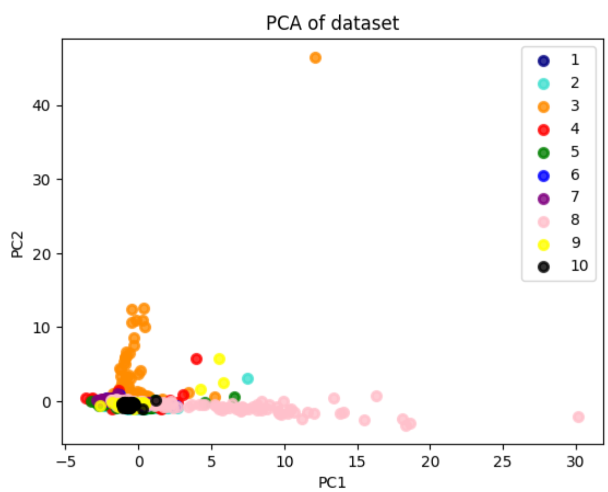
Problem 2 – dimensionality reduction (30%)

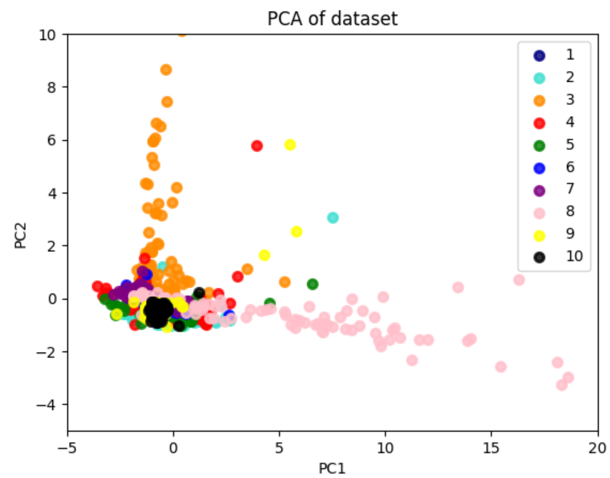
References for code:

- On how to apply [LDA and PCA](#) to datasets.
    - More on how to graph [LDA](#).
-

<u>10% credit. Perform PCA on the training set and generate a 2D scatterplot. To facilitate analysis, please color code each example according to its class label (even though PCA does not use class labels).</u>

We first perform PCA on the training set and generate the 2D scatterplot of the training set following dimensionality reduction. Coloring each example according to class label gives us the following, with the second graph being the zoomed in PCA model:





Interpret the results.

<u>*What information do the principal components capture?*</u>

- Principal components capture the directions of maximal variance, and in cases where PCA provides good separability, that means that variance is in difference of class means. However, in this case, PCA does not provide good discrimination. In this case, this result aligns with what we already know about PCA, as in the case of this 2d visualization of PCA, there appears to be two directions of variance in both the x and y direction. PC2 appears to show the direction of highest variance that favors visualization of class 3, or Soups, Sauces, and Gravies, as in the direction of PC2, there is a good spread of data points identified as class 3. PC1 appears to show direction of highest variance that favors visualization of class 8, or Poultry Products, which allows for us to see the clustering of the class.

- PC1 captures the most variation and PC2 captures the second most variance. The PCs are created by combining one or more variables to capture the directions of variance in the data, which helps us to better understand the data visually and undergo dimensionality reduction.

Is there any class separability?

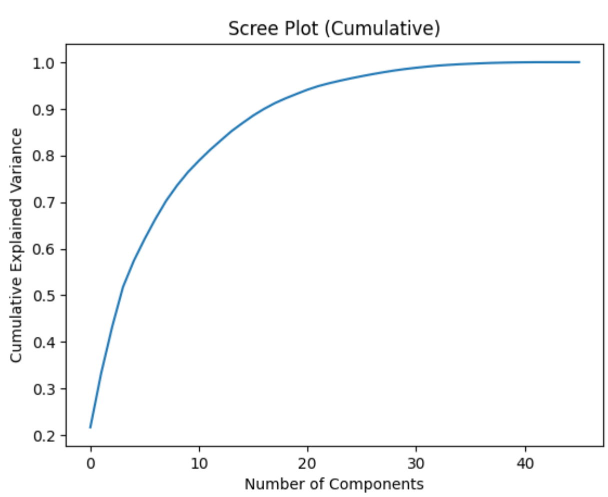
- In the case of applying PCA as a dimensionality reduction technique to the data, there is some class separability, but not much, as though we are able to somewhat distinguish between the classes 3 and 8, PCA does not prioritize separability of class.The other classes are clustered, meaning that those PCs chosen to graph the scatter plot do not properly separate the data into distinguishable classes that we can identify.

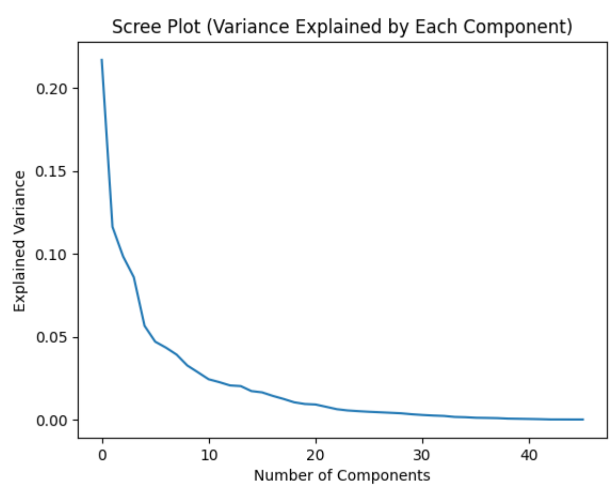
Is there any organization of classes into broader categories of food? …

- In terms of organization of classes into broader categories of food, there is not much organization, even when the scatter plot is zoomed in on. It is already difficult to determine distinguishable separation of classes even for class 3 and class 8, as there is still overlap between the two classes on the 2d scatter plot. Because there is no separability and organization of classes in the first place, it is impossible to organize classes into broader categories.

5% credit. Generate a scree plot and interpret the results. How many PCs are required to capture 95% of the variance in the data?

Generating a scree plot for the PCA training set (cumulative variance) gives us the following:





In order to capture 95% of the variance in the data, we thus require 23 components. Overall, this reflects the ineffectiveness of PCA for dimensionality reduction when it comes to class separation, as we require many components to capture the variance in the data. Overall, PCA might not be the best way to reduce our data.

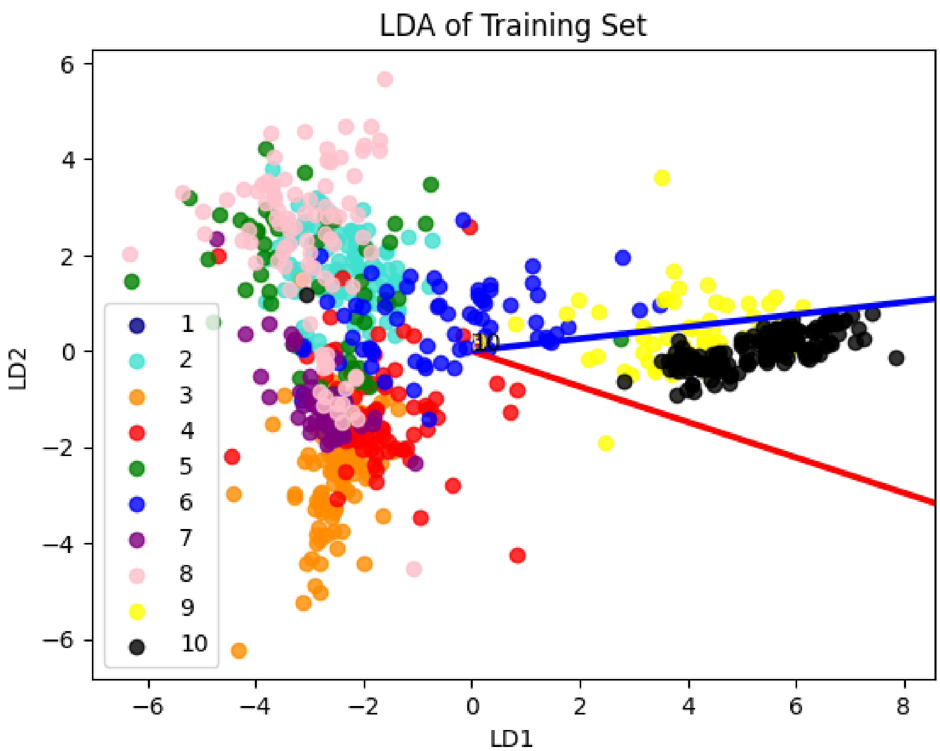
Interpret the results.

Does the scree plot suggest there is high or low collinearity between features?

- After around 5 components, we get diminishing returns on the amount of variation each component explains. There is a clear "elbow" point on the graph which shows that beyond the point, the eigenvalues become smaller. Because the cumulative explained variance increases slowly as we add components, as the elbow in cumulative variance starts at around components 5-7, the scree plot suggests that there is high collinearity between features. Additional components past 5 are not making significant contributions to explaining variance, thus, the scree plot above shows that there is high collinearity between features.
- Thus, this tells us that PCA is not a good fit as a dimensionality reduction method.

10% credit. Repeat part a. using Fisher's LDA. In this case, you will use the training data (x1.csv, c1.csv) to obtain the LDA eigenvector matrix (w1). Then, use matrix w1 to project the training examples (x1.csv), and generate a scatter plot. Finally, use matrix w1 to project the validation samples (x2.csv). As in problem 1, please color code the scatterplots according to class labels.

When repeating part a. For Fisher's LDA, we get the following eigenvector matrix attached below in the index for the sake of succinctness: Using matrix w1 to project the training examples samples gives us the following scatter plot, with eigenvectors graphed:



Interpret the results.

Discuss your findings.

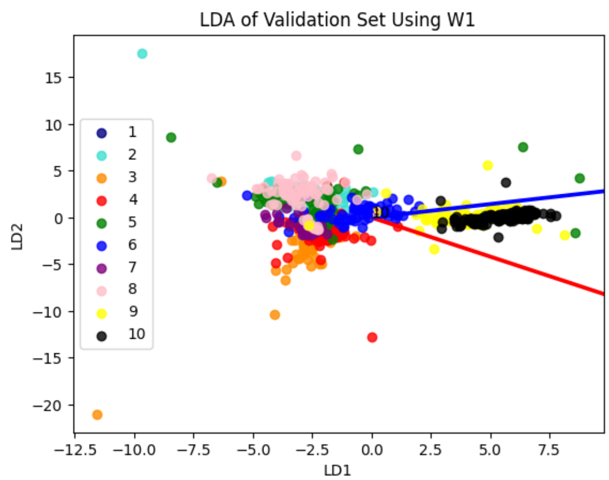
- Following LDA projection of the training data into a 2D scatter plot, we get the above. Overall, it looks like LDA is a better dimensionality reduction strategy as it is providing better separation of classes than that of PCA. There are now more distinguishable clusters that we can identify as classes, and though there is some overlap between classes, the result upholds expectations regarding LDA as LDA seeks to preserve as much classification information as possible.

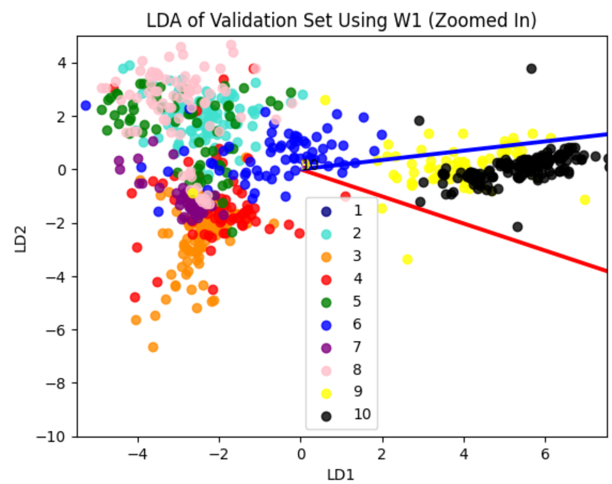
Does the LDA scatter plot for training data show any structure? Describe that structure.

- The LDA scatter plot does show structure, as it appears that there are more distinguishable clusters that can be identified as classes. In addition to this, there appears to be organization of classes into broader categories of food. Though we are visualizing the data using only the first two components, there is still enough information to see the general mean of each class, which is one of the requirements for LDA to be successful.
- For example, classes 9 and 10 are in distinguishable clusters, but they are also located generally close together. This organization structure makes sense as class 9 is Beef Products and class 10 is Lamb, Veal, and Game Products are close together. This aligns with the fact that beef and lamb have [similar nutritional profiles.](#)

- In addition to this, classes 2 and 6 are organized in a way where they are close to each other, which aligns with our understanding of fruits and vegetables, which are the classes that are located close to each other in the 2d scatter plot component projection. Because fruits and vegetables have similar [nutritional profiles](), this scatter plot is organized in a way that reflects real life nutritional values.

Then, using matrix w1 to project the validation samples gives us the following scatter plot, with the eigenvectors from the previous computation also graphed.

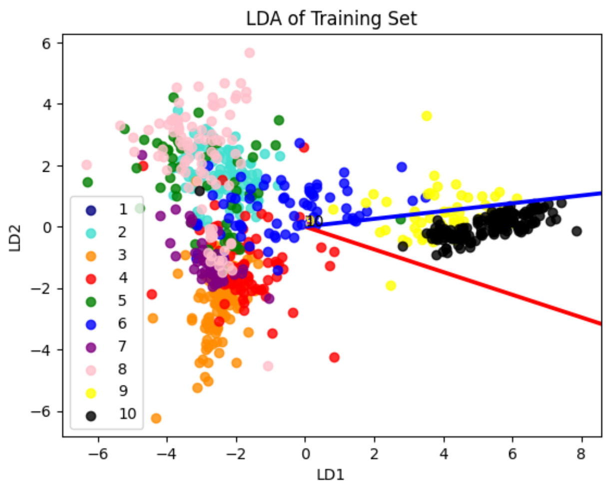


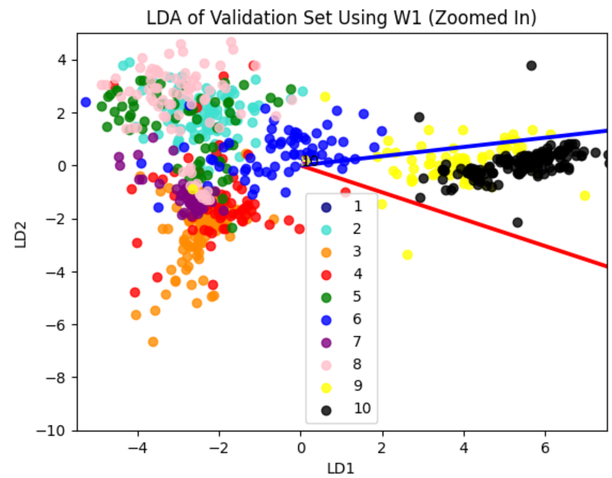


Interpret the results.

<u>Do these results hold when you project the (unseen) validation data? Please elaborate.</u>

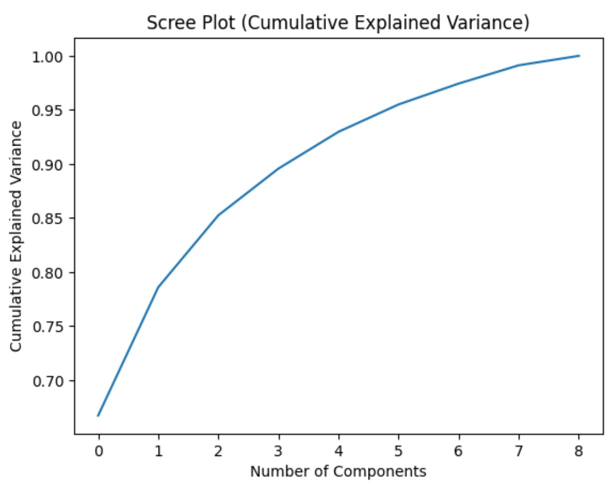
- When we project the unseen validation data, the results also hold, as the structure of the LDA scatter plot for the validation data similarly matches that of the Training set. This means that the data of the validation set similarly models a situation that matches that of the training set. This doesn't tell us if the data properly models real life, however. When comparing the structure of the projected validation set and the projected training set, we can see that there are a lot of similarities, including the general outline of the classes.

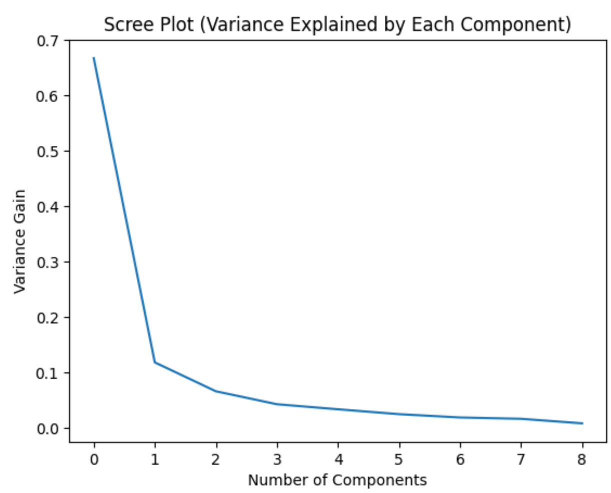




5% credit. Repeat part b. using Fisher's LDA.

Generating a scree plot of for the LDA training data gives us the following graph:





In order to capture 95% of the variance in the data, we thus require 6 components. Overall, this reflects the effectiveness of LDA for dimensionality reduction when it comes to class separation, as we require a lot less components to capture variance in comparison to that of PCA.

Investigating further shows us that in order to capture 100% of variance in the data, we require 9 components. This will be taken into account in problem 3.

Interpret the results.

Does the scree plot suggest there is high or low collinearity between features?

- The scree plot above suggests that there is low collinearity between features, as after the first component, there is diminishing returns on each component capturing variance in the data. Since most of the variance is explained by a low number of components (or at least much lower than that of PCA) this suggests that there is low colinearity between features.

- This is a good sign, as this implies that we can reduce the dimensionality of the dataset without losing a lot of information.