CSE512 Spring 2021 - Machine Learning - Homework 4

Your Name: Manish Reddy Vadala

Solar ID: 114190006

NetID email address: manishreddy.vadala@stonybrook.edu

Names of people whom you discussed the homework with: NA

1)

1.1) S.T LOOCV (Leave one out cross validation) error is bounded by $\frac{m}{n}$

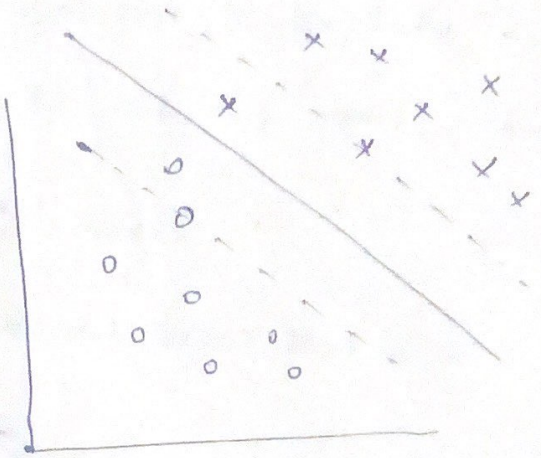
n - Data points

m - Support vectors

$$\text{Error} = \frac{\text{misclassifications among all folds}}{\text{Total no. of folds}}$$

↳ n

(∵ LOOCV is K-fold (K=N))



Linear Space

case 1: Picking non-Support Vector as a test set

In this case classifier works the same since all the support vectors are still intact, so margin width & the seperation line doesn't change.

So the test point (non-support vector) will be classified correctly. So There is zero error in this case.

case2: Picking support vector as a test set

In this case, seperating line & width might change. In linearly seperable, support vectors are placed on the correct side of support line but wrong side of the margin. After taking a point out for testing which is a support vector, if it close to the seperating line removing might move the line close towards it's class.

and classification of the test point gives wrong class.

And if all the test points end up as a support vectors covering all. Total no. of misclassifications can go up to m.

and error can be at max $\frac{m}{n}$ +

$\therefore$ error $< \frac{m}{n}$

So in both cases LOOCV error is always bounded by $\frac{m}{n}$

(1.2) General Kernel instead of linear kernel. Given the data is linearly seperable in high dimensional space w.r.t Kernel. Main change here would be hyper plane instead of seperable line which divided the n points. But everything is same here. working of LOOCV error when a support vector or non-support vector is taken as a test set.

so Even here:

case 1: Picking non-support vector as a test set.

This is same as for the linear kernel. Hyper plane doesn't change since all support vectors are in tact.

$\therefore$ The error would be nothing.

Case2 :- Picking Support vector as a test set.

This is same as for linear kernel case:

So Even if here if the hyperplane changes the maximum number of misclassifications is m.

So the total Error here at max would be $\frac{m}{n}$

So finally LOOCV Error is still bounded by $\frac{m}{n}$

▼ Q2

▼ 2.1

- Reporting the accuracy and the confusion matrix on both the train and test sets.
- for train set

  accuracy - 0.8665274408034151
  confusion matrix - [[23560 1160]
  [3186 4655]]
- for test set
  accuracy - 0.8661630120999939
  confusion matrix - [[11857 578]
  [1601 2245]]

▼ 2.2

- parameters that i'm tuning here are max_depth & min_child_weight
- Method: I have a defined a range for both max_depth and min_child_weight. for every possible combination of a pair i'm running the k-fold cross validation and getting the mean value of the accuracy score. and saving the params of the best mean accuracy score in all possible pairs.
- the best params that i have got here are max_depth = 9 & min_child_weight = 7.
- best accuracy in k-fold cross validation is 0.8723628242779302
- Using the best parameters we got and after training the xgb model, and running it on test data,
  - the accuracy we got is

    0.872366562250476
  - confusion matrix

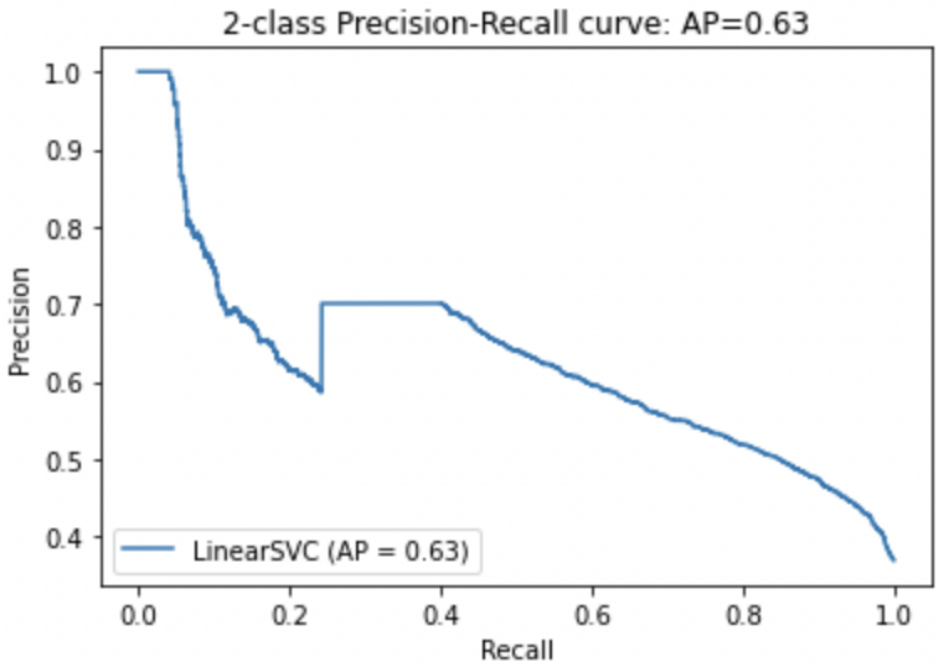
    [[11701 734]
    [1344 2502]]

▼ Q3

▼ 3.3.1

- Ap & precision recall curve on the validation data

  mAP: 0.00010264436423312873



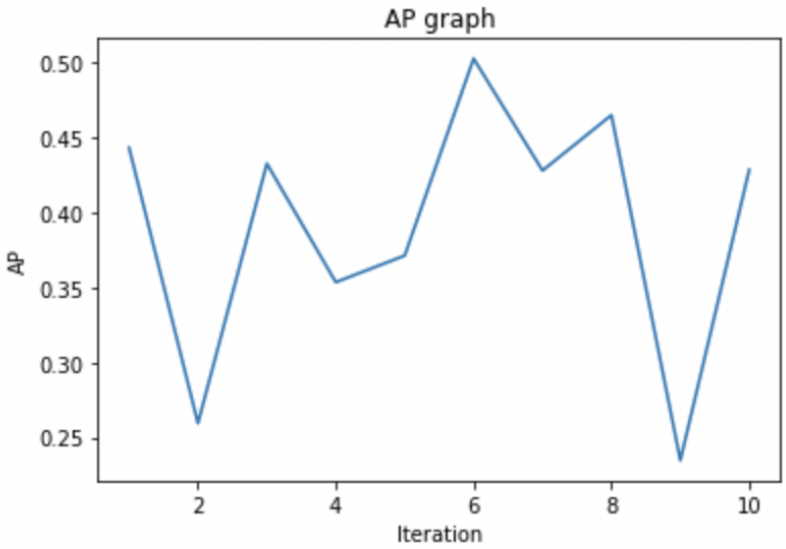
▼ 3.3.3

- Objection function i have used here is:

$$\frac{1}{2}||\mathbf{w}||^2 + C\sum_{j=1}^{n}\xi^j$$

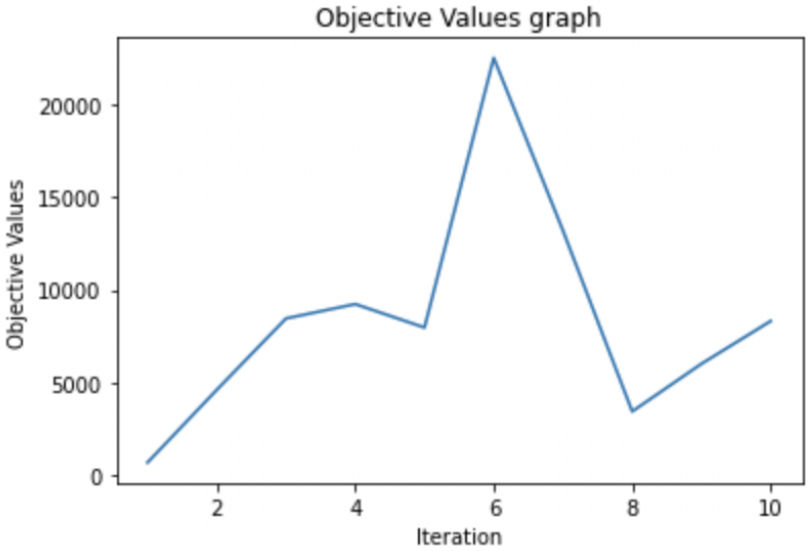
- plot of iterations vs AP
  - Accuracy precision on validation test for each iteration

    [0.44317889246320213, 0.2593701000220637, 0.4324943299997145, 0.35345005534834684, 0.37117352793954395, 0.502619001606597, 0.4278993168435851, 0.46487342200049087, 0.23461471974349477, 0.4284069876054457]



- plot of iterations vs objective values
  - Objective values for test set calculated for each iteration

    [660.5435950958606, 4591.654655187884, 8448.298555437541, 9223.674550508878, 7956.900511586875, 22536.495677293075, 13199.026460872585, 3420.743846330359, 5996.228710316962, 8307.526648469491]



▼ 3.3.4

- AP i got for the test dataset is 0.00013355551