2019150445 신백록

1.house 데이터에 적용한 linear regression, support vector machine(선형), 그리고 RANSAC에 의해 추정된 선형모형을 제시하고, RANSAC(그러므로 outlier가 제거된 OLS 선형모형)과의 비교를 통해 linear regression 모형과 support vector machine을 설명하라.

Estimated model for Linear Regression:

65.6978-0.137\*CRIM+0.031\*ZN+2.524\*CHAS-17.030\*NOX+1.240\*RM+0.031\*AGE-1.288\*DIS+0.262\*RAD-0.007\*TAX-0.828\*PTRATIO+0.005\*B-9.972\*LLSTAT-0.605\*LINDUS

Estimated model for RANSAC:

31.4280-0.340\*CRIM+0.026\*ZN+1.266\*CHAS-6.777\*NOX+3.205\*RM-0.037\*AGE-1.134\*DIS+0.243\*RAD-0.007\*TAX +0.600\*PTRATIO+0.012\*B-3.856\*LLSTAT-1.012\*LINDUS

Estimated model for Support Vector Linear Regression:

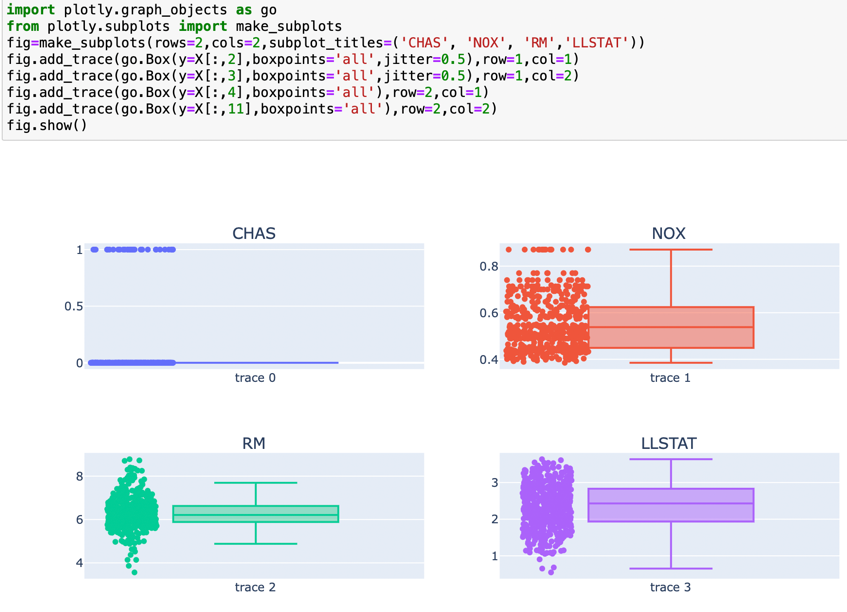
-0.105\*CRIM+0.023\*ZN+.663\*CHAS-2.290\*NOX+3.549\*RM-0.012\*AGE-0.760\*DIS+0.176\*RAD-0.010\*TAX+0.641\*PTRATIO+0.012\*B-5.780\*LLSTAT-0.607\*LINDUS

RANSAC의 Coef들과 Linear Regression, Support Vector Linear Regression의 Coef들의 차이를 계산하여 절댓값이 1이 넘는 column들만 추출하였다.

텍스트이(가) 표시된 사진

자동 생성된 설명

위에서 True가 나온 column들의 boxplot을 그려서 outlier들이 있는지 확인해보았다. 모든 column에서 outlier들은 존재하였지만 test mse를 비교해봤을 때 Support Vector Linear Regression의 test mse가 16으로 가장 작기 때문에 RANSAC보다는 SVL을 선택할 것이다. 아마 SVL에서 slack variable을 조정해주는 C 때문에 outlier가 얼추 제거된 것 같다.



2. house 데이터에 적용된 SVM에서 epsilon의 역할을 support vector의 관점에서 논의하라.

텍스트이(가) 표시된 사진

자동 생성된 설명

Epsilon이 0, 0.01, 0.1, 1, 5, 10, 100일 때 svl의 train mse와 test mse를 계산하였다. Epsilon이 작으면 support vector가 많아져 일반적으로 overfitting이 발생하는데 house 데이터를 보면 epsilon이 0일 때부터 점점 커지면 test mse도 점점 커지는 것을 볼 수 있다. Train mse는 5일 때가 가장 작고, 일반적으로는 epsilon과 비례한다. 당연한 것이 epsilon이 커지면 추정에 사용되는 support vector가 점점 줄어들기에 train mse는 커지고, 과소 적합이 일어난다. House 데이터에서는 epsilon이 0일 때 train mse, test mse 모두가 가장 작고 과대 적합도 일어나지 않았기에 epsilon이 0인 모델이 가장 좋아 보인다.