homework5

December 13, 2022

1 Homework 5

Subject sentence

1. Finish the algorithm from Lab 2. The code provided in class that is under Resources > Lectures presents the estimate after a fixed number of iterations. Instead, have it return the estimate that has the smallest SSE.

Lab 2 was about Boosting. Here is a boosting method. It scored 95.5% on a random dataset.

[5]:
```
from sklearn.ensemble import GradientBoostingClassifier
from sklearn.model_selection import train_test_split
from sklearn.datasets import make_classification

X, y = make_classification(n_samples = 1000, n_features = 10,n_informative = 2,␣
 ↪n_redundant = 0,random_state = 0, shuffle = False)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2,␣
 ↪random_state =1234)

clf = GradientBoostingClassifier(n_estimators=100, learning_rate=1.
 ↪0,max_depth=1, random_state=0).fit(X_train, y_train)
clf.score(X_test, y_test)
```

[5]:
```
0.955
```

2. Why is the smallest SSE not attained by the last iteration? It certainly is created with the most knots and thus in some sense the most flexible. Why does the algorithm not converge?

I am not sure what "last iteration" this question is asking about; I think it is related to the question 1 and this line in lab 2: "Pick the estimator that has the lowest sum of squares (a terrible idea if we thought that this might eventually fit noise)."

3. Why are we able to simply minimize SSE (not regularize, or need to hide data from the learner in a CV routine)?

We are able to simply minimize SSE because least squares has lots of nice properties and it is useful for estimating means at the undergraduate level. Boosting with weak learners also prevents overfitting.

4. Why does the smoothness vary across the unit interval when we only selected a single smoothing parameter (who's value, if we were doing traditional kernel smoothing, would seemingly

play connect-the-dots)?

The larger the smooth width, the greater the noise reduction, but also the greater the possibility that the signal will be distorted by the smoothing operation.

5. Run the learner on another simulated data set. Again, let the input be a grid (you'd need to change the code a lot to run this on unequally spaced data). What is the critical parameter that needs to be choosen for the method to work well? Can you interpret it?

The critical parameter that needs to be choosen for the method to work well is the minimum spacing between knots. No, I cannot interpret it.