**Assignment02**

|  |
| --- |
| * 코드는 함께 업로드 된 Jupyter notebook에 작성 * 코드를 작성하고 화면에 print된 것을 그대로 제출 * 결과는 제시된 질문에 따라 이 워드 파일에 기입 * 최종적으로 notebook 파일과 이 워드 파일을 e-class에 같이 제출 * 제출시 파일명 뒤에 학번\_이름을 추가 (‘Assignment02\_학번\_이름’) |

1. 무작위로 생성한 50개의 데이터를 기준으로 nearest neighbor를 찾으려고 한다. 데이터 생성 코드는 제공된 코드를 수정하지 않고 사용한다.

(1) 위의 생성한 데이터 중에서 (0,0)의 10 nearest neighbor를 찾으시오. 이 때 거리는 유클리디안 거리를 사용하고 10 nearest neighbor의 python 기준의 index를 쓰시오. (7점)

(2) 전체 데이터와 (0, 0)에 대한 scatter plot을 그리시오. 이 때 (0,0)은 무작위로 생성한 데이터와 구분되도록 하고, 10 nearest neighbor의 경우에도 nearest neighbor가 아닌 포인트와 구분되도록 scatter plot을 그리시오. (색이나 모양으로 구별) (7점)

참고: <https://matplotlib.org/examples/pylab_examples/scatter_star_poly.html>

(3) 기존 데이터에서 y축의 값을 5배한 다음에 다시 (0,0)의 10 nearest neighbor를 찾으시오. 이 때 거리는 유클리디안 거리를 사용하고 10 nearest neighbor의 python 기준의 index를 쓰시오. (7점)

(4) y축의 값을 5배한 데이터와 (0, 0)에 대한 scatter plot을 그리시오. 이 때 (0,0)은 무작위로 생성한 데이터와 구분되도록 하고, 10 nearest neighbor의 경우에도 nearest neighbor가 아닌 포인트와 구분되도록 scatter plot을 그리시오. (색이나 모양으로 구별) y축의 값을 변경한 후 어떻게 달라졌는지 쓰시오. (7점)

(5) 지금까지의 결과를 바탕으로 y축의 값을 5배한 데이터를 이용해서 nearest neighbor 방법을 이용해 분류나 회귀를 진행한다면 (비록 여기에 output target은 주어져 있지 않지만) 어떤 문제가 있는지 쓰고 이를 해결하기 위해서 어떻게 해야하는지 설명하시오. (5점)

2. 스팸 메일을 판별하는 모형을 decision tree 알고리즘을 이용해 만들려고 한다. 제공되는 spambase.txt 파일을 스팸과 스팸이 아닌 메일로부터 여러 변수들의 대한 값을 포함하고 있다. 맨 마지막 열의 target이 0이면 스팸이 아니고 1이면 스팸이다. 그리고 각 변수의 의미는 함께 첨부된 spambase\_feature.txt에 적혀있다.

(1) decision tree를 학습하는 단계에서 overfitting을 막기 위해서 여러 parameter를 이용하는데 어떤 조합이 가장 좋은지 stratified k-fold cross-validation을 이용해서 알아보고자 한다. 변경하고자 하는 옵션은 max\_depth, min\_samples\_split, min\_samples\_leaf 이고 사용하게 될 값은 다음과 같다.

max\_depth=[3, 4]

min\_samples\_split=[100, 200, 300]

min\_samples\_leaf=[50, 100]

위의 세 옵션에 대해서 5-fold cross-validation을 정확도를 기준으로 수행한다고 했을 때 아래 표에 각 validation step마다 validation set에 의한 정확도를 채우시오. (15점)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Max\_depth** | **Min\_samples\_split** | **Min\_samples\_leaf** | **Fold1** | **Fold2** | **Fold3** | **Fold4** | **Fold5** |
| **3** | **100** | **50** |  |  |  |  |  |
| **100** |  |  |  |  |  |
| **200** | **50** |  |  |  |  |  |
| **100** |  |  |  |  |  |
| **300** | **50** |  |  |  |  |  |
| **100** |  |  |  |  |  |
| **4** | **100** | **50** |  |  |  |  |  |
| **100** |  |  |  |  |  |
| **200** | **50** |  |  |  |  |  |
| **100** |  |  |  |  |  |
| **300** | **50** |  |  |  |  |  |
| **100** |  |  |  |  |  |

(2) 평균적으로 가장 정확도가 높은 조건으로 전체 데이터를 이용해서 decision tree를 학습하고 그 결과 tree를 그리시오. (8점)

(3) 스팸 메일과 아닌 메일을 판별하는 데 어떤 정보를 활용하면 좋은지에 대한 관점으로 최종 decision tree를 해석하시오. (5점)

3. Naïve Bayes 모형을 이용해서 스팸 SMS를 판별하는 모형을 만들고자 한다. 제공되는 spam\_sms.csv 파일은 각 SMS에 특정 단어가 등장한 빈도와 최종 output 값을 포함하고 있다. 최종 target을 ‘target’이라는 이름으로 되어 있고 여기서 스팸인 경우 1, 스팸이 아닌 경우 0이라고 되어 있다. target 앞 쪽의 모든 변수는 특정 단어를 의미한다.

(1) sklearn.model\_selection.train\_test\_split을 이용해서 데이터를 train set과 test set (validation set)으로 나누고자 한다. 먼저 주어진 데이터를 불러와 input에 해당하는 X와 output 값만 담고 있는 y 변수를 생성하고 여기서부터 train\_test\_split를 이용해서 Xtrain, Xtest, ytrain, ytest를 얻으시오. 이 때 test\_size=0.2, stratify=y, random\_state=150으로 설정하시오. Xtrain에 들어있는 데이터 중 맨 위에 있는 5개 데이터와 Xtest에 들어있는 데이터 중 맨 위에 있는 5개 데이터를 jupyter notebook 화면에 print하시오. (5점)

(2) Multinomial naïve Bayes 모형을 이용해서 세팅이 변경없이 train set으로 학습을 하고 test set을 이용해서 정확도(accuracy), recall, precision, f1 measure 값을 구하시오. (8점)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | accuracy | recall | precision | f1 |
| 결과 |  |  |  |  |

(3) 학습된 Multinomial naïve Bayes 모형으로부터 서로 다른 두 group(스팸 SMS와 스팸이 아닌 SMS)에서 확률이 높은 단어 상위 10개씩을 찾으시오. (10점)

참고: <https://docs.scipy.org/doc/numpy/reference/generated/numpy.argsort.html>

|  |  |  |
| --- | --- | --- |
|  | 스팸 SMS | 스팸이 아닌 SMS |
| 1 |  |  |
| 2 |  |  |
| 3 |  |  |
| 4 |  |  |
| 5 |  |  |
| 6 |  |  |
| 7 |  |  |
| 8 |  |  |
| 9 |  |  |
| 10 |  |  |

(4) 학습된 Multinomial naïve Bayes 모형에서 서로 다른 두 group에 대해서 개별 단어의 값들의 히스토그램을 그리시오. 이 때 히스토그램을 그리기 위한 막대의 수는 30으로 하시오. 히스토그램으로부터 알 수 있는 두 group의 차이점을 쓰시오. (8점)

(5) Bernoulli naïve Bayes 모형을 이용해서 세팅 변경 없이 train set으로 학습을 하고 test set을 이용해서 정확도(accuracy), recall, precision, f1 measure 값을 구하시오. (2)의 결과와 비교하여 어느 쪽이 더 나은 모형인지 설명하시오. 이 결과로부터 알 수 있는 사실은 무엇인가? (8점)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | accuracy | recall | precision | f1 |
| 결과 |  |  |  |  |