**CS302 Introduction to machine learning**



School of Undergraduate Studies, DGIST

201611161 정원균

Homework report 2.

2022. 05. 08

Q1. For MNIST data set, train Logistic regression models and find the best model that can achieve the highest accuracy on the test data set.

텍스트이(가) 표시된 사진

자동 생성된 설명

Figure 1. Library import and Linear regression code

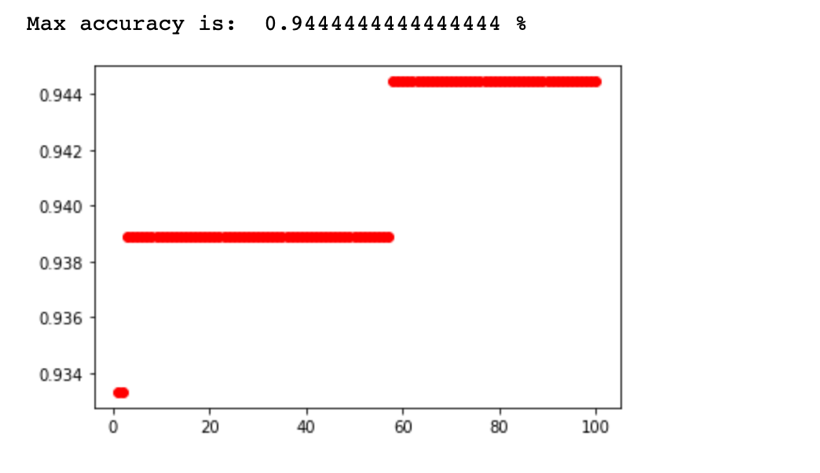


Figure 2. Accuracy of Linear Regression

Test에 사용할 Logistic Regression, KNN, SVM, Random forest 모듈은 sklearn library를 import 하여서 사용하였다. 또한 Mnist 데이터를 import하여서 train data와 test data를 만들었다.

다음 linear\_Reg( ) parameter에서 LogisticRegression(C, max\_iter)에서 규제강도를 의미하는 C 값을 0부터 100까지 나누어서 (x-label) Accuracy를 분석해보았다. 결과를 보면 C 값이 60이상일 때 94.4%정도의 정확도를 보임을 알 수 있다.

Q2. For the same data set, train K-NN classifiers and find the best model that can achieve the highest accuracy on the test data set

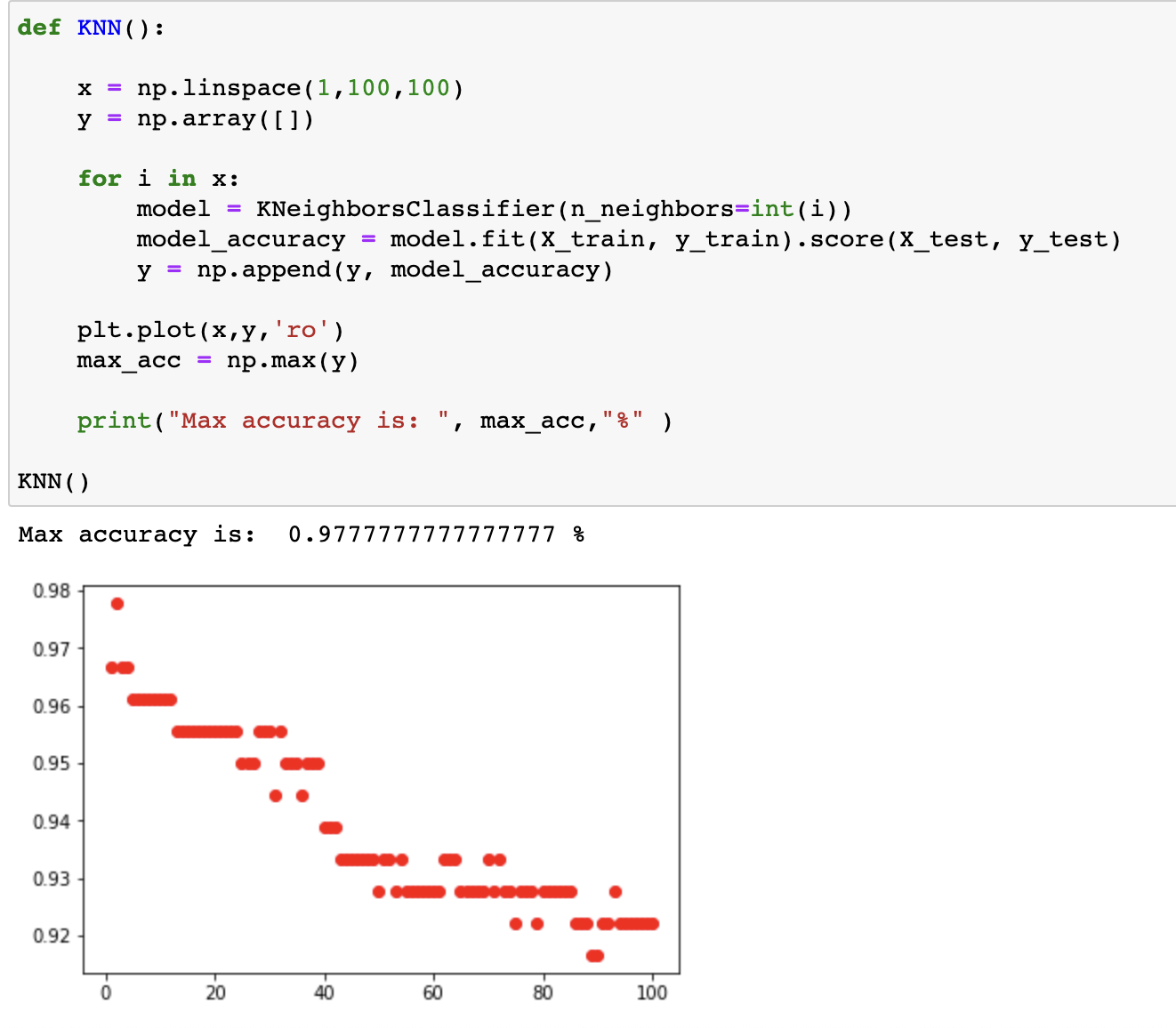


Figure 3. Code and Result of KNN

다음은 같은 Linear Regression에 사용하였던 같은 train, test data를 이용해서 KNN모델에 적용한 코드와 결과이다. 앞에서 규제강도에 따라 Accuracy를 확인했던 것과 달리 KNN model에는 몇 개의 neighbor을 비교할 것인지에 따라서 Accuracy를 분석해보았다. 그 결과 K = 2일때 가장 높은 정확도를 보였으며, K=2가 아닌 case에 대해서 비교할 Neighbor의 수가 증가할 수록 오히려 정확도가 떨어지는 모습을 볼 수 있다.

Q3. For the same data set, train SVM classifiers and find the best model that can achieve the highest accuracy on the test data set.

텍스트이(가) 표시된 사진

자동 생성된 설명

Figure 4. Code of SVM model

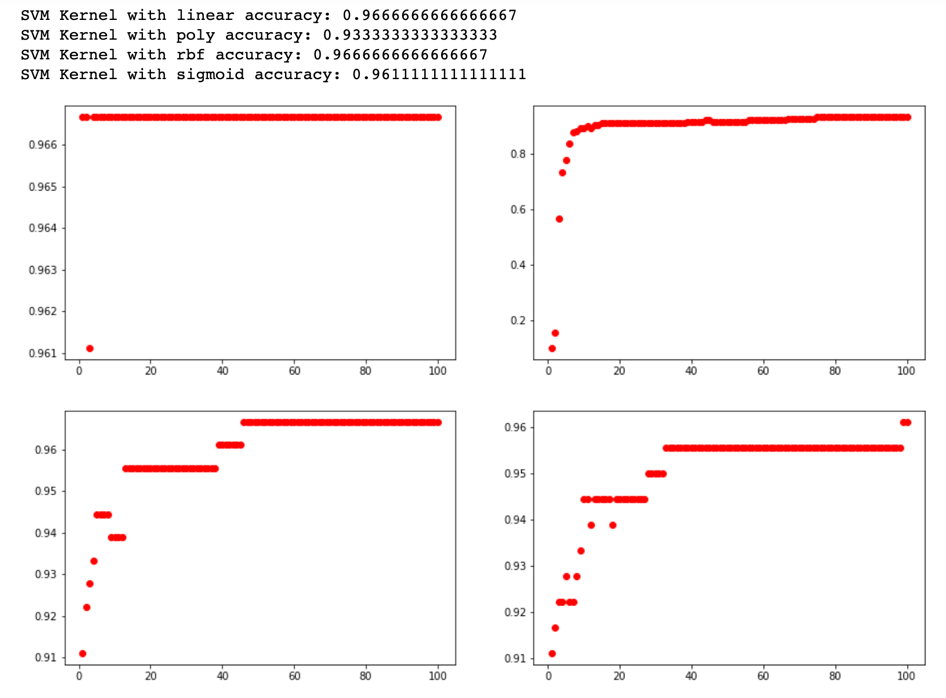


Figure 5. Result of SVM model

SVM에서는 총 두가지의 parameter의 변화를 주어서 Accuracy를 확인해보았다. Kernel은 총 4개 linear, polynomial, rbf, sigmoid의 결과를 확인하였고, 역시 각 kernel별로 규제정도 C를 1부터 100까지 바꾸어 가면서 각 kernel 별로 가장 높은 Accuracy를 출력하였다. 그 결과 linear model과 rbf model이 96.7% 정도로 가장 높음을 알 수 있고 C값은 50이상의 수를 parameter로 주면 문제없이 가장 높은 Accuracy가 나온다는 것을 알 수 있다.

Q4. For the same data set, train Random forest classifiers and find the best model that can achieve the highest accuracy on the test data set.

텍스트이(가) 표시된 사진

자동 생성된 설명

Figure 6. Code & Result of Random Forest model

다음은 max depth = 20, Estimator 수를 100 ~ 1000까지 준 Random forest model학습 결과이다. Estimator를 약 180개 이상 대입하였을 때 약 93.9% 의 정확도를 보임을 알 수 있다.

**Conclusion**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Linear Reg. | KNN | SVM | Random Forest |
| Accuracy | 94.4% | 97.8% | 96.7% | 93.3% |

총 4개 유형의 model을 바탕으로 각 model당 변화시킬 수 있는 parameter를 바꾸면서 최대 정확도를 구해보았을 때 KNN model에서 n of neighbor가 2일 때 가장 높은 정확도를 보여준다는 것을 알 수 있다.