Report :: TIPR Assignment - I

NIDHI KUMARI (15127)

11-02-2019

1 INDRODUCTION

Python version -3
In this assignment we have implemented Random projection,naive bayes, K nearest neighbor, Locality sensitive Hashing. For each of the following i have used 10 folds cross validations. Dataset to work upon are

- Dolphins
- Pubmed
- Twitter

2 Task1

we have to implement Random projection. Random projection is a method to reduce the dimensionality of the data. For implementation i have created a random matrix with zero mean and unit variance and multiply the random matrix with the original data matrix. The output matrix which we will get is the reduced dimension matrix

3 Task2

Here we have to implement K Nearest Neighbor and Naive bayes. I have implemented Gaussian naive bayes for dolphins and pubmed whereas multinomial naive bayes for twitter dataset.

4 Task3

plotting of accuracy and f1-score vs dimensions for various dataset using implemented methods.

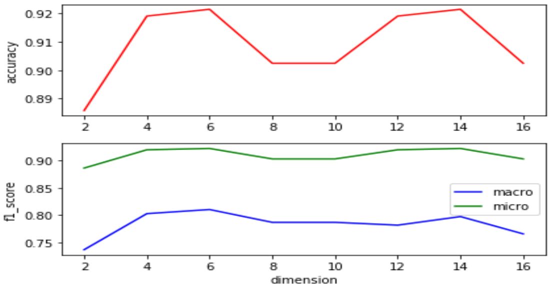
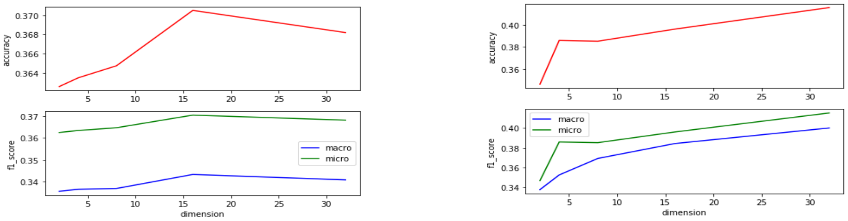


Figure 1: KNN for dolphins

since dolphins dataset is small hence the accuracy fluctuates a lot. so we can't predict any relation between dimension and accuracy

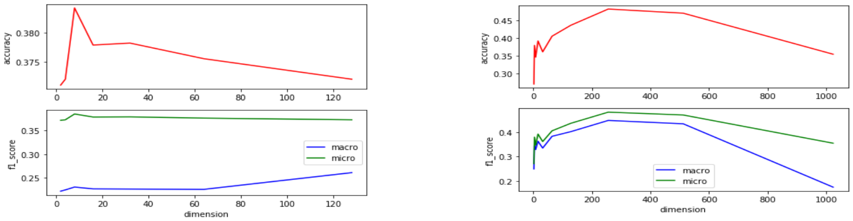


(a) KNN

(b) Naive Bayes

Figure 2: Pubmed dataset

for pubmed dataset KNN is performing good for lower dimension but not good for higher dimension. Whereas Naive bayes is performing good for higher dimension and overall Naive bayes is performing better than KNN hence for Naive Bayes i an on Amar side.



(a) KNN

(b) Naive Bayes

Figure 3: Twitter dataset

For Twitter dataset also Naive bayes is performing good as compared to KNN and KNN is performing good for low dimension and Naive bayes is performing good also on not so high dimension. so Akbar side is safe for this case.

5 Task4

plotting of accuracy and f1-score vs dimensions for various dataset using library methods.

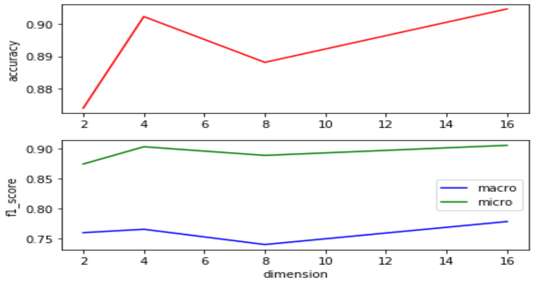
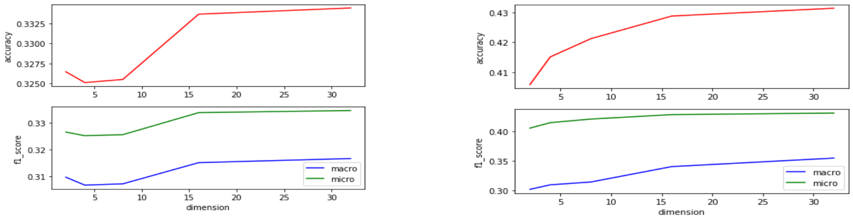


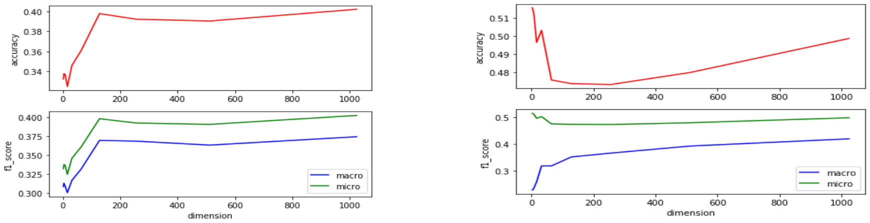
Figure 4: KNN (Library) for dolphins
inbuilt library KNN is performing good on high dimension.



(a) KNN

(b) Naive Bayes

Figure 5: Pubmed dataset
for pubmed dataset for KNN accuracy is first decreasing then it is increasing. Whereas in case of Naive bayes acuuracy keeps increasing with dimensions.



(a) KNN

(b) Naive Bayes

Figure 6: Twitter dataset
For Twitter dataset also Naive bayes is performing good as compared to KNN. it is giving aprrox .50 accuracy on lower dimension

6 Task5

In task 5 we have to compare our implemented method with the inbuilt library method

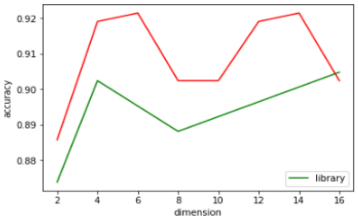
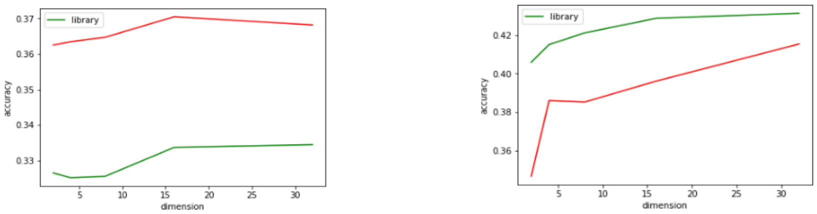


Figure 7: KNN for dolphins

for dolphin data library method is performing a little bit bad as compared to implemented method.

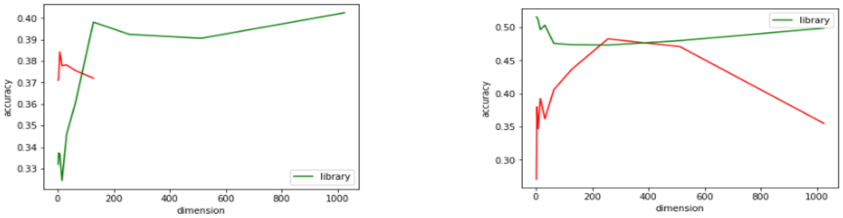


(a) KNN

(b) Naive Bayes

Figure 8: Pubmed dataset

inbuilt library for Naive bayes is performing better which is not the case with KNN for pubmed.



(a) KNN

(b) Naive Bayes

Figure 9: Twitter dataset

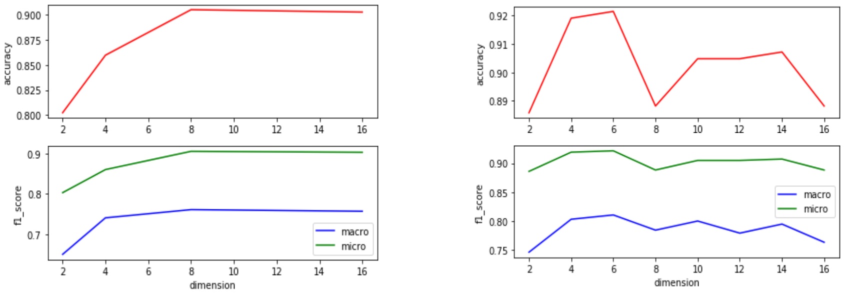
In both cases library is performing good for twitter

7 Task6

I have used inbuilt Scikit library LSHForest to implement Locality sensitive Hasing.

8 Task7

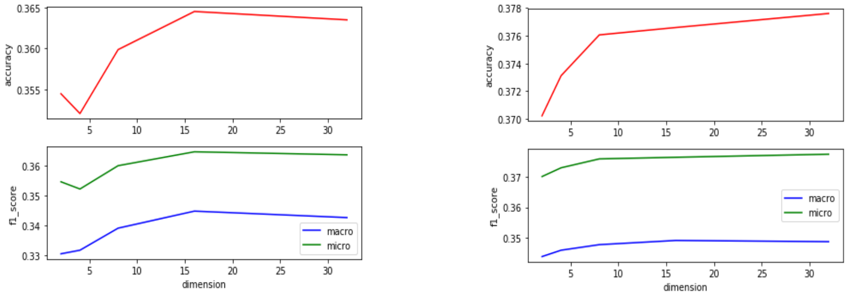
here we have to compare LSH with PCA. for PCA i have used KNN as a classifier after reducing the dimensions.



(a) LSH

(b) PCA

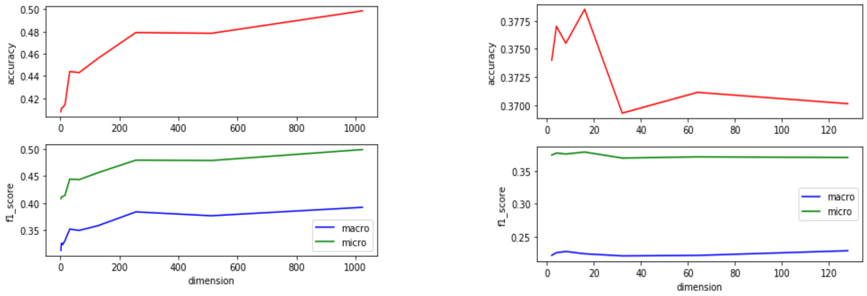
Figure 10: dolphins dataset
LSH accuracy is increasing as dimension is increasing but decreasing in case of PCA.



(a) LSH

(b) PCA

Figure 11: Pubmed dataset
accuracy is increasing in both the cases.



(a) LSH

(b) PCA

Figure 12: Twitter dataset

accuracy is increasing in the case of LSH whereas decreases in the case of PCA.

9 Github link

`https://github.com/Nidhi-kumari/tipr-first-assignment`