2019150445 신백록

1. 선형 SVM에서 c값의 변화에 따른 accuracy의 변화와 과대적합에 미치는 영향을 논의하라.

For 문을 통하여 C가 0.01, 0.1, 1, 5, 10, 100인 경우에 대하여 model을 적합 시켰고, 각각의 train acc와 test acc를 출력하였다.

텍스트이(가) 표시된 사진

자동 생성된 설명

결과를 보니, C가 적을 때는 train acc와 test acc가 모두 낮게 나오는 것을 확인할 수 있다. Overfitting이 일어나지는 않지만, 애초에 C를 작게 준 탓에 오차를 많이 허용해 모델의 성능이 좋지 않게 나온 것이다. C가 1정도에서 가장 좋은 성능의 모델이 만들어지는 것을 볼 수 있다. C가 5일 때부터는 Overfitting이 발생하기 시작한다. C가 커지면 오차를 적게 허용하고, bandwidth가 작아져 train set에 모델이 과적합되는 것이고, 실제로도 그런 것을 accuracy를 통해 확인할 수 있다.

텍스트이(가) 표시된 사진

자동 생성된 설명 텍스트이(가) 표시된 사진

자동 생성된 설명

마지막으로, gridsearchcv를 통해 C를 조정해가며 가장 성능이 나은 C를 찾아보았고, 대략 C=0.16정도일 때 SVC가 가장 좋은 성능을 보였다.

2. 교재의 비선형 SVM 모형에서 kernel=’rbf’, ‘poly’, ‘sigmoid’으로 변경하고 c와 gamma를 변경하여 accuracy와 과대적합에 미치는 영향을 논의하라

책에서 Kernel=’rbf’, C=1.0, gamma=0.2 였을 때, train acc는 0.99인 반면, test acc는 0.95정도로 overfitting을 해결할 필요가 있었다.

Overfitting을 해결하기 위하여, 커널의 분산을 조절하는 감마를 줄이면 비선형성이 줄어들어 좀 더 General한 모델을 만들 수 있다. 따라서 감마를 줄여보면, 다음과 같이 test acc는 감소하거나 유지되고, train acc 마저도 줄어드는 것을 볼 수 있다.

텍스트이(가) 표시된 사진

자동 생성된 설명텍스트이(가) 표시된 사진

자동 생성된 설명

따라서 감마를 줄이며 C를 늘려 모델의 bias를 막으며 Overfitting까지 해결을 해보았다. Train acc는 99%, test acc는 대략 98%로 좋은 성능을 보인다.

텍스트이(가) 표시된 사진

자동 생성된 설명

또한 감마를 늘리게 되면 kernel의 분산이 작아져 조금 더 가파른 정규 분포가 만들어지기에 비선형성이 강해진다. 따라서 다음과 같이 overfitting이 심하게 발생하게 된다.

텍스트이(가) 표시된 사진

자동 생성된 설명

Sigmoid와 Ploy로도 모델 적합을 해보았는데, 다음과 같이 kernel=sigmoid를 사용하면, gamma를 매우 작게 주어야 그나마 작동이 가능한 수준이 되었다.

텍스트이(가) 표시된 사진

자동 생성된 설명

Kernel=’poly’를 사용할 때, gamma와 C를 잘 조절해가며 주며 bias와 overfitting을 해결해보았더니 test acc가 1이 나왔다. 물론 이것 또한 CV를 통해 kernel, C, gamma를 조절해가며 확인하면 좋은 parameter들을 추출해낼 수 있을 것이다.

텍스트이(가) 표시된 사진

자동 생성된 설명