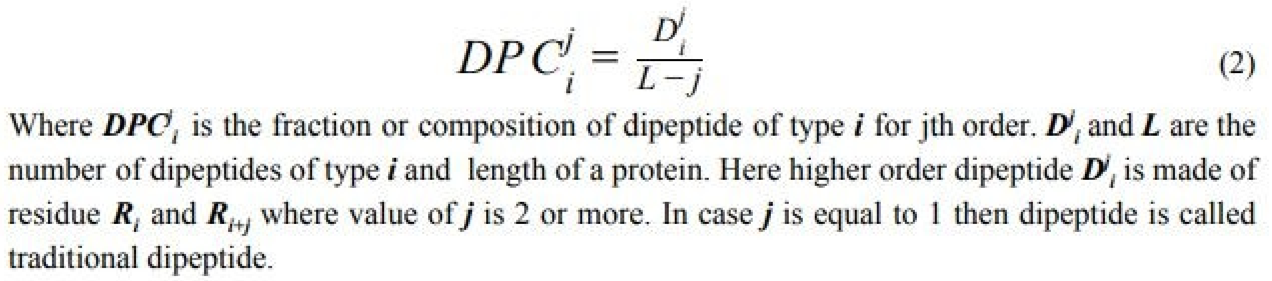
BDMH Assignment 2

Preprocessing: From Training and Test dataset take all the sequence and use p-feature extraction and extract all the features as
1. Di-peptide
2. Tri-peptide
3. Amino acid composition

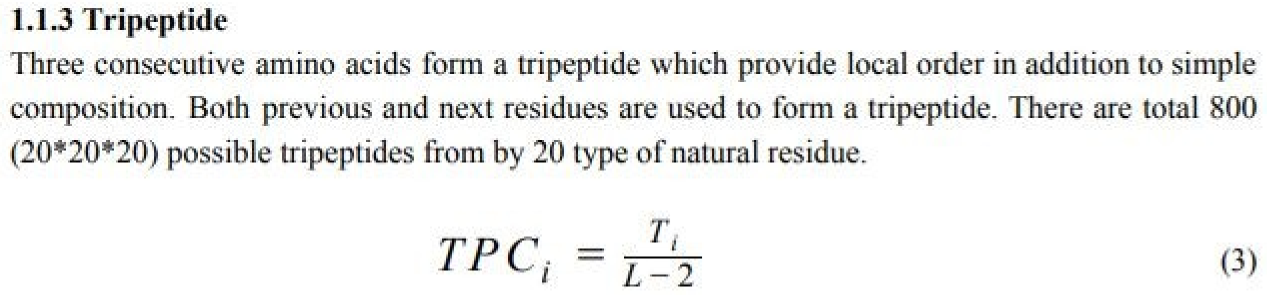
1. Dipeptide sequence:- By using following formula i extract percentage of all the possible 400 Dipeptides formed by 20 amino acids.



$$DPC_i^j = \frac{D_i^j}{L-j} \qquad (2)$$

Where $\boldsymbol{DPC^j_i}$ is the fraction or composition of dipeptide of type $\boldsymbol{i}$ for jth order. $\boldsymbol{D^j_i}$ and $\boldsymbol{L}$ are the number of dipeptides of type $\boldsymbol{i}$ and length of a protein. Here higher order dipeptide $\boldsymbol{D^j_i}$ is made of residue $\boldsymbol{R_i}$ and $\boldsymbol{R_{i+j}}$ where value of $\boldsymbol{j}$ is 2 or more. In case $\boldsymbol{j}$ is equal to 1 then dipeptide is called traditional dipeptide.

2. Tripeptide sequence:- By using following formula i extract percentage of all the possible 8000 Tripeptides formed by 20 amino acids.



1.1.3 Tripeptide

Three consecutive amino acids form a tripeptide which provide local order in addition to simple composition. Both previous and next residues are used to form a tripeptide. There are total 800 (20*20*20) possible tripeptides from by 20 type of natural residue.

$$TPC_i = \frac{T_i}{L-2} \qquad (3)$$

3. Amino acid composition: In this i computed all the amino acid composition as the percentage of a perticular amino acid.

Now I combined all the features and it becomes 8420 features.

Now i remove all the features which has all the values zero in the entire dataset. Because feature which has zero values does not make any effect on the prediction.

Now i used Select K Best features and select 3000 best features. Now I have a dataset with 3000 features.

Methodology:-

Notebook 1:- I used Keras sequential model with two dense layers with 512 neuron with relu activation function and optimizer='adam', init='glorot_uniform' on Dipeptide and Amino acid composition with 420 features. I trained this model with 12 epochs and i got 88.169% accuracy on public leaderboard.

Notebook 2:- In this i also used Keras sequential model with 2 layers with 2048 neurons with relu activation function and optimizer='adam', init='glorot_uniform' on Dipeptide and Tripeptide and amino acid compositions with 8420 features then removal of all the features with 0 values now using select K Best features i choose 3000 best features. I got 88.549% accuracy by using this configuration on the public leaderboard.

Analysis:- I used different models to train with different variations of models and different number of neurons and no of layers and different optimizers and init where i got different results and efficiency of the model. According to their validation performance i choose my submissions. Because of randomly initializing the parameters, output is slightly changing .