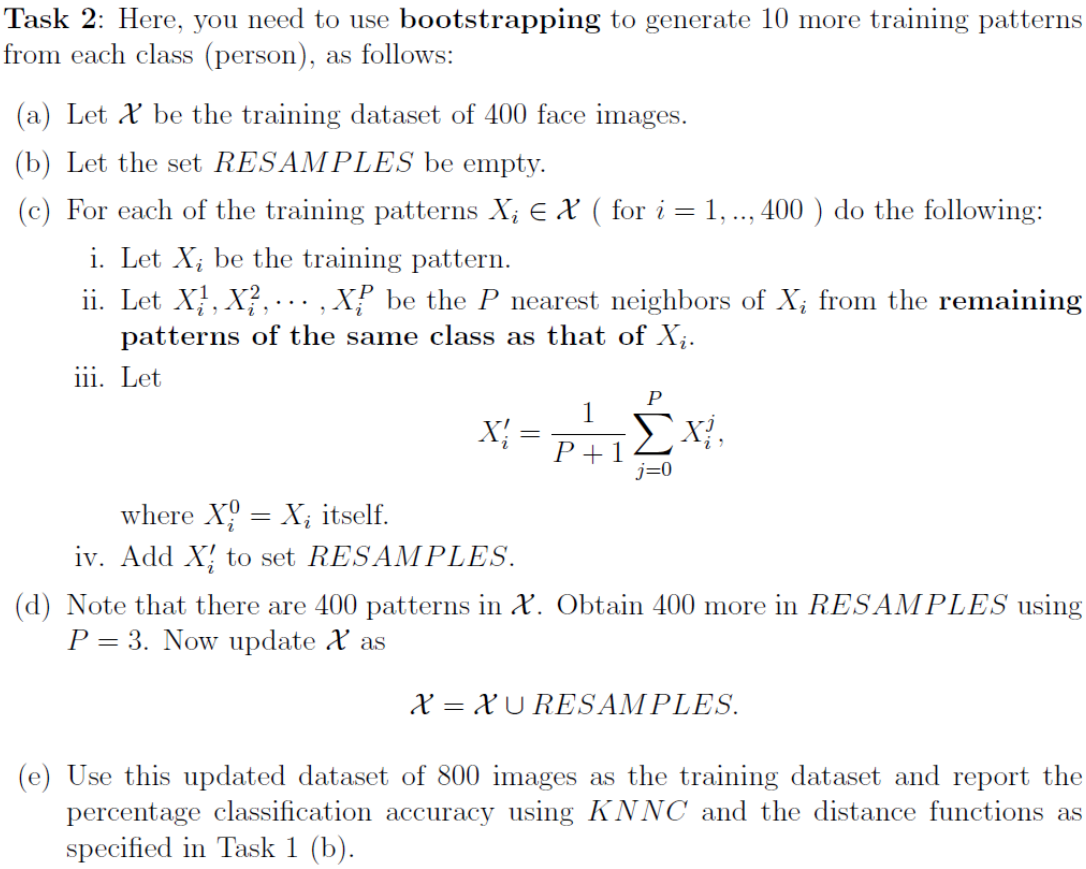
Machine Learning

KNNC Task – 2



Task 2: Here, you need to use **bootstrapping** to generate 10 more training patterns from each class (person), as follows:

(a) Let $\mathcal{X}$ be the training dataset of 400 face images.

(b) Let the set $RESAMPLES$ be empty.

(c) For each of the training patterns $X_i \in \mathcal{X}$ (for $i = 1, .., 400$) do the following:

i. Let X_i be the training pattern.

ii. Let $X_i^1, X_i^2, \cdots, X_i^P$ be the P nearest neighbors of X_i from the **remaining patterns of the same class as that of** X_i.

iii. Let

$$X_i' = \frac{1}{P+1}\sum_{j=0}^{P} X_i^j,$$

where $X_i^0 = X_i$ itself.

iv. Add X_i' to set $RESAMPLES$.

(d) Note that there are 400 patterns in $\mathcal{X}$. Obtain 400 more in $RESAMPLES$ using $P = 3$. Now update $\mathcal{X}$ as

$$\mathcal{X} = \mathcal{X} \cup RESAMPLES.$$

(e) Use this updated dataset of 800 images as the training dataset and report the percentage classification accuracy using $KNNC$ and the distance functions as specified in Task 1 (b).

(a) SOLUTION

CODE:

Please find the code attached for BOOTSTRAPPING as *KNNC_OlivettiFaceData_Bootstrapping_impl.py*

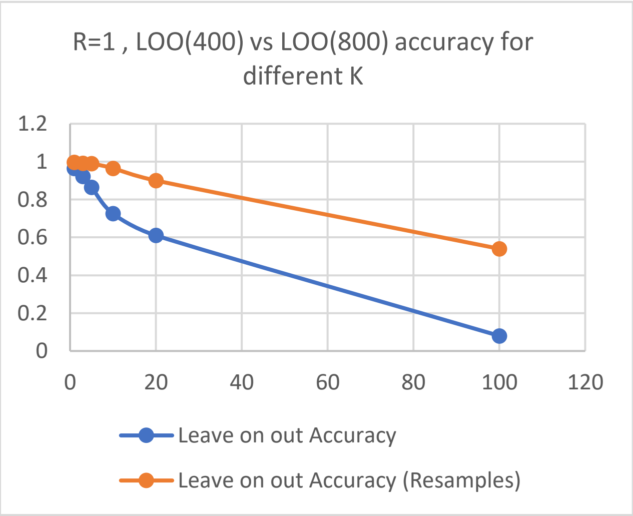
- **BOOTSTRAPPING** algorithm is implemented as per the description above to increase the number of samples per person.
- Using Resampled set of 800 images hence, KNNC for each value of r and K is determined.
- **RESULT captures the LOO accuracy after resampling. In addition a tabular data is illustrated below to show and compare LOO accuracy with 400 samples and LOO accuracy with 800 samples after resampling is done using BOOTSTRAPPING.**

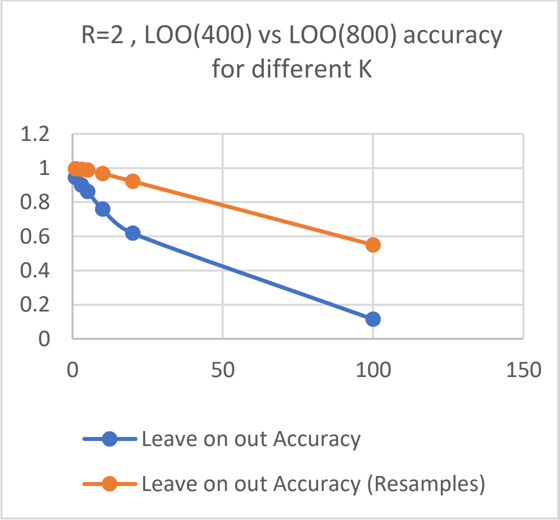
RESULT:

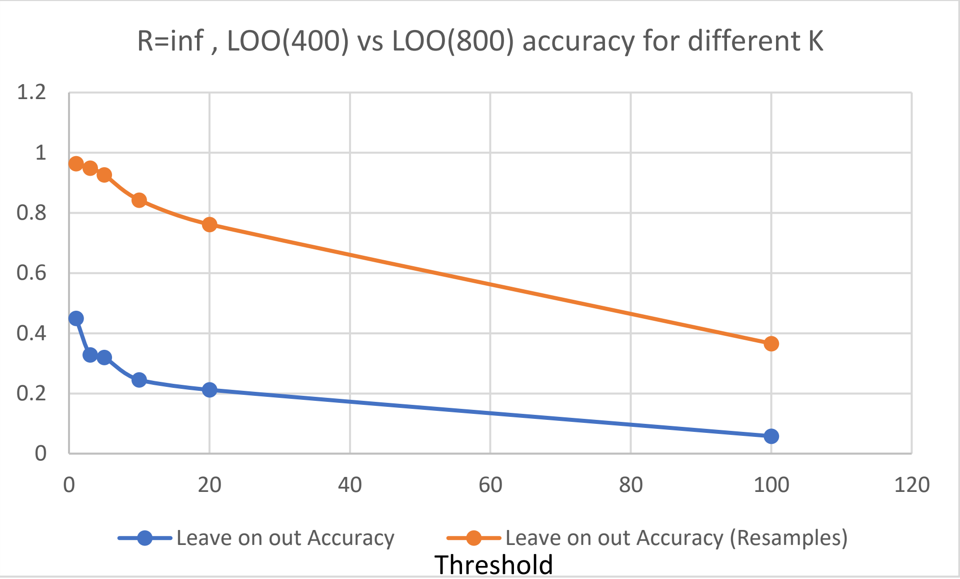
Pls find below a tabulation of results from TASK1 (400 samples) and TASK2 (800 samples) . The result shows LOO accuracy for R=1,2,inf and K=1,3,5,10,20,100.

r - (exp value in Minkowski distance)	K	Leave on out Accuracy(n= 400 samples) – TASK 1 result	Leave on out Accuracy (Resamples, n=800 samples)
1	1	0.965	0.996
	3	0.922	0.991
	5	0.865	0.99
	10	0.725	0.965
	20	0.61	0.9
	100	0.08	0.54
2	1	0.945	0.995
	3	0.9	0.991
	5	0.863	0.988
	10	0.76	0.968
	20	0.618	0.922
	100	0.115	0.55
Infinity	1	0.45	0.964
	3	0.328	0.949
	5	0.32	0.927
	10	0.245	0.843
	20	0.212	0.762
	100	0.058	0.366

PLOT:







INFERENCE/ANALYSIS:

- The **bootstrapping** process of resampling increases the number of samples per person (Class) to 20. This is achieved by adding the mean value of 3 nearest neighbours and itself for each sample belonging to the same class.
- **The KNNC model applied over the entire resampled data (with 800 samples) shows that the accuracy and loo accuracy is much higher when compared to original data (of 400 samples) for all values of K and R.**
- **With resampled data, we can see that even for K=20 and lesser, the accuracy is very high compared to original data where accuracy was high only to K<10. This also clearly shows that when 20 samples are added per person, K value upto 20 still shows higher accuracy.**
- **Therefore in general it shows that when the sample data set is increased with resamples using bootstrapping method, the accuracy of the model also increases.**

RESOURCES USED FOR THE ASSIGNMENT:

- **Environment:**
  Anaconda, Jupyter notebook
- **Software :**
  Python
  Python libraries/modules: Pandas, Numpy, SkLearn etc