Assignment 1

2015313254 노인호

September 29th 2019

1 Progress

1.1 Problem

붓꽃(Iris)의 품종을 분류하기 위한 문제이다. 붓꽃의 품종은 setosa, versicolor, virginica로 나눠지는데 어떤 품종인지 구분하기 위해 꽃받침(sepal)의 길이(length)와 너비(width), 그리고 꽃잎(petal)의 길이(length)와 너비(width)를 이용한다. 길이와 폭의 단위는 cm이다. 품종의 개수가 3개이고 각 품종에 대한 정보를 미리 알고 있으므로 multi-class classfication문제라는 것을 알 수 있고, supervised learning을 이용해서 학습을 진행한다.

1.2 Data Load

scikit-learn에 있는 iris데이터를 'load_iris()'를 이용하여 로드하고 데이터의 형태를 파악한다. Target이 3개, Feature가 4개, 데이터는 총 150개가 numpy.ndarray의 형태로 저장이 되어있음을 확인할 수 있다.

1.3 Data preprocessing

(1) iris data를 학습시키기 위한 입력변수와 종속변수는 다음과 같게 된다.

- input variables(x) : sepal length, sepal width, petal length, petal width
- output variables(y) : species (setosa, versicolor, virginica)

(2) scikit-learn의 train_test_split함수를 이용해서 먼저 train(train and valid)과 test set을 8:2로 나누고 validation을 위해 train(train and valid) set을 train과 valid set으로 8:2로 나눈다. random_state=0으로 두었다. 그래서 data set의 비율은 train : valid : test = 0.64 : 0.16 : 0.2 가 된다.

(3) data scaling은 데이터를 정규화 시켜주는 StandardScaler함수를 사용하였다.

1.4 Learning Algorithm(KNN)

Scikit-learn의 KNeighborsClassifier 함수를 사용하였다. main hyperparameter는 다음과 같다.

- n_neighbor (the number of neighbors k)
- metric (distance metric p : Euclidean(p=2), Manhattan(p=1). Minkowski(p=p) ...)
- weights (weighting scheme : uniform, distance(inverse of Euclidean, Manhattan. Minkowski ...))

1.5 Hyperparameter

n_neighbor 값을 1-30개 까지 바꾸고, 5가지의 metric과 2개의 weights를 변경시켜가며 valid accuracy값이 제일 잘 나오는 Hyperparameter값을 찾았다. valid accuracy가 가장 높게 나오는 sweetspot은 n_neighbors가 22였고 metric='minkowski' p가 3, weights='uniform', valid accuracy=1 였다.

2 Conclusion

1.6에서 나온 hyperparameter로 계산해본 test accuracy=0.9가 나왔다.

3 Python code in jupyter notebook

웹사이트 https://nbviewer.jupyter.org/ 에서 다음의 gist 주소를 입력하면

- https://gist.github.com/nosy0411/9d3e2a1c029c8f3eb7439a52ec01cacb

assignment1.ipynb 파일과 assignment1.py 파일의 코드를 볼 수 있다.