Fall 2016
CSCE 666 Pattern Analysis
Homework #3

Dennis Rodrigo da Cunha Silva

In recognition of the Texas A&M University policies of academic integrity, I certify that I have neither given nor received dishonest aid in this homework assignment.

Name: Signature:

Problem 1

The average face computed from the set of images provided can be seen in Figure 1.1. As it can be seen, some data details are still preserved in the average image, such as glasses contours and a predominately male face (this is expected since most faculties are men). This average face is used in the snapshot principal component analysis (Snapshot PCA) method. The first six eigenfaces found can be seen in Figure 1.2.

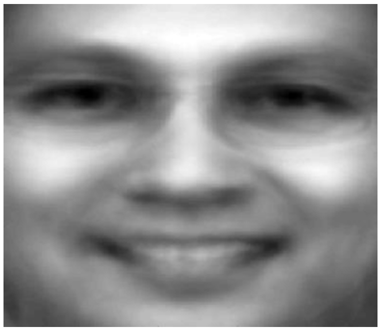

Figure 1.1: Average face.

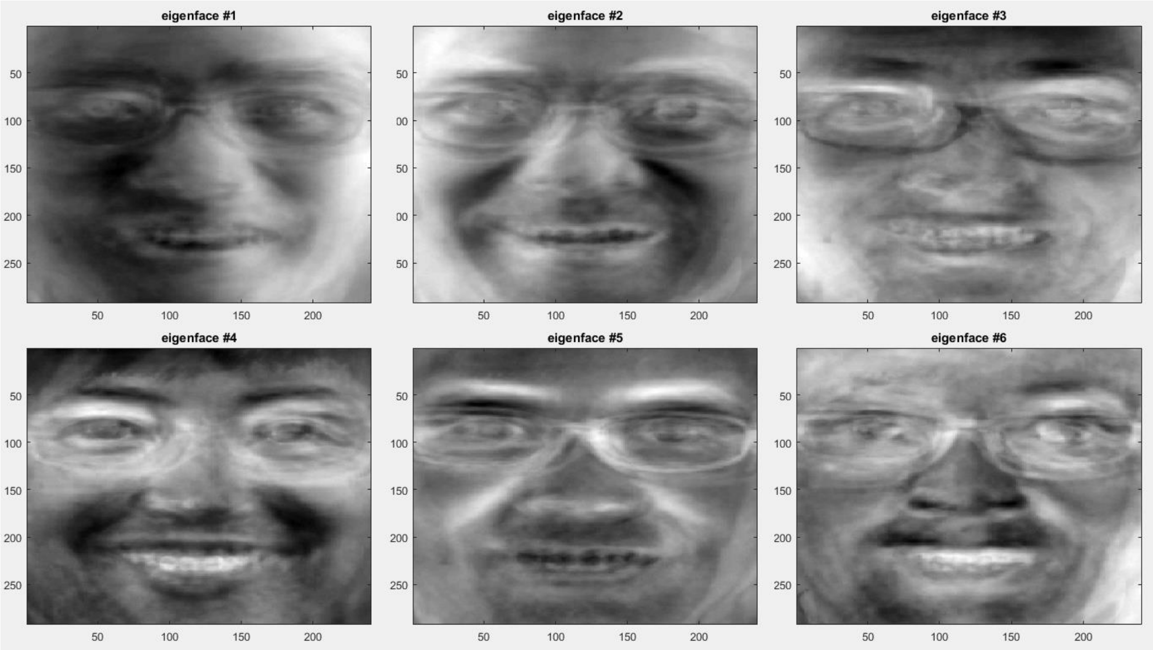

Figure 1.2: First six eigenfaces found.

The eigenfaces shown in Figure 1.2 can be interpreted as base faces for the corresponding dataset. One can note that details such as glasses contour, lighting, smile, laugh lines, etc are preserved on those eigenfaces. Eigenface #1 and #2, for example, capture lighting in different directions: eigenface #1 has a shadow on the left hand side and is brighter on the right hand side, and eigenface #2 has the opposite characteristics. Also, the eigenfaces shown in, its majority, resemble male faces. This is also expected, since the majority of faculties (33/42) are male. Eigenfaces #1 and #4 also shows more noticeable laugh lines. Eigenface #6 captures a darker region around the mouth that looks like goatee – the faculties with more noticeable facial hair have goatee rather than full beard or mustache only, such as Dr. Ioerger, Dr. Gooche and Dr. Walker, as can be seen in Figure 1.3.

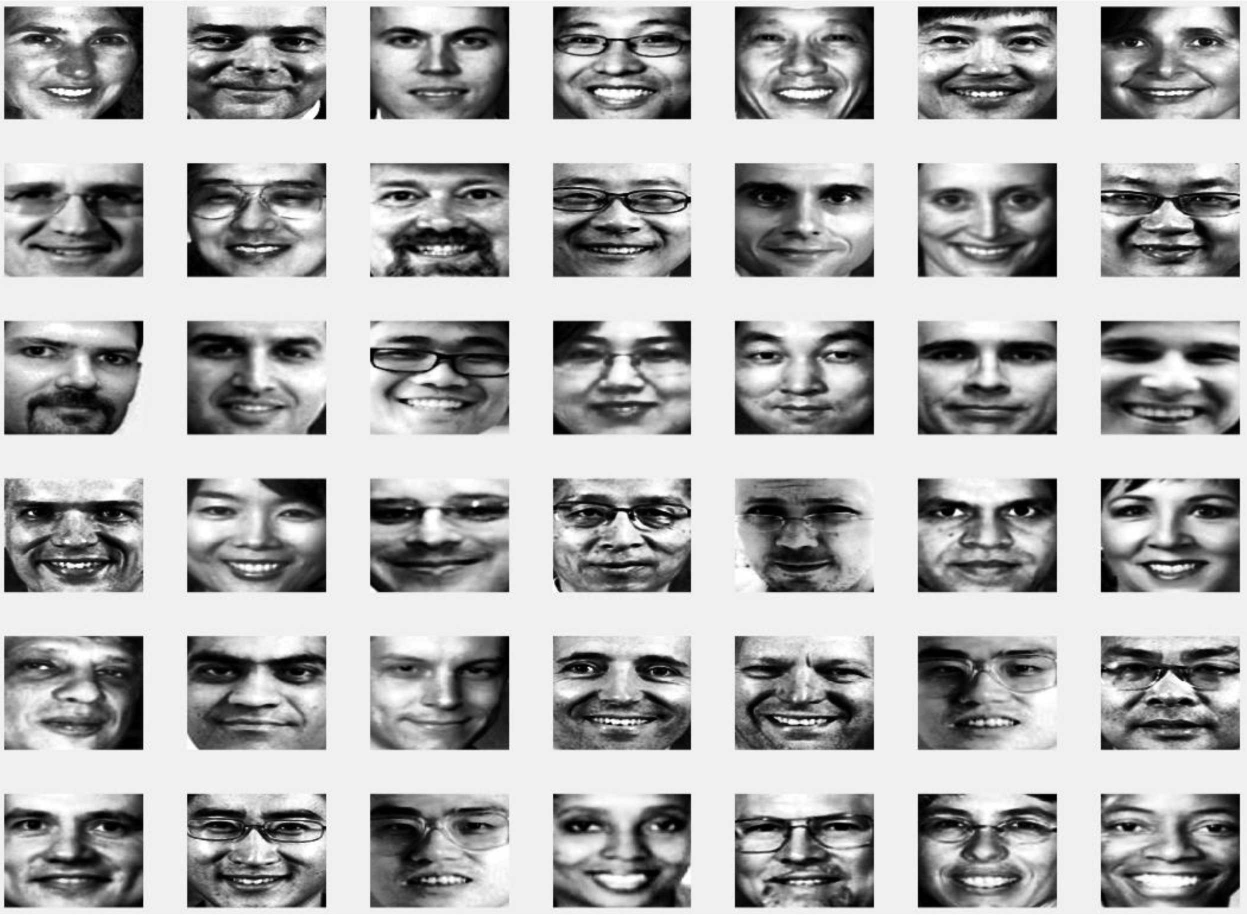

Figure 1.3: Faculty images in the data provided.

The 2D scatter plots for the first six principal components are shown in Figure 1.4. Most of the scatter plots shown in Figure 1.4 do not seem to carry any structural information. However, the scatter plot for principal components 1 and 2 has interesting results. As it can be seen in Figure 1.5, I manually divided the data points into to two clusters. The corresponding images divided into these two clusters can be seen in Figure 1.6.

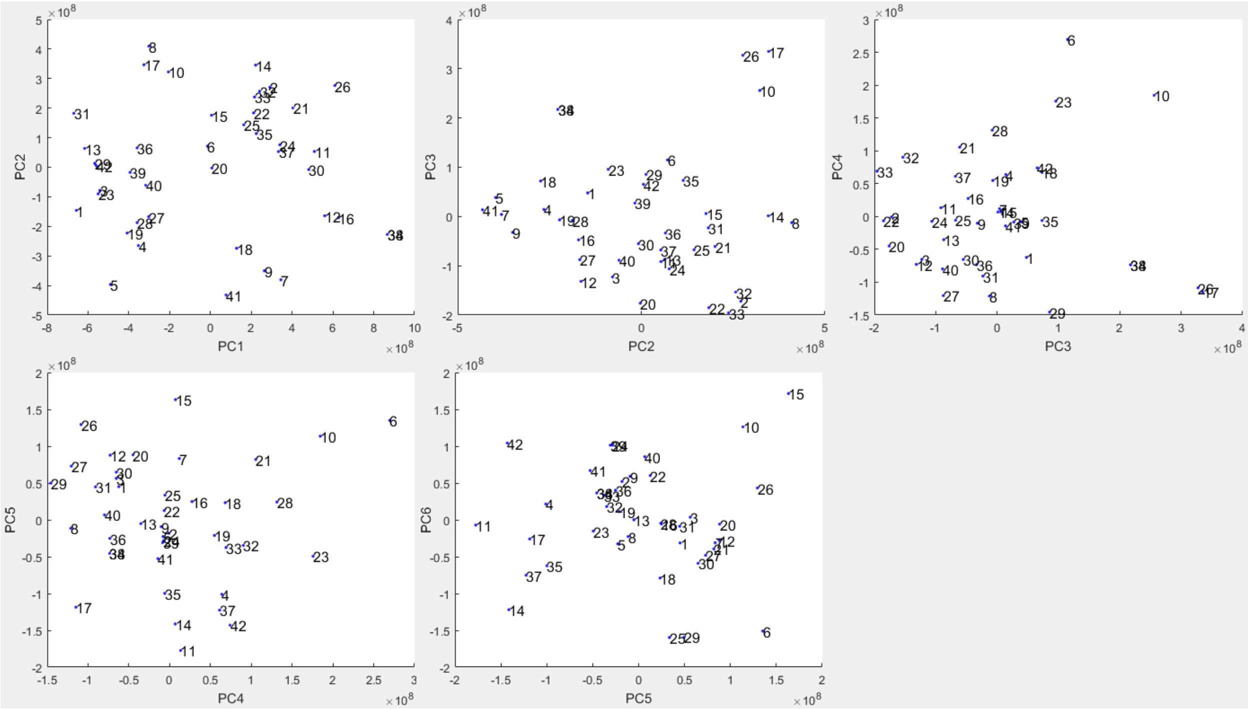

Figure 1.4: 2D scatter plots for the first six principal components.

The most interesting fact about the two clusters divided is that all female faculties were assigned to cluster 1, as it can be seen in Figure 1.6a. This could be used in face recognition applications. Also, all faculties with their hair partially shown in the images are also assigned to cluster 1 (except for Dr. Choe). Most faculties with darker regions around the eyebrows were assigned to cluster 2. Most faculties with the darker (more noticeable) frame glasses were also assigned to cluster 2 (except for Dr. Chai).

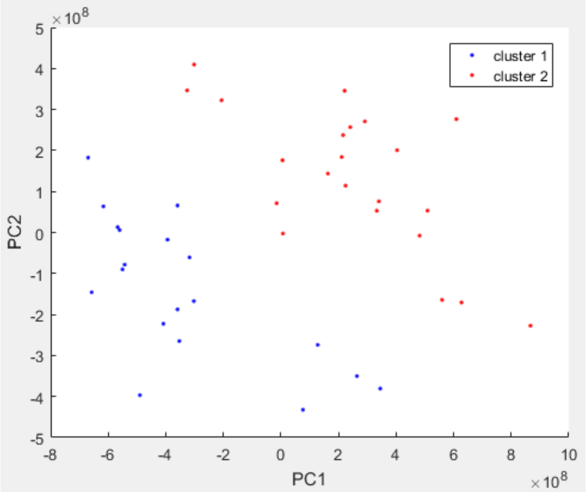

Figure 1.5: Clusters manually divided for principal components 1 and 2.

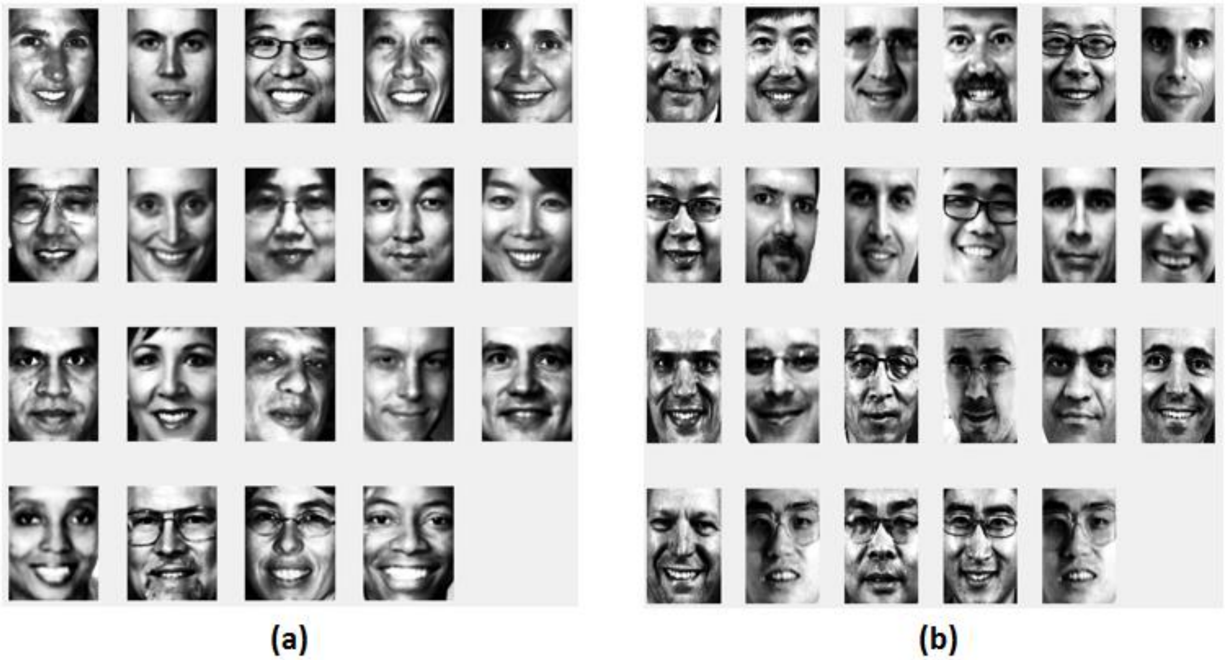

Figure 1.6: Images corresponding to clusters (a) 1 and (b) 2.

Problem 2

Figure 2.1 shows the eigenvalues found by using the ISOMAP technique applied to the provided United States (US) data. As it can be seen, the eigenvalues 1 and 2 carry most of the variance in the dataset and by plotting the 2D and 3D maps that could be confirmed.

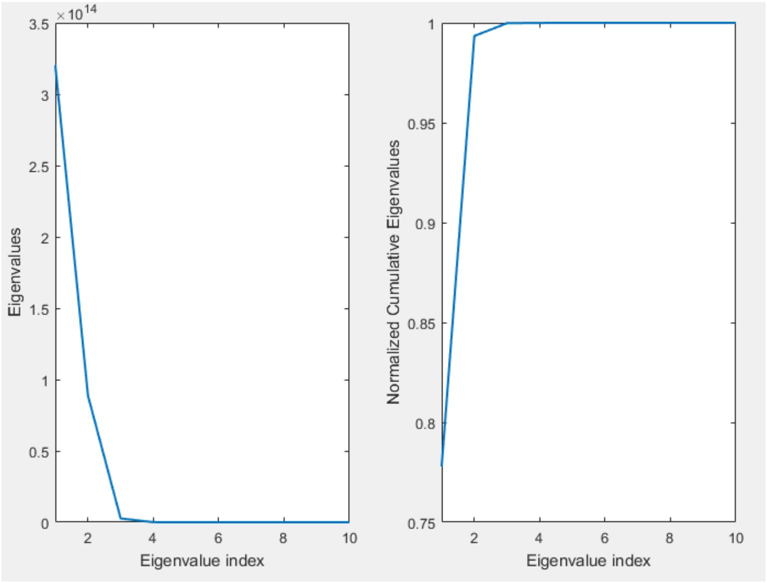

Figure 2.1: Eigenvalues found by using ISOMAP into the US map data.

The 2D manifold found for the US data provided is shown on Figure 2.2. As one can notice, the manifold accurately resembles the shape of the US. Figures 2.3a and 2.3b shows zoomed in regions in the manifold shown in Figure 2.2. As it can be seen in Figure 2.3a, cities such as, Austin, El Paso and New Orleans are relatively close, as well as Memphis, Birmingham and Atlanta. As for Figure 2.3b, St. John is close to Eastport, and Toronto is close to Buffalo.

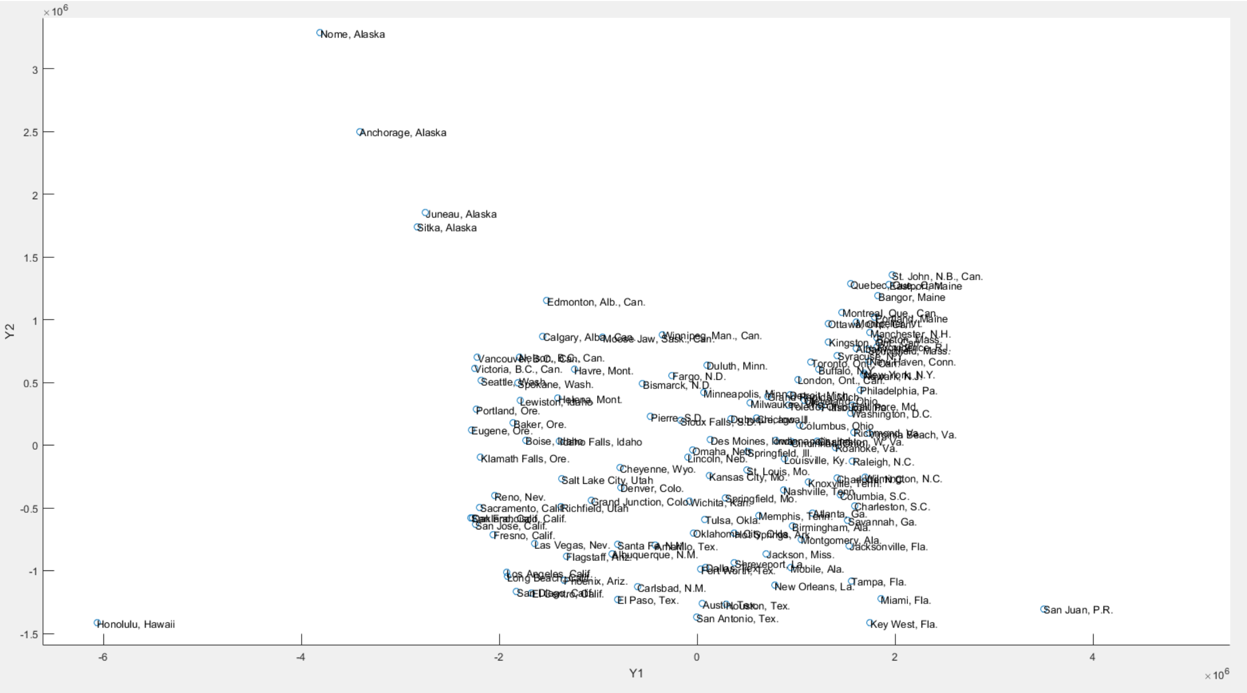

Figure 2.2: 2D US manifold found with ISOMAP.

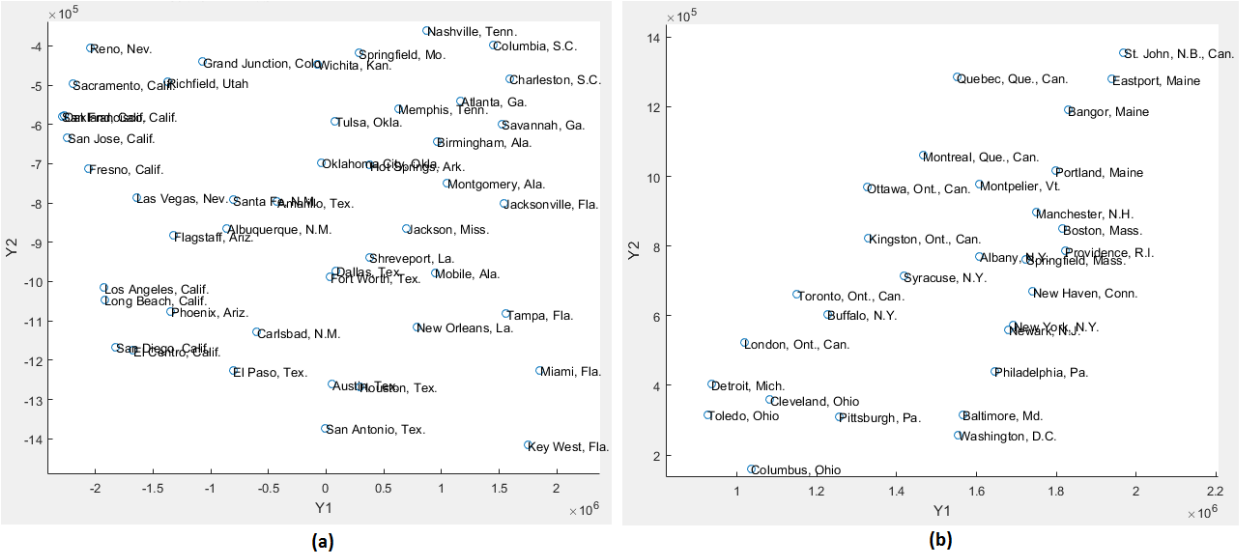

Figure 2.3: Portions of the manifold shown in Figure 2.2. (a) shows southern US and (b) shows northeastern US and part of Canada.

Figure 2.4 shows the 3D manifold found for the US data. Although some similarities are maintained, using the 3D manifold for visualization might not help as much as the 2D one.

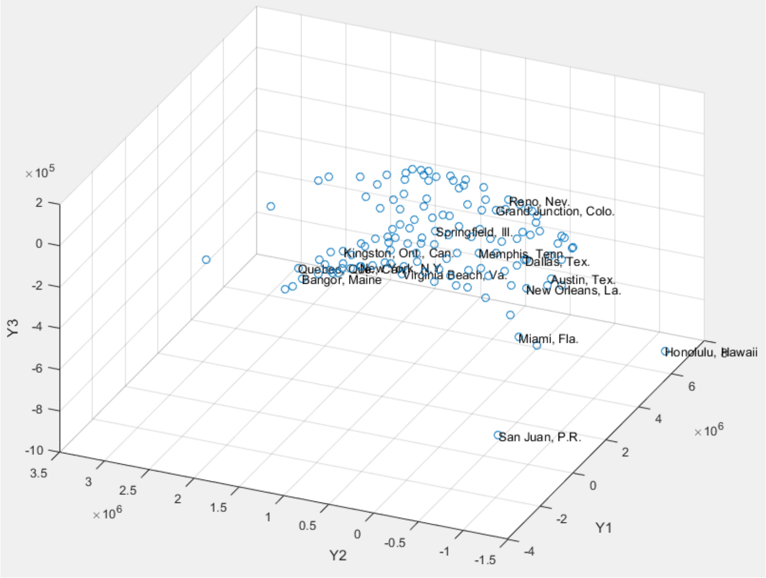

Figure 2.4: 3D US manifold found with ISOMAP.

Figure 2.5 shows the scree plot for the eigenvalues found for the World map. As it can be seen, the first three eigenvalues carry most variance, and therefore should be considered for creating the final manifold. This is also confirmed with the 2D and 3D manifold plots.

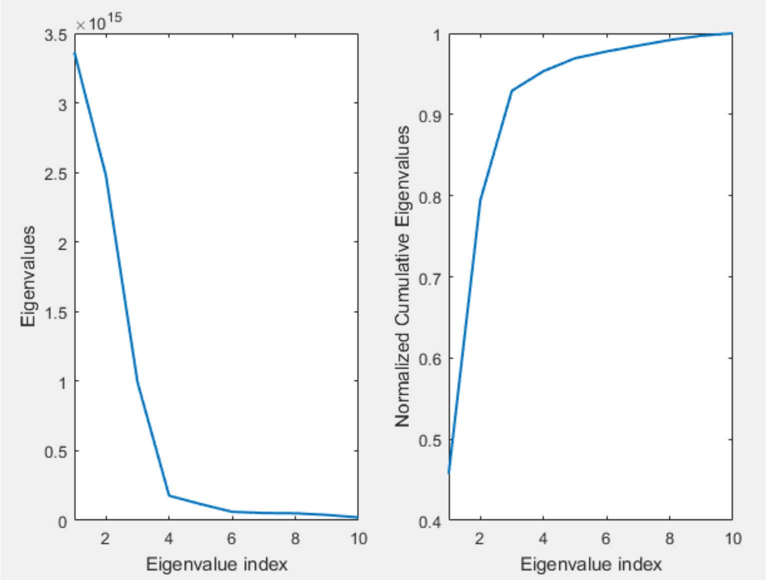

Figure 2.5: Eigenvalues found by using ISOMAP into the World map data.

The 2D manifold for the World dataset is shown in Figure 2.6. As it can be seen, most of the distances are preserved and make sense in the 2D plot. If we were to divide the regions into clusters (Americas, Europe+Africa and Asia+Oceania), most of the cities would be correctly assigned. Figure 2.6 shows the 3D manifold found for the World data. The manifold found resembles the world globe.

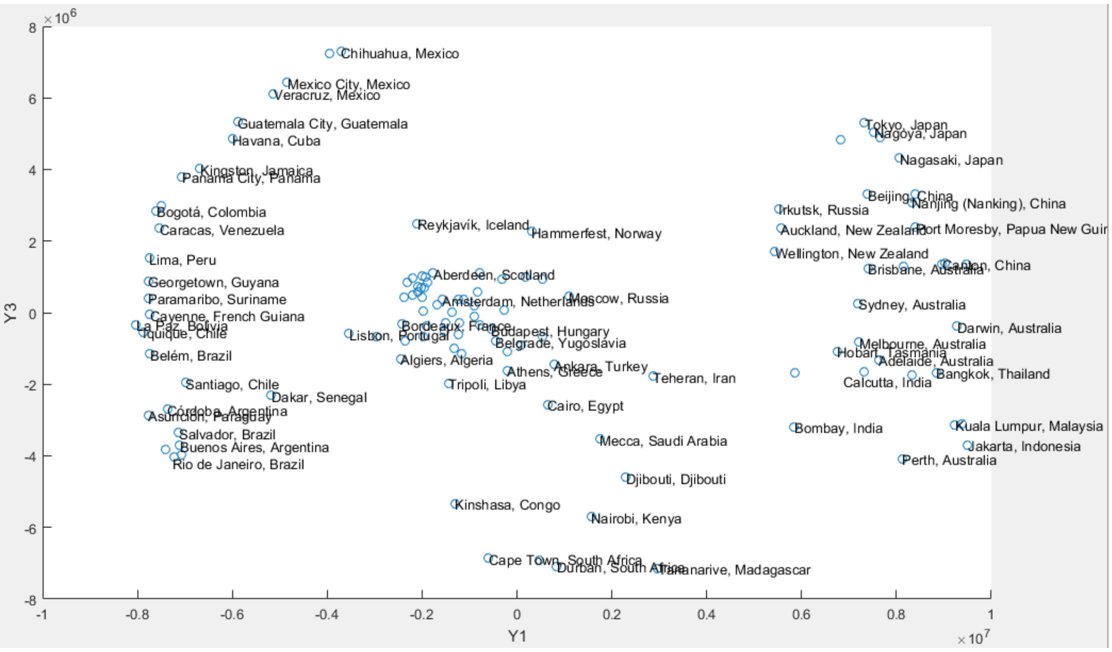

Figure 2.6: 2D World manifold found with ISOMAP.

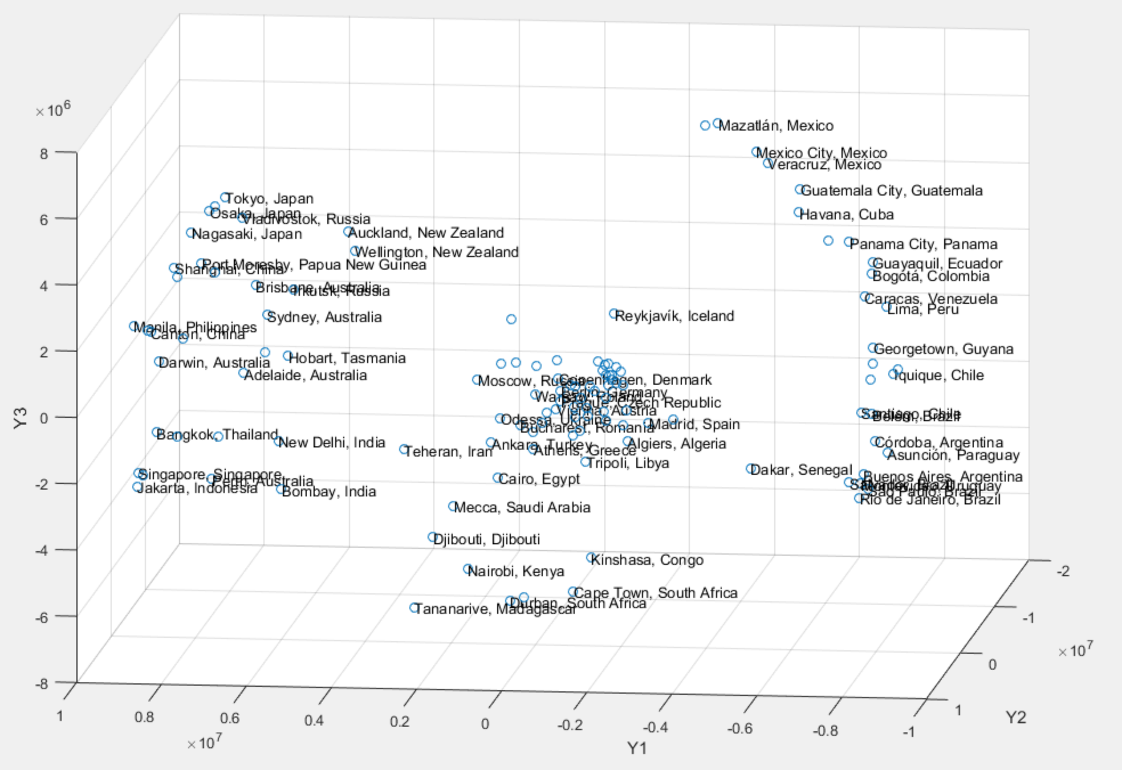

Figure 2.7: 3D World manifold found with ISOMAP.

Figures 2.8 shows zoomed in regions found in the 3d World manifold. Figure 2.8a shows a zoom into Europe, northern Africa and part of the middle east. As one can notice, Berlin, Prague and Vienna are close to one another, as well as Odessa, Bucharest and Ankara. Figure 2.8 shows the cities in America. Rio de Janeiro, Buenos Aires and Asunción are really close, but far from another cluster formed by Veracruz, Mexico City and Guatemala City.

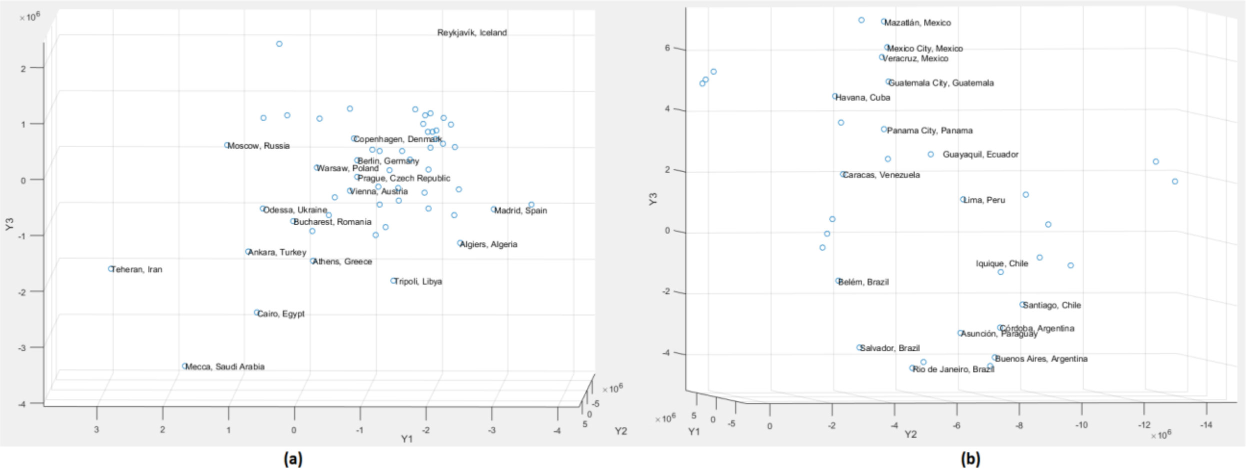

Figure 2.8: Portions of the manifold shown in Figure 2.7. (a) shows Europe, northern Africa and part of the middle east and (b) shows America.

Problem 3

The reconstructed images using codebooks for values of k = 1, 2, …, 10 are shown in Figure 3.1. As it can be seen, the higher the value of k, the more details can be seen in the reconstructed image. For k = 1, the image reconstructed is fully green (the most common color in the original image), for k = 2, the images are reconstructed with green and white (the two most common colors in the original image), and so on. This is done so that the error decreases as the length of the codebook increase. Therefore, the codewords that emerge as the length of the codebook increases are those that will make the error small. For that to happen, the codeword chosen has to be a color similar to the next set of more noticeable colors in the image – first is green, second is white, third is purple (red+blue), and so on. The changes seen as k increase can be seen in Table 3.1.

Table 3.1: Details shown in the reconstructed images for particular values of k.

k	Details
1	Green field
2	Partial Real Madrid uniform
3	Partial Barcelona uniform
4	Players' skin color and goal area field details are shown
5	Players' shadows can be seen
6	Barcelona uniform main colors (red and blue) are shown
7	Shadows have more details
8	Players' hairs are properly shown
9	Barcelona uniform details are improved. Faces have more details
10	Shadows details are improved. Penalty box line quality is improved

Image 3.2 shows the sum-squared-error (SSE) as a function of k. The image was subsampled with (2:1) ratio to speed up computations. As one can notice, the SSE decreases as the values of k increases. This is expected, since the k-means criterion function is to minimize the sum-squared-error and, as the number of k increases, the SSE can be further minimized.

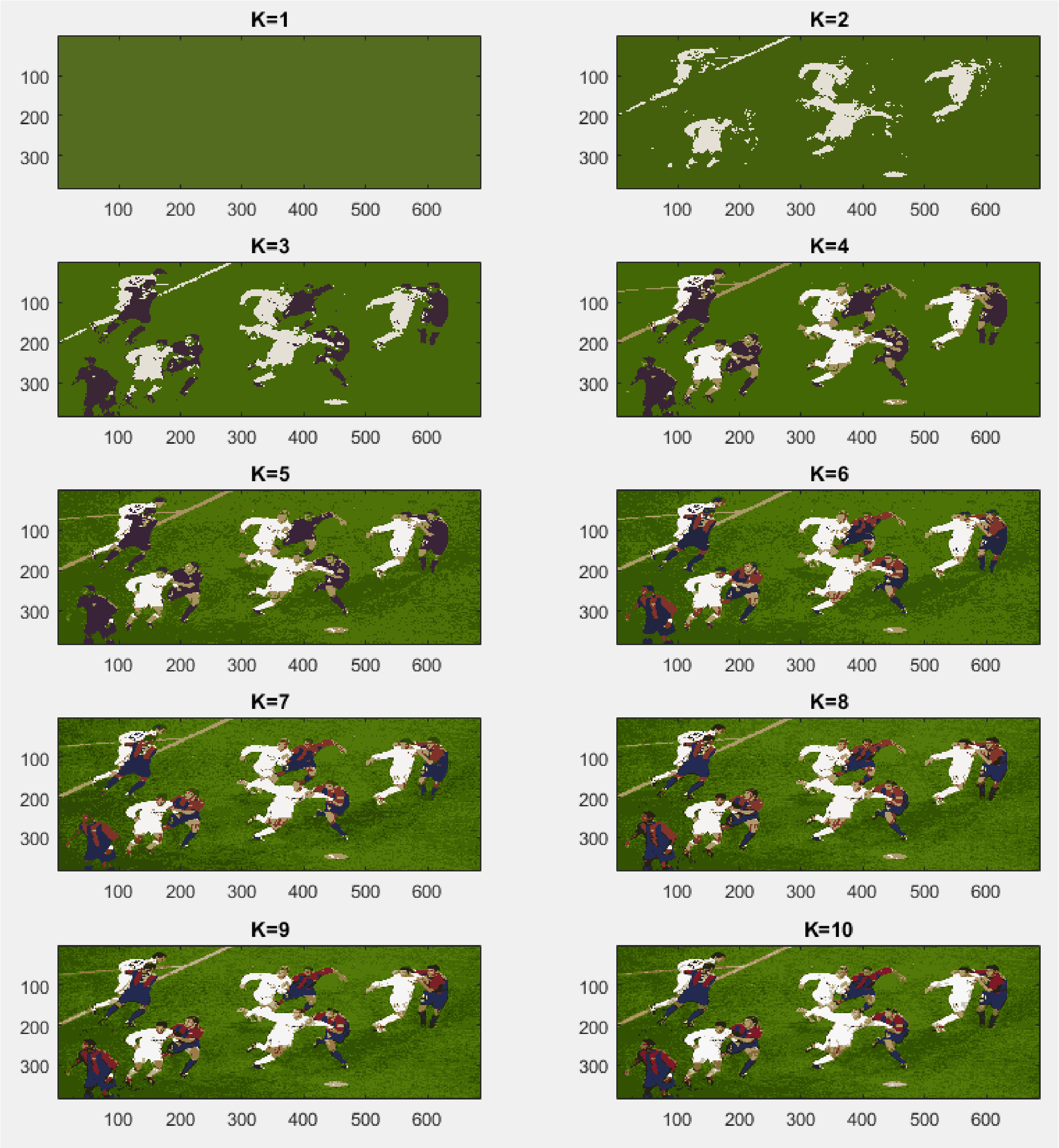

Figure 3.1: Images reconstructed using codebooks for values of k = 1, 2, ..., 10.

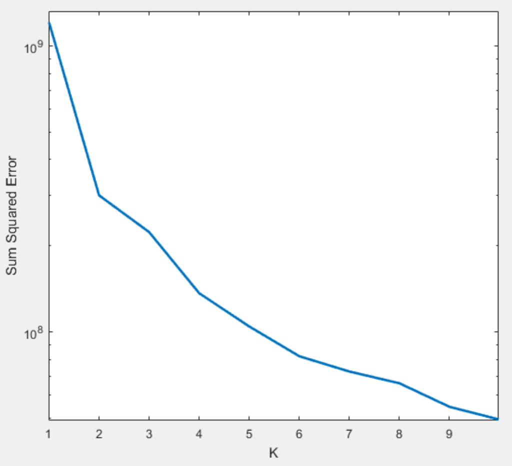

Figure 3.2: Sum-squared-error as a function of k. The image provided was subsampled with ratio (2:1) to speed up computations.

Problem 4

3 Nearest Neighbors Sequential Forward Selection Wrapper

The dataset provided has several missing points. Therefore, the first step for solving this problem was to preprocess the data and handle missing data. The following steps were performed: compute the mean value for each feature (missing points were not considered in this step) of the dataset. Replace all missing data with the respective mean value calculated in the previous step.

For solving this problem, I first used a wrapper to select the subset of features for the data set provided. The classification algorithm used in the wrapper was 3 Nearest Neighbors (3NN). The search procedure employed was sequential forward selection (SFS). For determining the features subset, I used subsampling with 10 experiments to estimate subset performances. Also, for estimating the performance of each classifier on the reduced complexity data, three-way data partition with 30 subsampling experiments was used.

The average accuracies for different classifiers are shown on Table 4.1. As it can be seen, for 3NN SFS wrapper and using 1NN as classifier, the accuracy was the best one among all when the number of features M = 10. Also, the reason why the quadratic classifier performs poorly must be related do the intrinsic data characteristics – most likely the underlying data distribution is not Gaussian. Moreover, for most of the cases, as the number of K increases, the performance decreases – 1NN has shown the best overall results.

Table 4.1: Average accuracy (%) found for different classifiers for the reduced complexity data found with the 3NN SFS wrapper. Number of iterations = 30.

	SFS. M = 6	SFS. M = 7	SFS. M = 8	SFS. M = 9	SFS. M = 10
Quadratic classifier	30.03	29.33	30.22	29.76	29.41
1NN	77.65	75.11	78.61	83.21	**84.30**
2NN	76.44	73.98	77.74	80.37	81.72
5NN	77.31	74.58	78.22	81.14	83.96
20NN	72.39	68.97	70.41	73.82	75.11
50NN	64.44	63.11	63.12	65.48	65.88
100NN	55.16	53.62	52.08	50.15	53.76
200NN	43.11	45.57	44.47	42.90	42.70

The features selected for the best case (3NN SFS, 1NN, M = 10) were: Carbohydrate (g/100g), Sodium (mg/100g), Total folate (µg/100g), Zinc (mg/100g), Riboflavin (mg/100g), Food Folate (µg/100g), Thiamin (mg/100g), Vitamin B6 (mg/100g), Vitamin B12 (µg/100g), Folate (µg diet folate equivalent /

100g). The accuracy achieved by testing the best configuration on the test data provided was **84.12%,** a value similar to the one found during the validation phase.

Figures 4.1 and 4.2 show the scatter plots for some combinations of features after applying SFS. As it can be seen, especially when feature 6 is considered, the classes are relatively scattered around determined regions. Therefore, the reduced complexity data set provides good discriminatory information, making it possible to obtain relatively good classification rates (~84%) even when using a simple classifier such as 1NN.

The corresponding code for this problem was copied to unix.cse.tamu.edu and tested.

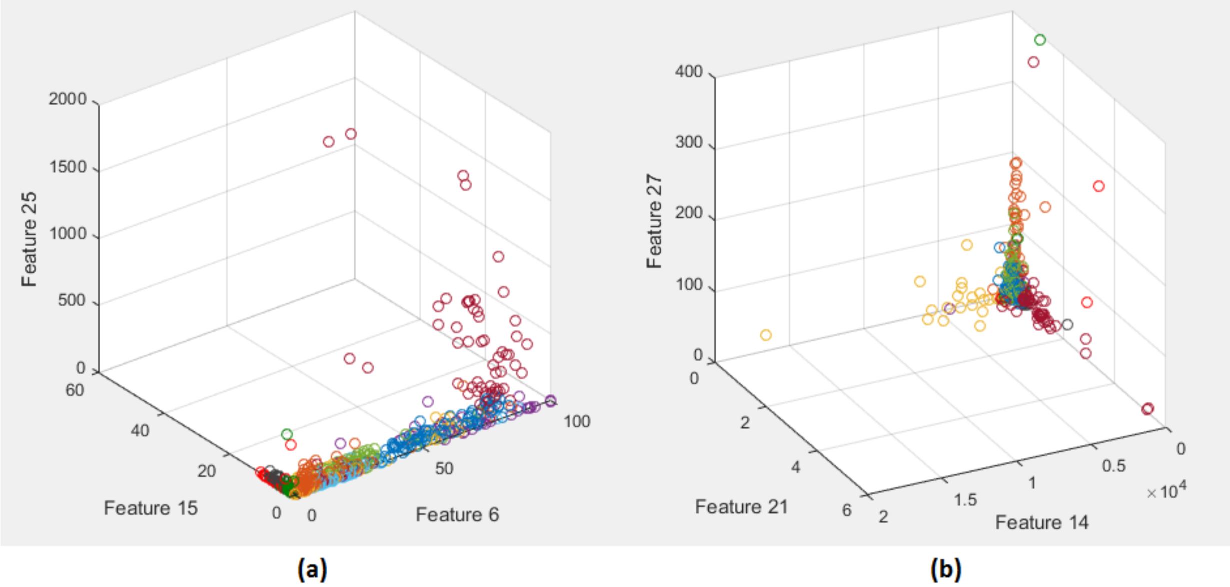

Figure 4.1: Scatter plots for features selected from the data provided. (a) shows the 3D scatter plot for feature 6, 15 and 25 and (b) shows 3D scatter plot for features 14, 21 and 27. Different colors represent different classes.

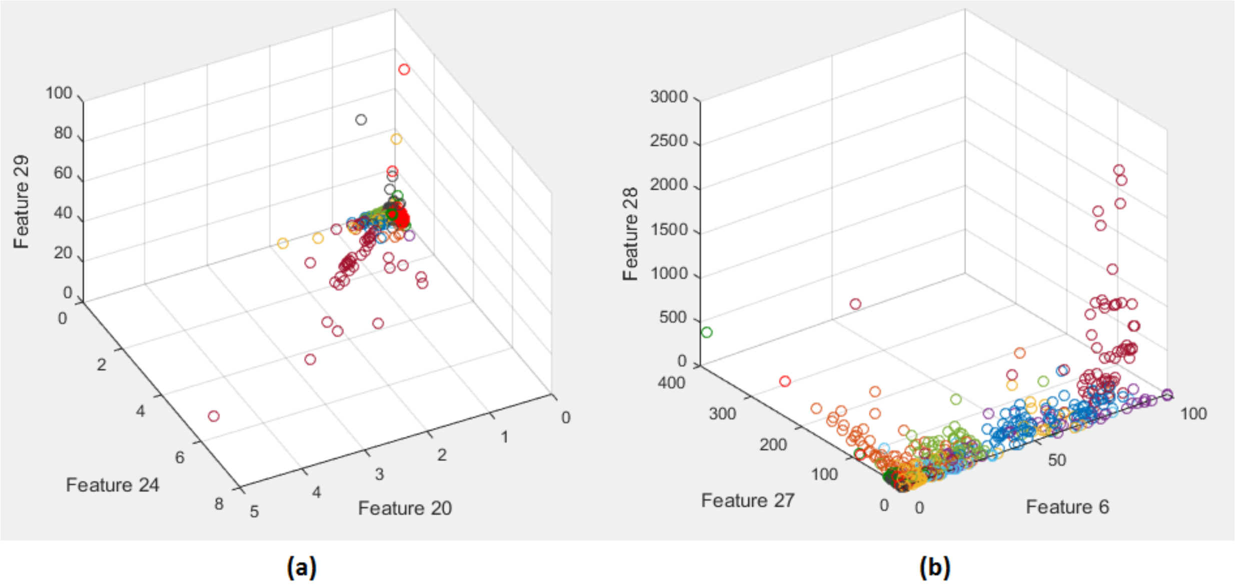

Figure 4.2: Scatter plots for features selected from the data provided. (a) shows the 3D scatter plot for feature 20, 24 and 29 and (b) shows 3D scatter plot for features 6, 27 and 28. Different colors represent different classes.