Name: Mian Usman Naeem Kakakhel

Id: 21701015

Section: 02

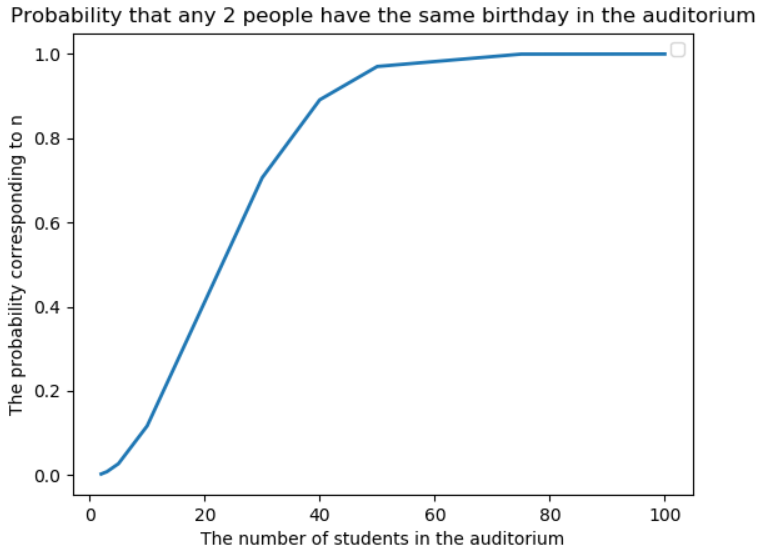
HW 1

03/03/2020

Q1.1) Probability that any 2 students from n have the same birthday.

Since leap years are not considered, there are 365 days in a year.

Now for 2 people to have the same birthday the probability would be $\frac{365}{365} * \frac{1}{365}$. To find the probability of at least 2 people having the same birthday, we can calculate this by (1 - at least two people not having the same birthday). For 2 people not to have the same birthday, we get the probability $\frac{365}{365} * \frac{364}{365}$ further expanding this to 3 people, for 3 people not having the same birthday is $\frac{365}{365} * \frac{364}{365} * \frac{363}{365}$. Thus n people not having the same birthday is $\frac{365!}{(365-n)! * 365^n}$. Finally, at least two people having the same birthday is $1 - \frac{365!}{(365-n)! * 365^n}$.



Q1.2) Find the n where this probability is 100%.

n should be equal to 366 to make sure that probability is 100% because if there are upto 365 people, there is a chance that all of them might be born on all 365 days of the year but when there are 366 people, then it is a must than at least two people will have the same birthday.

Q2.1) Obtain and report the following probabilities:

P(S = disease) : 0.011
P(S = healthy) : 0.989
P(T = positive|S = disease) : 0.94
P(T = negative|S = disease) : 0.06
P(T = positive|S = healthy) : 0.02
P(T = negative|S = healthy) : 0.98

Q2.2) Test returns a positive result, see whether a person is diseased.

Technically we have to find P(S = disease|T = positive).

Since $P(S = disease|T = positive) = \frac{P(S = disease \bigcap T = Positive)}{P(T = Positive)}$ and

$$P(T = positive|S = disease) = \frac{P(S = disease \bigcap T = Positive)}{P(S = disease)}.$$

Thus $P(S = disease|T = positive) = \frac{P(T = positive|S = disease) * P(S = disease)}{P(T = Positive)}$.

The only information missing from this formula is $P(T = Positive)$ and that can be found by using bayes theorem. Thus,

$$P(S = disease|T = positive) = \frac{P(T = positive|S = disease) * P(S = disease)}{P(T = positive|S = disease) * P(S = disease) + P(T = positive|S = healthy) * P(S = healthy)}$$

$$P(S = disease|T = positive) = \frac{0.94 * 0.011}{0.94 * 0.011 + 0.02 * 0.989}$$

Finally, **P(S = disease|T = positive) : 0.343.** Since the probability is less than 0.5 we will not diagnose the patient as diseased.

Q2.2) Number of tests required to diagnose the patient.

Since we are given the information that the first test is positive, we need to find out the number of successive positive tests we should get that would diagnose the patient. Thus, we need to find,

$$P(S = disease|T_1 = positive \bigcap T_2 = positive \bigcap T_3 = positive \bigcap ....\bigcap T_n = positive)$$

Since,

$$P(S = disease|T_1 = positive \bigcap ....\bigcap T_n = positive) = \frac{P(T_1{=}positive\bigcap....\bigcap T_n{=}positive \bigcap S = disease)}{P(T_1{=}positive\bigcap....\bigcap T_n{=}positive)}$$

and,

$$P(T_1 = positive \bigcap ....\bigcap T_n = positive|S = disease) = \frac{P(T_1{=}positive\bigcap....\bigcap T_n{=}positive \bigcap S = disease)}{P(S = disease)}$$

Thus,

$$P(S = disease | T_1 = positive \bigcap .... \bigcap T_n = positive) = \frac{P(T_1 = positive \bigcap .... \bigcap T_n = positive | S = disease) * P(S = disease)}{P(T_1 = positive \bigcap .... \bigcap T_n = positive)}$$

We will come back to the above formula but first calculate its different parts.
Using Bayes theorem,

$$P(T_1 = positive \bigcap .... \bigcap T_n = positive) = (P(T_1 = positive \bigcap .... \bigcap T_n = positive | S = disease) * P(S = disease)) + (P(T_1 = positive \bigcap .... \bigcap T_n = positive | S = healthy) * P(S = healthy))$$

Since all the tests are independent events, we can show,

$$P(T_1 = positive \bigcap .... \bigcap T_n = positive | S = disease) = P(T = positive | S = disease)^n = 0.94^n$$

$$P(T_1 = positive \bigcap .... \bigcap T_n = positive | S = healthy) = P(T = positive | S = healthy)^n = 0.02^n$$

Thus,

$$P(T_1 = positive \bigcap .... \bigcap T_n = positive) = (0.94^n * 0.011) + (0.02^n * 0.989)$$

Finally,

$$P(S = disease | T_1 = positive \bigcap .... \bigcap T_n = positive) = \frac{(0.94^n * 0.011)}{(0.94^n * 0.011) + (0.02^n * 0.989)}$$

If we put n = 1 in the above formula, we get P(...) = 0.343 and when n = 2, we get P(...) = 0.96.
Since there is a 96% chance of a disease we can safely diagnose the patient at **n >= 2**.

Q3.1) Test Set Accuracy: 94.886 %

Q3.2) All the locations where HIV-1 PT would cleave this polyprotein

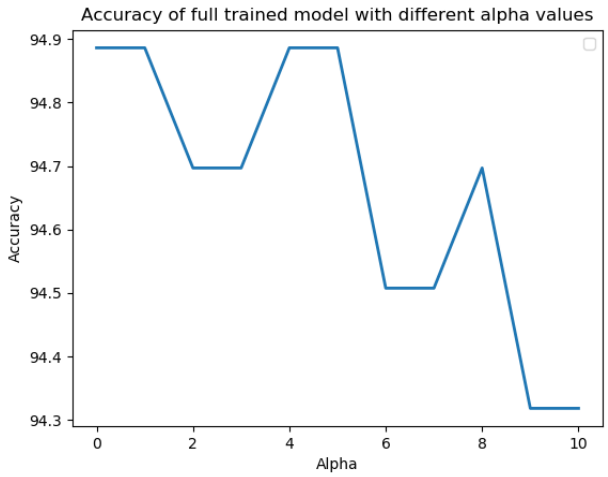
[[5 6]
[40 41]
[42 43]
[60 61]
[79 80]
[131 132]
[168 169]
[183 184]
[196 197]
[215 216]
[295 296]
[315 316]
[320 321]
[341 342]
[342 343]
[362 363]
[366 367]
[376 377]

[447 448]
[465 466]
[467 468]
[482 483]]

Q3.3) 8-mer for which your model assigns it to class 1 with highest probability:
['S' 'A' 'V' 'L' 'L' 'E' 'A' 'T'] This means that in the class 1, this 8-mer has the highest probability to break in the middle.
8-mer for which your model assigns it to class 0 with lowest probability:
['W' 'W' 'W' 'W' 'W' 'W' 'W' 'W'] This means that in the class 0, this 8-mer has the lowest probability to not break in the middle.
['L' 'E' 'A' 'A' 'A' 'A' 'A' 'K'] on the other hand has the highest probability of being in class 0 but the lowest probability to break overall.

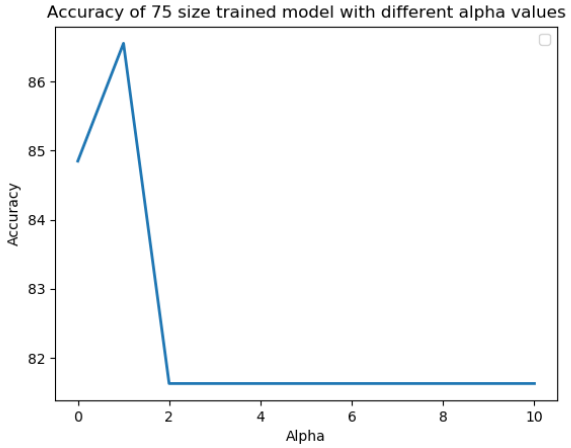
Q3.4) Smoothing with different alpha values:
Smoothing of different alphas with for full dataset:

Alpha	Accuracy %
0	94.88636363636364
1	94.88636363636364
2	94.6969696969697
3	94.6969696969697
4	94.88636363636364
5	94.88636363636364
6	94.50757575757575
7	94.50757575757575
8	94.6969696969697
9	94.31818181818183
10	94.31818181818183



Technically smoothing is required to remove the possibility of having zero probability for any class in the data set. Since the data set has around 5000 - 6000 samples, it has a very low chance of having a class with zero probability. That is why, the overall trend of accuracy while increasing alpha is getting low and the highest probability is at alpha 0 as we got in 3.1. The decreasing accuracy is due to high values of alpha creating noise in the samples.
Smoothing of different alphas with for 75 samples from dataset:

Alpha	Accuracy %
0	84.84848484848484
1	86.5530303030303
2	81.62878787878788
3	81.62878787878788
4	81.62878787878788
5	81.62878787878788
6	81.62878787878788
7	81.62878787878788
8	81.62878787878788
9	81.62878787878788
10	81.62878787878788



Technically smoothing is required to remove the possibility of having zero probability for any class in the data set. Since the data set has around 75 samples, it has a very high chance of having a class with zero probability. That is why, the accuracy at alpha at 0 is less than the accuracy of alpha at 1. After alpha 1, the accuracies decrease a lot due to the noise in the samples created by the high values of alpha.

Q3.5) Mutual Information

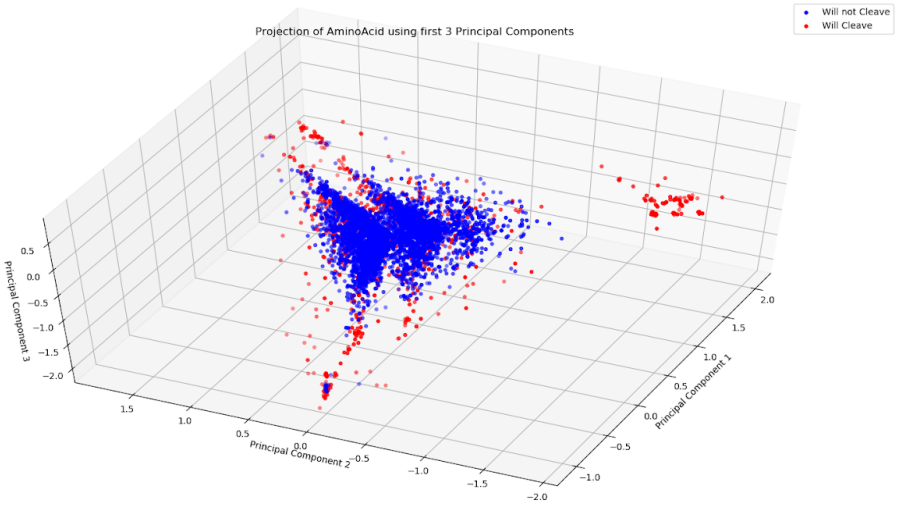
Highest Accuracy: 95.265 % at K=76(using 76 MI features). Yes this is greater than the 94.88 % accuracy that we got in 3.1.

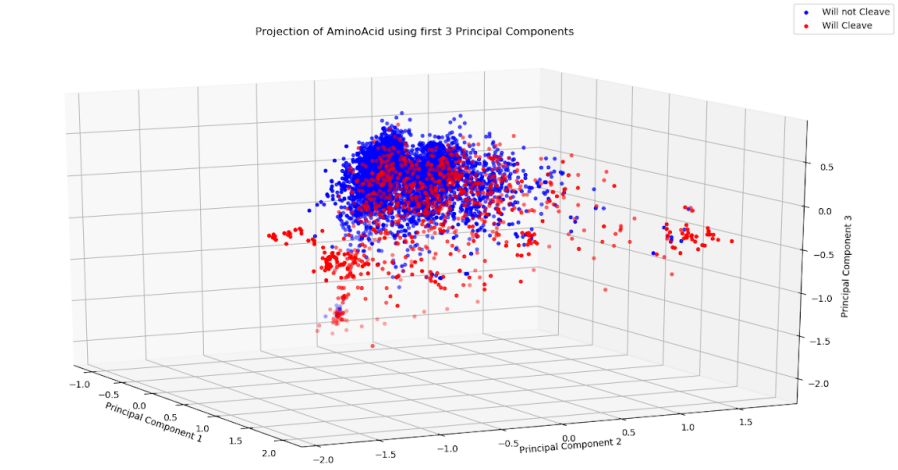
Q3.6) PCA

The features reduced to 3 dimensions from 160 can now be viewed as shown on the next page:

Proportion of Variance PVE for the 3 Principal Components are: 2.52%, 2.30% and 2.14% respectively.

Since the Cumulative information retained by the 3 Principal components is 6.98% and the accuracy given by training with this data is: 81.63 % we can see that the loss of accuracy from part 3.1 is 13.25 %. Since the time taken to train the model with 160 features and time taken to train with 3 features is not really very far apart, it is not feasible to lose 13.25 % of the accuracy of our model by using PCA k=3.


Projection of AminoAcid using first 3 Principal Components
Will not Cleave
Will Cleave
Principal Component 1
Principal Component 2
Principal Component 3


Projection of AminoAcid using first 3 Principal Components
Will not Cleave
Will Cleave
Principal Component 1
Principal Component 2
Principal Component 3