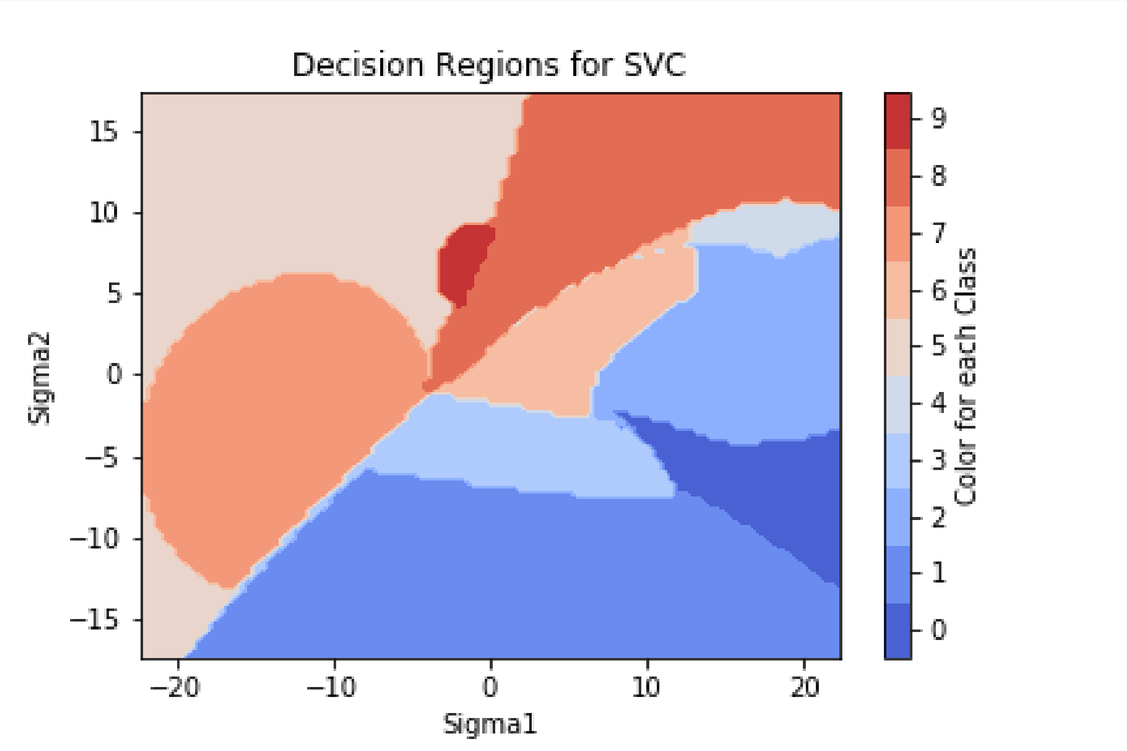
$$\begin{pmatrix}
845 & 2 & 8 & 51 & 4 & 4 & 72 & 0 & 14 & 0 \\
4 & 951 & 7 & 31 & 5 & 0 & 1 & 0 & 1 & 0 \\
15 & 2 & 748 & 11 & 137 & 0 & 79 & 0 & 8 & 0 \\
32 & 6 & 12 & 881 & 26 & 0 & 40 & 0 & 3 & 0 \\
1 & 0 & 98 & 36 & 775 & 0 & 86 & 0 & 4 & 0 \\
0 & 0 & 0 & 1 & 0 & 914 & 0 & 57 & 2 & 26 \\
185 & 1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

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



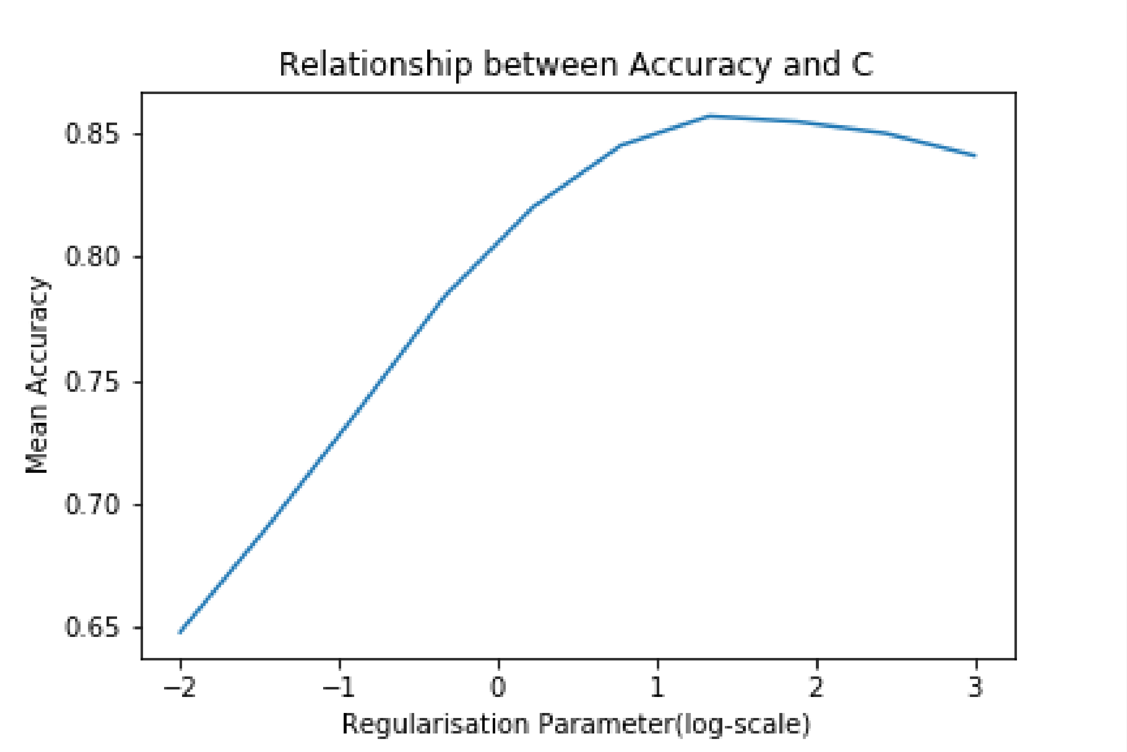
The points are classified into only 9 classes and the decision regions for Class 0 and Class 1 are very insignificant.

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 points are classified into 10 classes, which is more than that in 2.3. The size and shape of the decision regions for each class are also different from those in 2.3. The decision boundaries are more convoluted than those in 2.3. Overall SVC does a better job than Linear Regression in classification.

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 accuracy obtained is 0.857 and the corresponding value of C is 1.33.

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.

The accuracy for the training set is 0.908. The accuracy for the test set is 0.877.

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

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

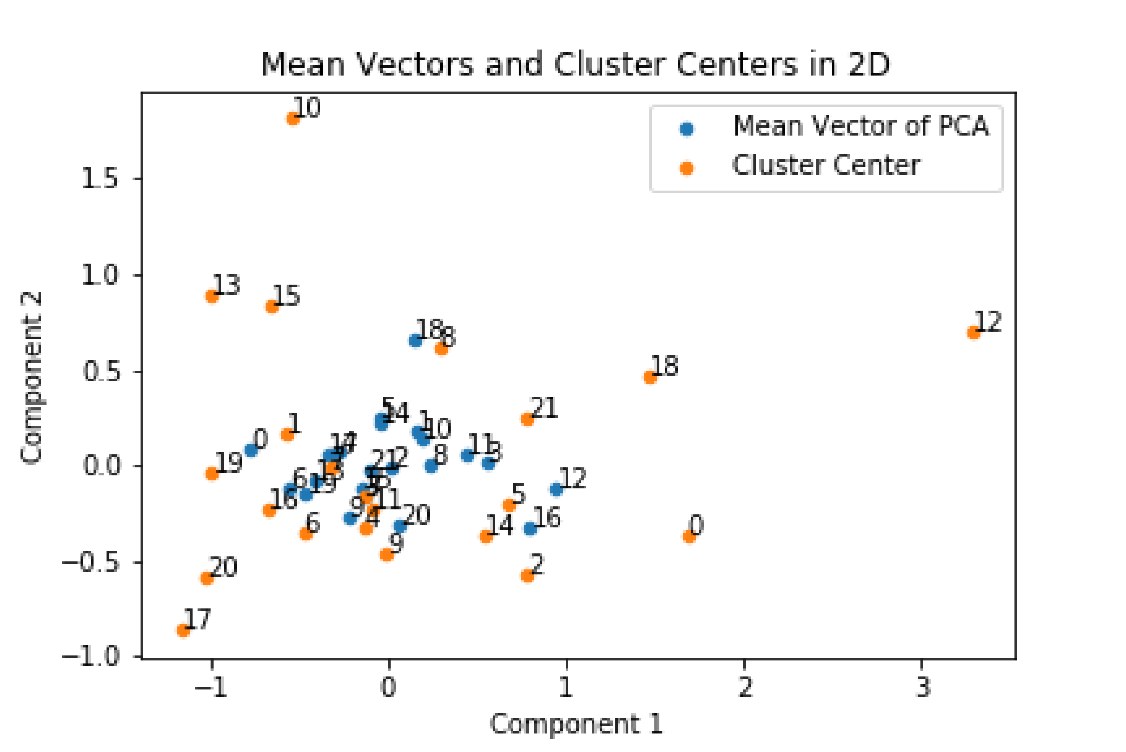
3.1 (3 points) Apply k-means clustering on `Xtrn` for $k = 22$, where we use sklearn.cluster.KMeans with the parameters `n_clusters=22` and `random_state=1`. Report the sum of squared distances of samples to their closest cluster centre, and the number of samples for each cluster.

The sum of squared distances of samples to their closest cluster center is 38186. The number of samples for each cluster is given by

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

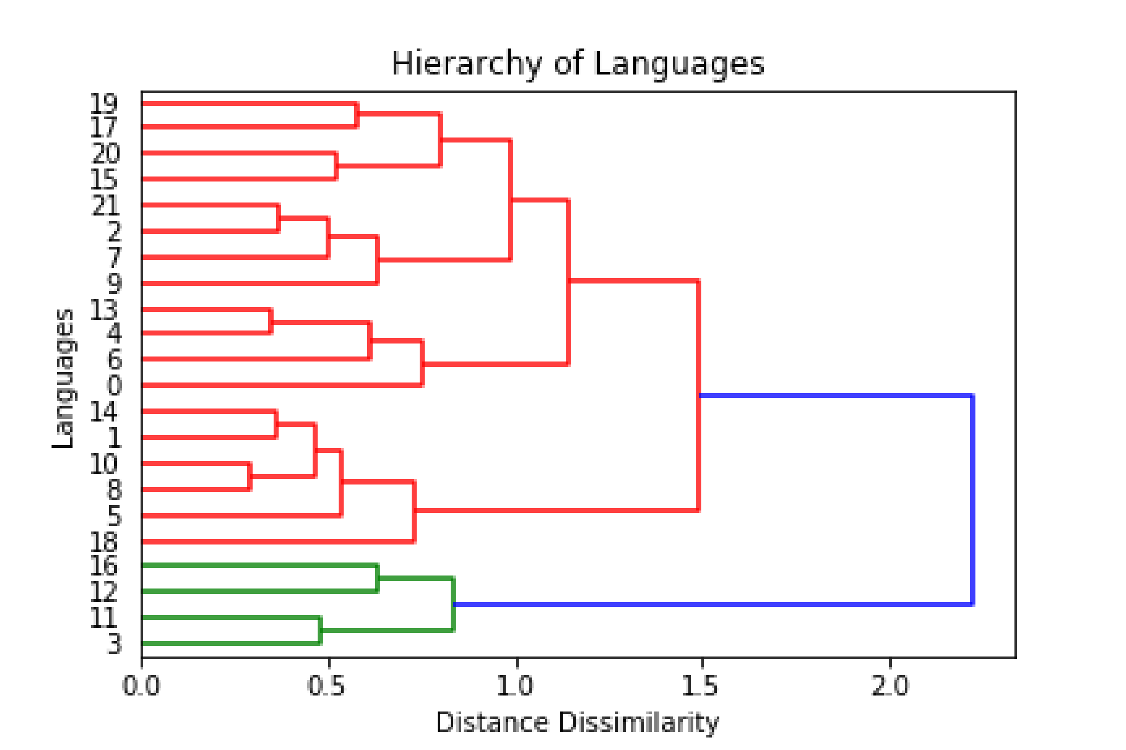
Question 3 continued on next page. . .

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.



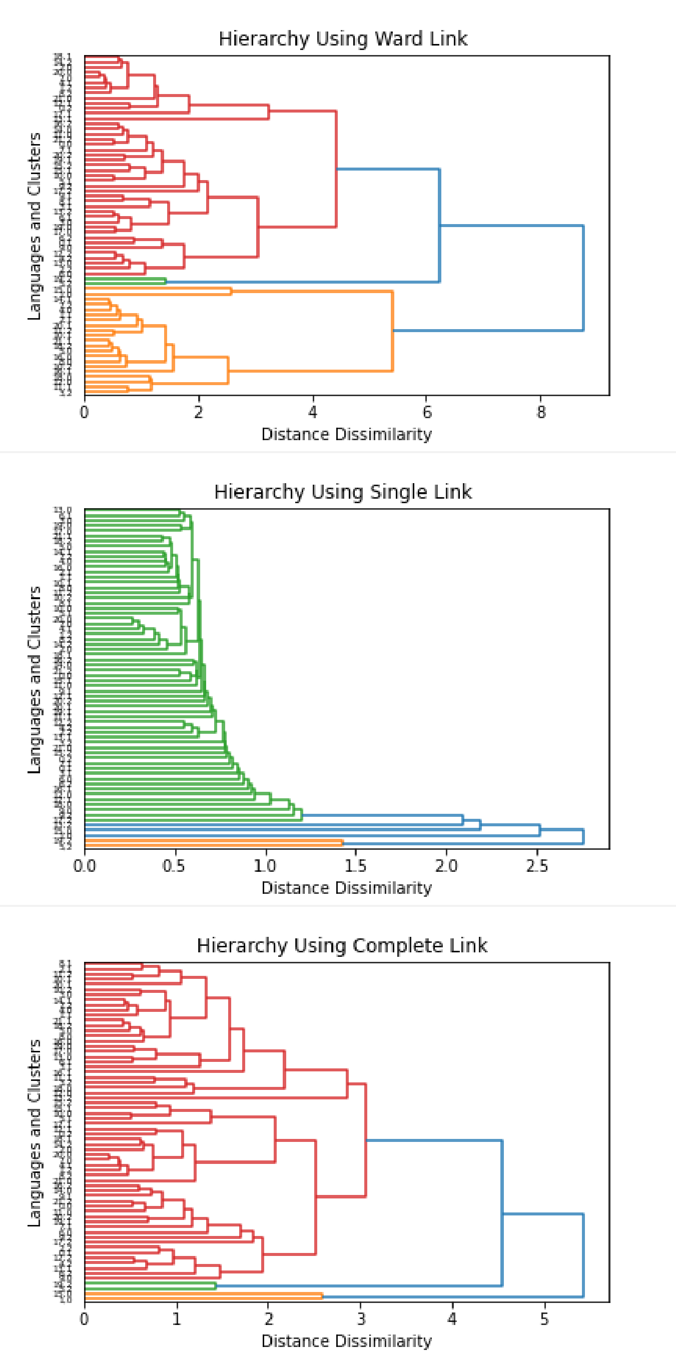
The mean vectors are more concentrated while the cluster centers are sparser. This is because some samples with the same language do not necessarily belong to the same cluster. Samples of different languages may belong to the same cluster if they are relatively close to each other.

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.



Language 3,11,12,16 belong to a cluster while the other 18 languages belong to another cluster.

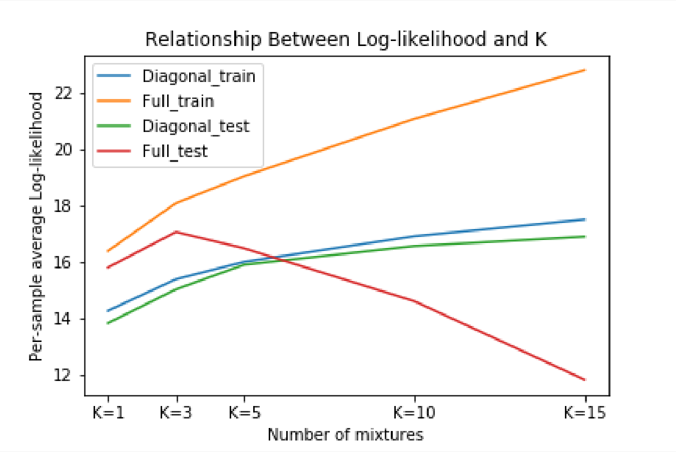
3.4 (5 points) We here extend the hierarchical clustering done in Question 3.3 by using multiple samples from each language.



19-2 and 5-2 would end up in the same small cluster when all three methods are used. When 'ward' is used, there will be another two large clusters whereas when the other two methods are used, there will be a very small cluster besides the one mentioned earlier,and a very large cluster. Samples of the same language may end up in completely different clusters.

Question 3 continued on next page. . .

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 table below shows the per-sample average log-likelihood for each case.

——	K=1	K=3	K=5	K=10	K=15
Full,Train	16.4	18.1	19	21	22.8
Full,Test	15.8	17.1	16.5	14.6	11.8
Diagonal,Train	14.3	15.4	16	16.9	17.5
Diagonal,Test	13.8	15	15.9	16.6	16.9

When the full covariance matrix is used on test data, the log-likelihood would first increase when K increases from 1 to 3, then decrease as K increases. For the other three cases, the log-likelihood would increase as K increases.