Bilkent University

Department of Computer Engineering

CS - 464

Introduction to Machine Learning

Homework 1

Name : Osman Burak İntişah

Id : 2160243

Section: 2

Question 3.1)

After training my Bernoulli Naive Bayes model on training set, I have evaluated my model on test set. In the end, I calculated the accuracy.

My test set accuracy is : **94.88636363636364 %**

Question 3.2)

There would be in total **22** cleave occur.

For finding the exact indices I found the corresponding index of the 8mer and added 4 to it. Because since every 8mer includes 8 aminoacid and if that 8mer is labeled with one, the corresponding place of cleave occur is 4 more than the index of 8mer.

The exact amino acid indices are;

[6, 41, 43, 61, 80, 132, 169, 184, 197, 216, 296, 316, 321, 342, 343, 363, 367, 377, 448, 466, 468, 483]

Question 3.3)

Output:

The 8-mer for which your model assigns it to class 1 with highest probability:

['s', 'a', 'v', 'l', 'l', 'e', 'a', 't']

The 8-mer for which your model assigns it to class 0 with lowest probability:

['w', 'w', 'w', 'w', 'w', 'w', 'w', 'w']

For finding the first answer, for every indices I have found the amino acid which has the highest probability to be there when label is 1. So I found the answer. For finding the second answer, for every indices I have found the amino acid which has the lowest probability to be there when label is 0.Then, I found the answer.

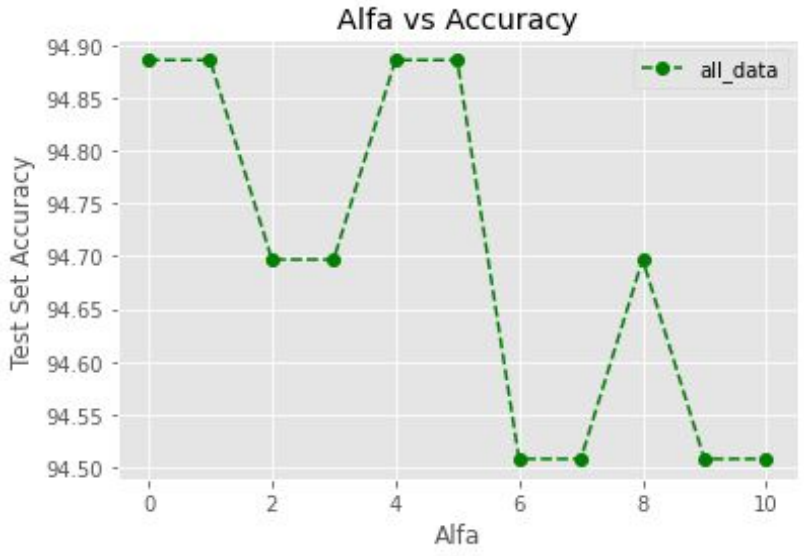
From the first answer, we can understand that for the first index ‘S’ amino acid has the highest probability to, and for the second ‘A’ has the highest when label 1. And it goes like this. These are the amino acids with the highest probabilities to have the label 1.

From the second answer, we can understand that if amino acid ‘W’ is in anywhere in the 8mers it generally labeled 1. So that amino acid 0 has the lowest probability to classifying the data with label 0.

Question 3.4)

3.4.1) Results when training my classifier using all of the training set

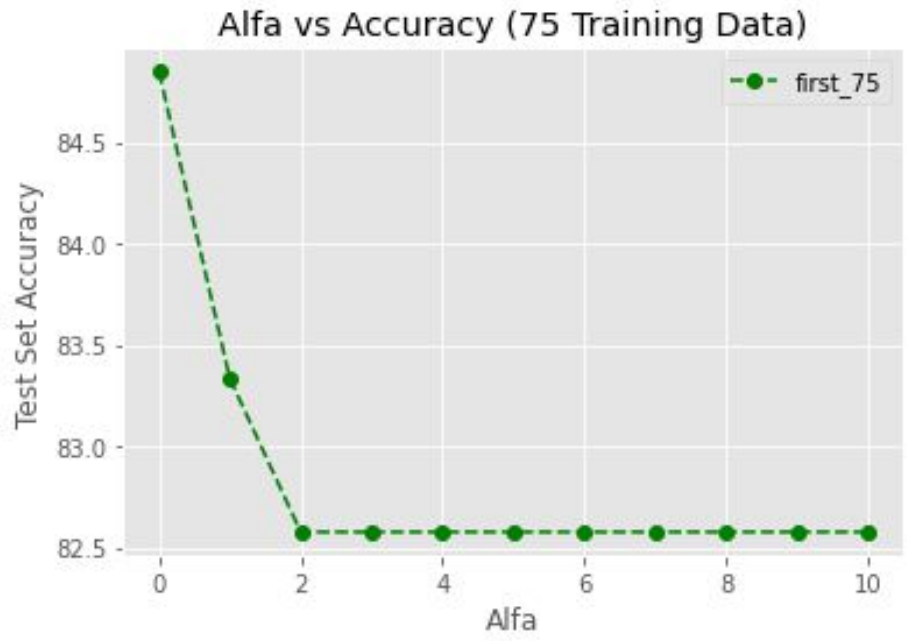
Output:



Plot 1: Alfa vs Accuracy (Using all training set)

The accuracy when alfa = 0
94.88636363636364
The accuracy when alfa = 1
94.88636363636364
The accuracy when alfa = 2
94.6969696969697
The accuracy when alfa = 3
94.6969696969697
The accuracy when alfa = 4
94.88636363636364
The accuracy when alfa = 5
94.88636363636364
The accuracy when alfa = 6
94.50757575757575
The accuracy when alfa = 7
94.50757575757575
The accuracy when alfa = 8
94.6969696969697
The accuracy when alfa = 9
94.50757575757575
The accuracy when alfa = 10
94.50757575757575

3.4.2) Results when training my classifier using only first 75 rows



Plot 2: Alfa vs Accuracy (Using first 75 rows)

The accuracy (when training size is 75) when alfa = 0
84.84848484848484
The accuracy (when training size is 75) when alfa = 1
83.33333333333334
The accuracy (when training size is 75) when alfa = 2
82.57575757575758
The accuracy (when training size is 75) when alfa = 3
82.57575757575758
The accuracy (when training size is 75) when alfa = 4
82.57575757575758
The accuracy (when training size is 75) when alfa = 5
82.57575757575758
The accuracy (when training size is 75) when alfa = 6
82.57575757575758
The accuracy (when training size is 75) when alfa = 7
82.57575757575758
The accuracy (when training size is 75) when alfa = 8
82.57575757575758
The accuracy (when training size is 75) when alfa = 9
82.57575757575758
The accuracy (when training size is 75) when alfa = 10
82.57575757575758

When I applied additive smoothing to the whole datas in my training set, I have found the results in the part 3.4.1. From the output I have provided it is clearly seen that with additive smoothing we cannot have more accuracy than before (which is given in part 3.1). Moreover when we increase the alfa our accuracy is decreasing.

Secondly, when I applied additive smoothing to the first 75 rows in my training set, at the beginning I have found less accuracy (which is understandable since we use less data). Then when I continue increasing alfa value my accuracy started to decrease for 2 alfa points then it stayed the same. Therefore, I concluded that with insufficient data MAP estimates does not change my accuracy.

Question 3.5)

Maximum accuracy value: **95.26515151515152**

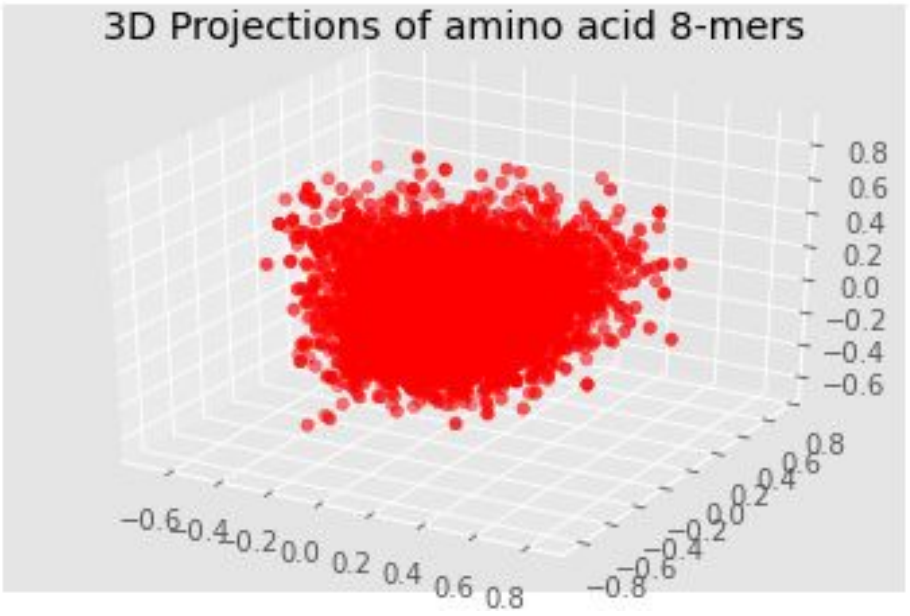
k value: **76**

It is the first k value obtained the maximum accuracy. But I have also reached the maximum accuracy some more k values as well which are; **[77, 78, 79, 80, 81, 96, 97, 98, 99, 100, 101, 104, 105, 120, 121, 122, 123, 124, 125]**

Yes, the maximum accuracy found is higher than the accuracy value you obtained for 3.1.

Question 3.6)

3.6.1)



Plot 2: 3D Projections of amino acid 8-mers

3.6.2)

Proportion of variance explained (PVE):

[0.025294651313728718, 0.023087831826153853, 0.021420185386614846]

3.6.3)

In my opinion applying PCA is feasible for this data set. Because with using PCA we can achieve;

- find the latent features driving the patterns for labeling
- reduce the dimensionality.

These achievements may helps us to find which amino acids or which positions are driving the labeling process. So after PCA is used we can decide on the coefficients for that ones etc.

Additionally, there are many data and many features in this data set. By using this approach we can reduce the noise.

Lastly, visualising the data helps us to understand which features are more effective for labeling.