Bilkent University

Computer Science Department

CS464 Introduction to Machine Learning
Spring 2020
Homework 1

Muhammed Musab OKŞAŞ, 21602984

Question 3.1

In this question, given test dataset is evaluated according to their trained results.

$$\mathbf{P}(X_j \mid Y = 1) \quad \mathbf{P}(X_j \mid Y = 0) \quad \mathbf{P}(Y = 0) \quad \mathbf{P}(Y = 1)$$

Evaluation included two possible predictions for label 1 and 0 according to given 8-mer order. Naïve Bayes approach is applied to these orderings of 8-mers.

For the label 1, possibilities of amino acids of 8-mer from 1 to 8 in one hot encoded way given that label is 1 are added in log forms. Possibility of label is 1 is added in the log form to result that is found before.

Similar approach for the label 0 is applied for the current 8-mer. Two results are compared and final prediction for this 8-mer is made according to this comparison.

There are four possibilities of this prediction. It can be false negative, true negative, false positive, true negative.

$$accuracy = \frac{\mathrm{tn} + \mathrm{tp}}{\mathrm{tn} + \mathrm{tp} + \mathrm{fn} + \mathrm{fp}}$$

The results that I found according to the given test dataset length of 528 are:

tn: 422 tp: 79
fn: 18 fp: 9

accuracy: 0.9489 or 94.89%

Question 3.2

Given gag sequence is modified into 493 x 160 one hot encoded version and printed in “one_hot_encoded_gag.txt”.

This modified data is tested according to the trained dataset. Indices expected to be cleaved are shown below.

Indices to be cleaved: 5 6
Indices to be cleaved: 42 43
Indices to be cleaved: 60 61
Indices to be cleaved: 79 80
Indices to be cleaved: 131 132
Indices to be cleaved: 183 184
Indices to be cleaved: 196 197
Indices to be cleaved: 215 216
Indices to be cleaved: 295 296
Indices to be cleaved: 315 316
Indices to be cleaved: 320 321
Indices to be cleaved: 341 342
Indices to be cleaved: 342 343
Indices to be cleaved: 362 363
Indices to be cleaved: 366 367
Indices to be cleaved: 376 377
Indices to be cleaved: 447 448
Indices to be cleaved: 465 466
Indices to be cleaved: 467 468
Indices to be cleaved: 482 483

Question 3.3

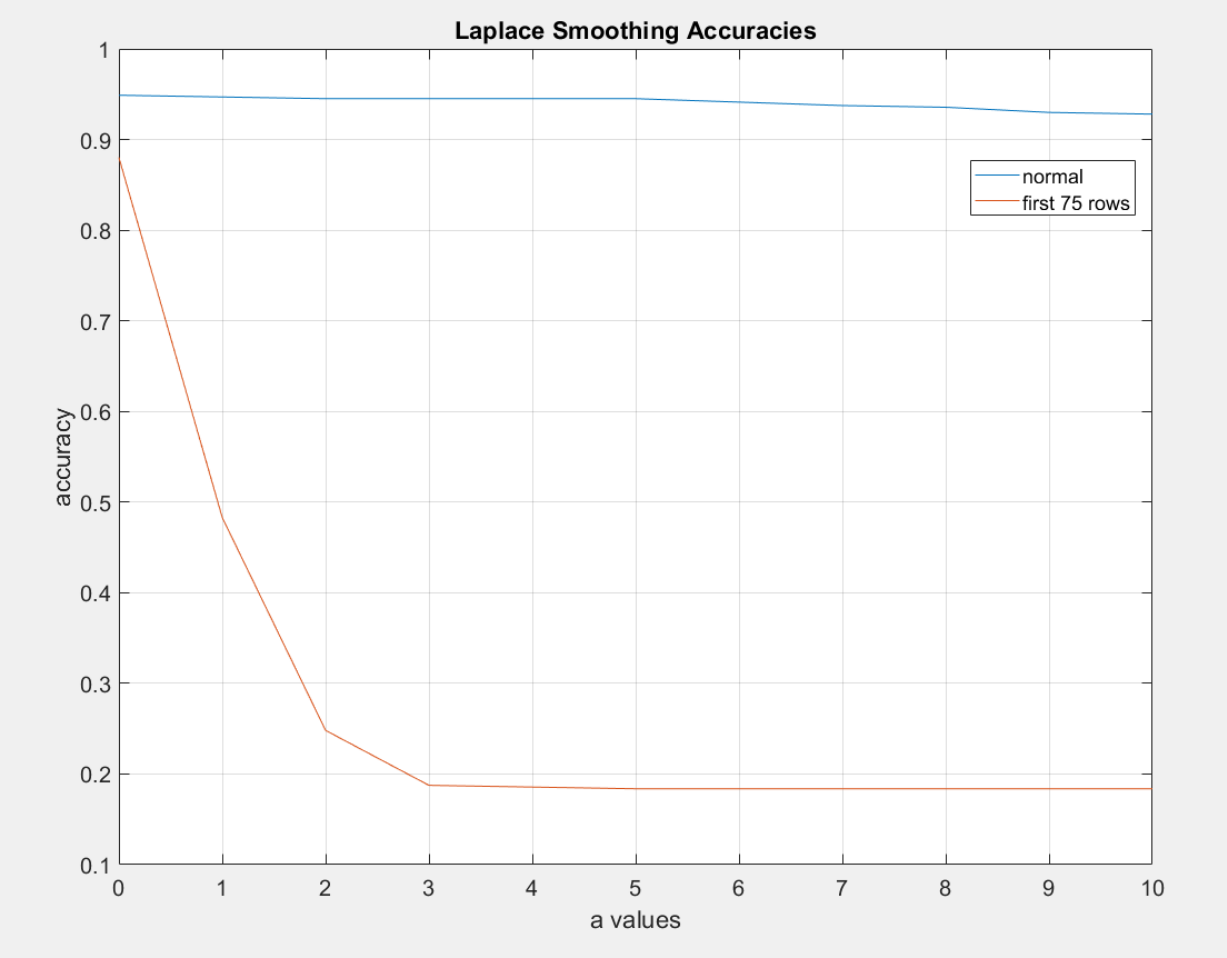
For the given gag sequence

The 8-mer labeled as class 1 with highest probability is: arvlaeam

The 8-mer labeled as class 0 with lowest probability is: rkkgcwkc

Question 3.4

Laplace smoothing is applied for both all of the data and first 75 rows of the data and the results are as shown in the chart below:



Accuracy for all of the training set goes from 94.89% to 92.80%

Accuracy for the first 75 rows of the training set goes from 88.02% to 18.37%

There are some points to make:

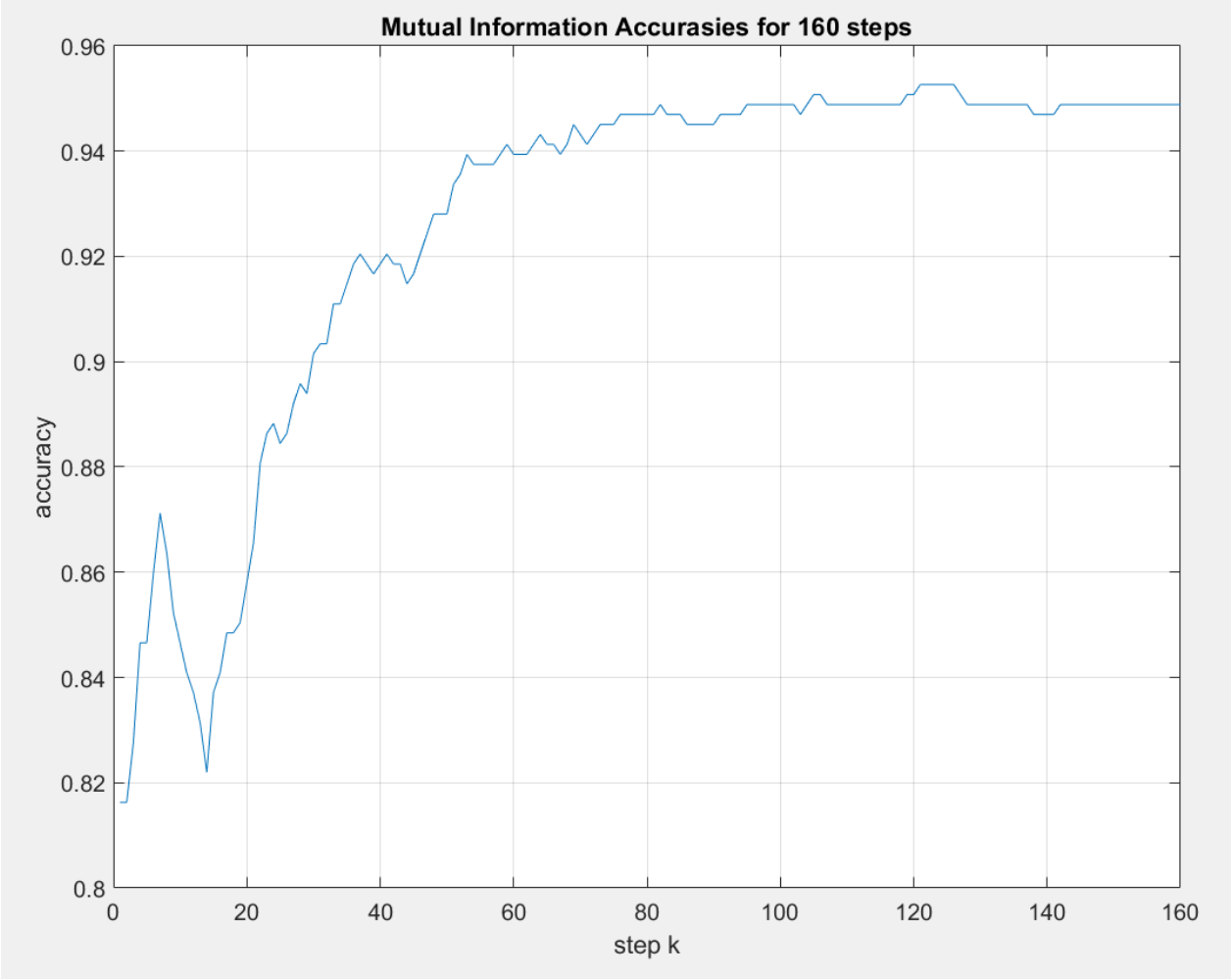
- Considering that accuracy without Laplace Smoothing is 0.9489, it can be stated that when a value is small its effect is considerably less.
- When a value increases accuracy tends to decrease according to the chart.
- When training set is small the effect of Laplace Smoothing is dramatically high. It can be seen from the chart that after some value of a, our trained data predicts same label for all test set. This makes the model unreliable.

Question 3.5

Indices of features according to their mutual information with the label is given in sorted order:

53, 62, 44, 70, 65, 69, 117, 127, 2, 19, 85, 89, 160, 34, 101, 64, 42, 104, 98, 80, 102, 73, 93, 81, 67, 90, 118, 122, 59, 153, 96, 4, 113, 55, 75, 66, 41, 5, 106, 99, 56, 147, 84, 103, 95, 24, 77, 154, 97, 110, 116, 22, 29, 119, 36, 6, 150, 133, 144, 155, 54, 112, 57, 45, 13, 35, 61, 137, 114, 46, 26, 142, 109, 39, 92, 63, 71, 121, 72, 88, 50, 159, 135, 3, 82, 78, 123, 15, 111, 151, 52, 40, 49, 9, 16, 17, 51, 83, 68, 21, 129, 136, 7, 100, 157, 11, 47, 37, 108, 94, 76, 10, 87, 130, 107, 23, 156, 148, 125, 74, 141, 31, 146, 115, 128, 143, 131, 149, 27, 38, 33, 145, 1, 152, 86, 79, 105, 25, 32, 124, 58, 18, 43, 139, 8, 30, 132, 120, 48, 158, 140, 12, 138, 20, 28, 126, 91, 134, 14, 60

According to this order, the training data is trained for the first k feature (k goes from 1 to 160) and the plot is shown as below considering their accuracies:



According to the given plot above there are some points to make:

- Even with just one feature 81.62% accuracy level can be reached.
- Just with seven feature 87.12% accuracy level can be reached.
- Even though mutual information of features with the label is high there is no indication that these features should correlate together. This is can be seen from step 8 to 14 since the accuracy decrease as we increase feature number.
- Maximum accuracy reached is 95.26% at the step 121. This shows that not always more feature is the best since the accuracy reached with 121 features is better than the accuracy with 160 features.

Question 3.6

Principal Component Analysis is applied to the given dataset. Instead of finding covariance matrix and finding its eigenvectors and eigenvalues, Singular Value Decomposition is applied. If we say X is centered version of given dataset with size of m x n, and C is covariance matrix. The reason to apply SVD is to avoid $C = X^TX/(m-1)$ matrix multiplication.

Let

$C=VLV^T$ where V is eigenvectors and L is eigenvalues

$X=USV^T$ (Singular Value Decomposition) where U is a unitary matrix and S is the diagonal matrix of singular values s_i

$$\mathbf{C} = \mathbf{V}\mathbf{S}\mathbf{U}^\top\mathbf{U}\mathbf{S}\mathbf{V}^\top/(n-1) = \mathbf{V}\frac{\mathbf{S}^2}{n-1}\mathbf{V}^\top$$

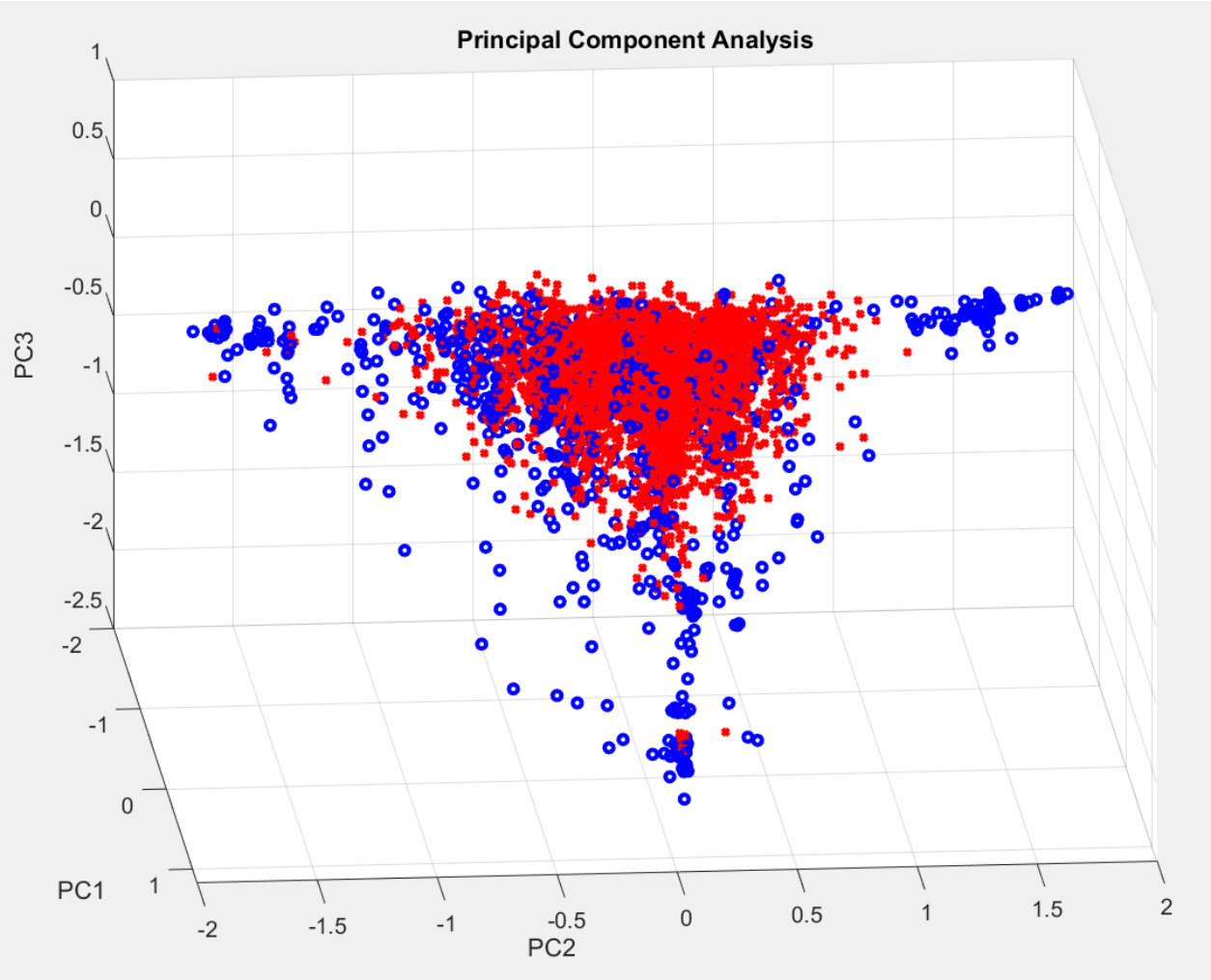
Therefore, principal components are

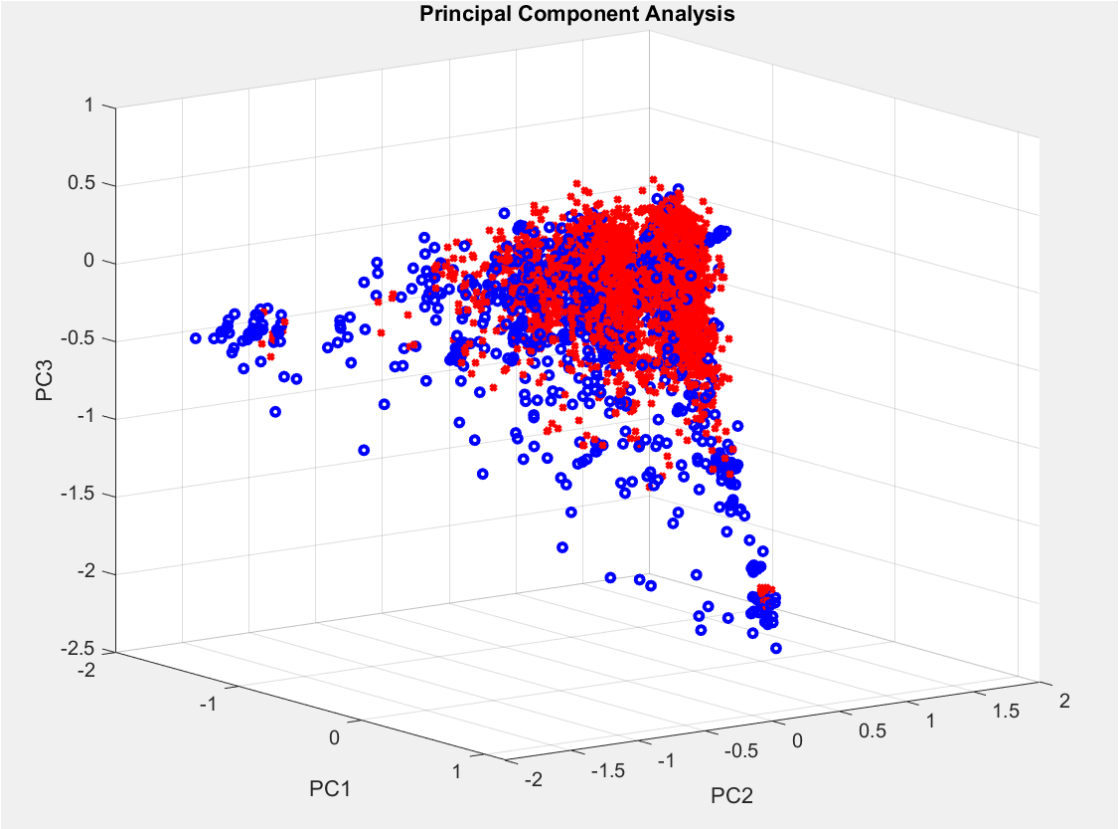
$$\mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{S}\mathbf{V}^\top\mathbf{V} = \mathbf{U}\mathbf{S}$$

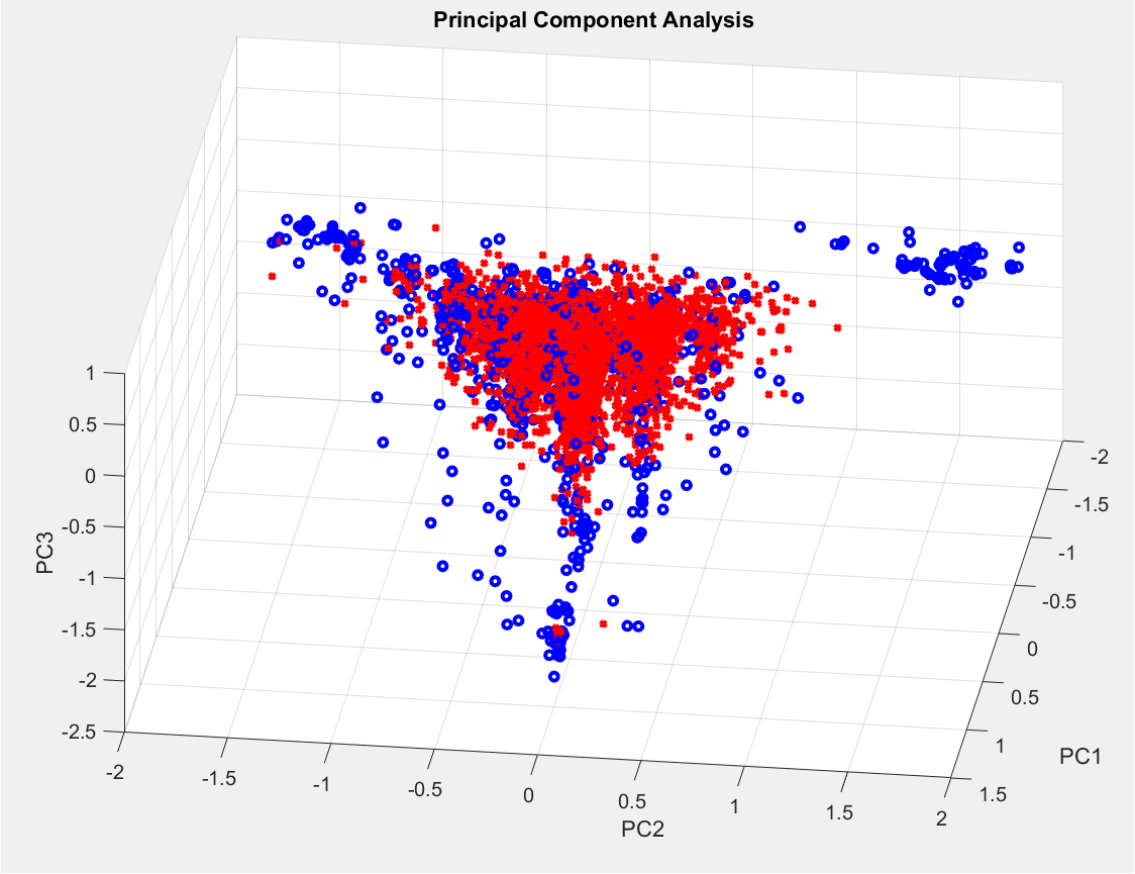
Eigenvalues of covariance matrix via λ_i (diagonal values of L) is

$$\lambda_i = s_i^2/(n-1)$$

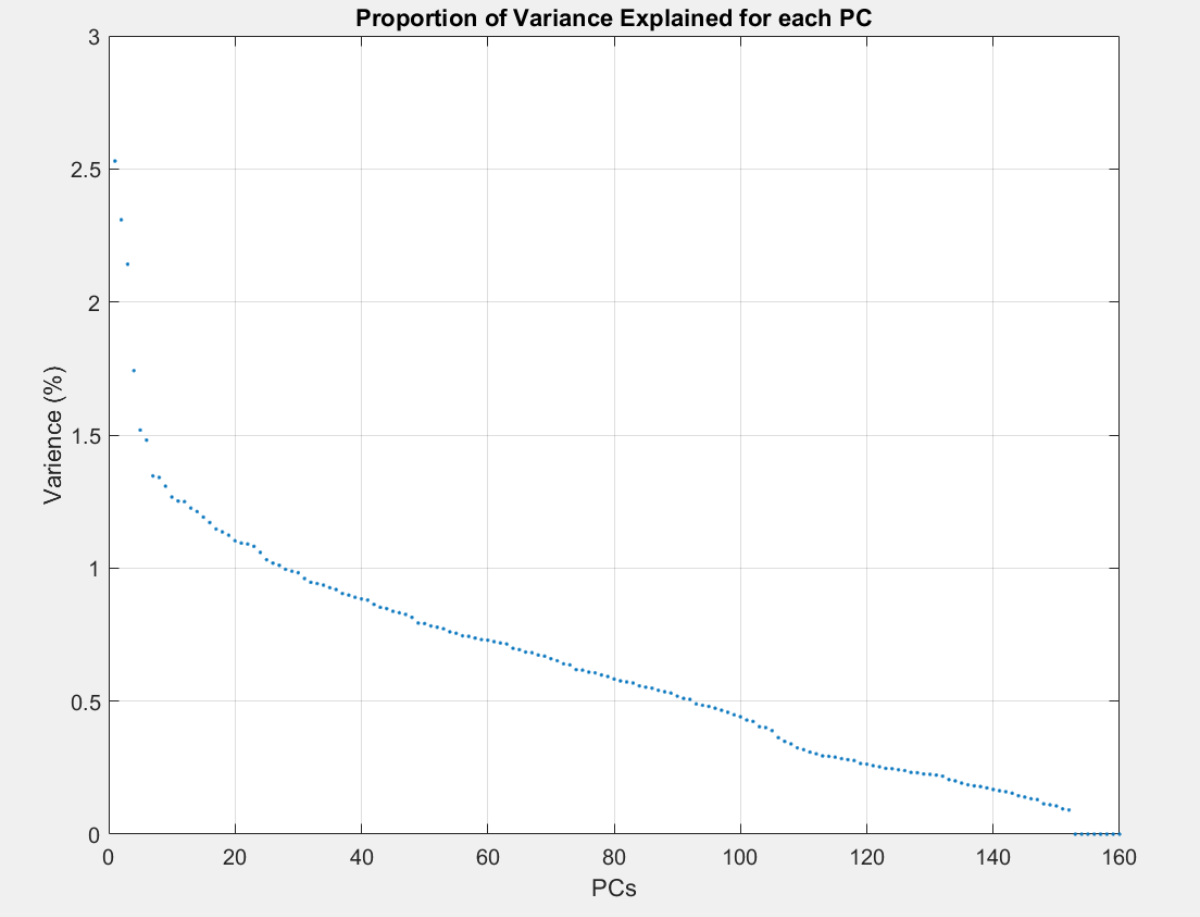
In the lights of equations explained above, 3D projections of amino acid 8-mers by using first three principal components is plotted below. Different views of the same plot are available:




Principal Component Analysis
PC3
PC1
PC2


Principal Component Analysis
PC3
PC2
PC1

Proportion of variance explained (PVE) for the first three obtained principal components is 6.980%.
All percentages of variances for principal components can be seen as shown:



According to the given plots above considering question 3.6 there are some points to make:

- Considering the first three principal component analysis plot, it is clear that the first three component doesn't cluster labels effectively. Label 1 and 0 are still twine together.
- Considering that PVE for the first three obtained component is 6.980%, it is reasonable for PCA to not cluster labels effectively because only small percentage of the variance of the all data is included in the first three PCs.
- As a result, PCA is not feasible to apply to given dataset. The reason for this result could be because our features are one hot encoded meaning that they don't vary much which is fundamental to apply PCA.