|  |  |
| --- | --- |
| [Machine Learning]  [2021-1] |  |
| Homework 3  Lec 8, 9, 10 |  |
| [Date] 2021.05.14  Student ID : 2016112158  Name : KimHeeSuf  Professor : Juntae Kim | logo-placeholder |

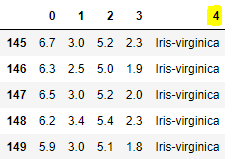
1. Explain what Feature Selection and Dimensionality Reduction are, and why they are needed in machine learning. You can use examples for explanation. (10pts)

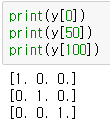
|  |
| --- |
| Your Answer |
| feature selection은 데이터의 여러 feature 중에서 예측에 있어서 해당 feature가 유용한지 유용하지 않은지 확인하는 과정을 뜻한다. 유용하다면 해당 feature를 사용하고, 유용하지 않다면 버린다. 불필요한 특징을 제거하여 간결한 특징 집합을 만드는 것이다. 예를 들어, 사람의 나이를 예측할 때 feature로 얼굴 주름의 개수, 흰머리의 개수,피부색, 발의 크기 등이 있다고 가정하자. 이때 얼굴 주름의 개수나 흰머리의 개수는 나이를 예측할 때 유용하지만 발의 크기나 피부색 등은 유용하지 않다고 볼 수 있다. 따라서 나이 예측에 있어서 얼굴주름의 개수, 흰머리 개수 만으로 feature를 설정하는 것을 feature selection으로 볼 수 있다.  Dimesionality Reduction은 관찰 대상을 잘 설명할 수 있는 잠재 공간(latent space)은 실제 공간( observation space)보다 작을 수 있다. 이렇게 관찰 공간 위의 샘플을 기반으로 잠재공간을 파악하는 것을 차원축소라 한다. iris 데이터의 feature들인 sepal\_length, sepal\_width, petal\_length, petal\_width에 대해서 PCA를 통해 2차원 공간으로 만들 수 있다 |

2. Answer following questions. (20pts)

2-1. Perform one-hot encoding on the class of Iris data.

Expected Output

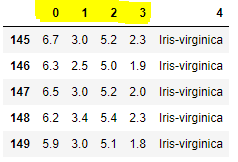


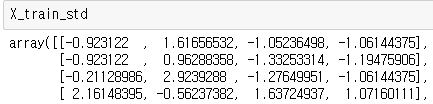


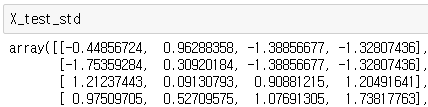
|  |
| --- |
| Code |
| import pandas as pd  import numpy as np  iris = pd.read\_csv("./iris.csv")  df = pd.DataFrame(iris)  df.columns=[0,1,2,3,4]  from sklearn.preprocessing import LabelEncoder  le = LabelEncoder()  df[4] = le.fit\_transform(df[4])  onehot\_df = pd.get\_dummies(df,columns=[4])    y = np.array(onehot\_df[["4\_0","4\_1","4\_2"]], dtype=np.float32)  type(y)  print(y[0])  print(y[50])  print(y[100]) |
| Result(Captured images) |
|  |
| Description |
| iris.csv를 읽어와서 column을 설정해준 후, LabelEncoder로 3개의 label을 0,1,2로 바꾼다. 그 후 get\_dummies로 onehot encoding해주면 “4\_0”, “4\_1”, “4\_2”로 인코딩된다. 해당 부분만 np.array로 가져오면 된다 |

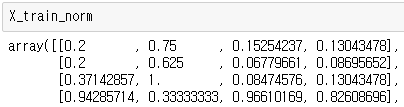
2-2. Perform Standardization and Normalization on the Iris data(except class. split train and test set)

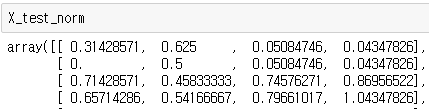
Expected Output







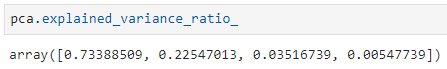


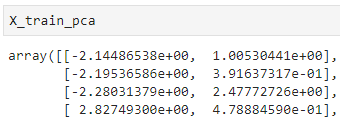
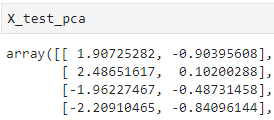


|  |
| --- |
| Code |
| from sklearn.model\_selection import train\_test\_split  X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y,test\_size=0.3, random\_state=0, stratify=y)  from sklearn.preprocessing import MinMaxScaler  from sklearn.preprocessing import StandardScaler  mms = MinMaxScaler()  X\_train\_norm = mms.fit\_transform(X\_train)  X\_test\_norm = mms.transform(X\_test)  sc = StandardScaler()  X\_train\_std = sc.fit\_transform(X\_train)  X\_test\_std = sc.transform(X\_test) |
| Result(Captured images) |
|  |
| Description |
| MinMaxScaler객체 mms를 선언하여 X\_train\_norm, X\_test\_norm으로 변환하였고, StandardScaler 객체 sc를 선언하여 X\_train\_std, X\_test\_std로 변환하였다 |

2-3. Perform PCA on the iris data, check the explained variance ratio and choose 2 components.

Expected Output



|  |
| --- |
| Code |
| from sklearn.decomposition import PCA  pca = PCA()  X\_train\_pca = pca.fit\_transform(X\_train\_std)  X\_test\_pca = pca.transform(X\_test\_std)  pca.explained\_variance\_ratio\_ |
| Result(Captured images) |
|  |
| Description |
| PCA객체 pca를 선언하여 차원축소 하여 X\_train\_pca, X\_test\_pca로 변환하였다. 예시에서, explained\_variance\_ratio\_는 4개인데 X\_train\_pca[0]의 차원은 2D이다. explained\_variance\_ratio\_가 4개라면 X\_train\_pca[0]의 차원은 4D여야 하는거 아닌가.. |

3. Explain what PCA(Principle Component Analysis) is and why you need it in machine learning. (10pts)

|  |
| --- |
| Your Answer |
| PCA는 데이터의 variance를 최대한 보존하면서 서로 직교하는 축을 찾아 고차원 공간의 표본을 저차원으로 투영하는 차원축소 기법이다. 데이터의 feature가 많아 차원이 늘어나면 데이터 공간의 부피가 기하급수적으로 증가하여 데이터의 밀도가 차원이 증가할수록 희소(sparse)해진다. 밀도가 희소해진다는건 데이터포인트 간의 거리가 증가한다는 것이고, 이를 이용해 머신러닝 모델을 학습시키면 모델이 오버피팅될 위험이 커진다. 이를 해결하기 위해 분산을 최대한 보존하는PCA를 통해 차원의 수를 줄일 필요가 있다. |

4. Explain what Bias and Variances are and how they affect the performance of machine learning models. (10pts)

|  |
| --- |
| Your Answer |
| Bias는 지나치게 단순한 모델로 인한 error이다. 예측값과 실제값의 차이로 볼 수 있다. Bias가 높으면 underfitting이 발생한다.  Variance는 지나치게 복잡한 모델로 인한 error이다. Variance는 주어진 데이터로 학습한 모델이 예측한 값의 변동성을 의미한다. Variance가 높으면 overfitting이 일어난다. |

5. Apply Bagging and Adaboost algorithms on Iris dataset. Set the base\_estimator to “DecisionTreeClassfier”. We recommend you use “BaggingClassifier” and “AdaBoostClassifier” in Scikit-learn. (20pts)

Iris data Setting(Use this Code)

import numpy as np

import pandas as pd

df = pd.read\_csv('https://archive.ics.uci.edu/ml/'

'machine-learning-databases/iris/iris.data', header=None)

X = df.iloc[50:150, [2, 3]].values

y = df.iloc[50:150, 4].values

y = np.where(y == 'Iris-virginica', -1, 1)

Expected Output

print("Decision Tree train/test accuracies %.3f/%.3f"

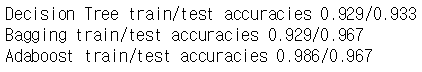
% (tree\_clf.score(X\_train, y\_train), tree\_clf.score(X\_test, y\_test)))

print("Bagging train/test accuracies %.3f/%.3f"

% (bag\_clf.score(X\_train, y\_train), bag\_clf.score(X\_test, y\_test)))

print("Adaboost train/test accuracies %.3f/%.3f"

% (boost\_clf.score(X\_train, y\_train), boost\_clf.score(X\_test, y\_test)))



|  |
| --- |
| Code |
| hw3-5.ipynb 참고 |
| Result(Captured images) |
|  |
| Description |
| X\_train, X\_test,y\_train,y\_test를 생성하고 DecisionTreeClassifier 객체 tree, BaggingClassifier객체 bag, AdaBoostClassifier 객체 ada를 선언하고 학습한 후, score를 측정하였다. |

6. Apply Logistic Regression and Decision Tree Classifier on 20newsgroups Dataset(Document Classification). (30pts)

20newsgroup dataset is a news dataset consisting of 20 categories.

In this question, we only use samples from 4 categories.

Follow this process:

1) Load the data.



2) Make preprocessor function & porter stemmer tokenizer function.

3) Apply the TF-IDF(TFidfVectorizer) on the data.

- Use the preprocessor function & porter stemmer tokenizer

- Use stop-words

- Drop terms occurred in more than 10% of docs

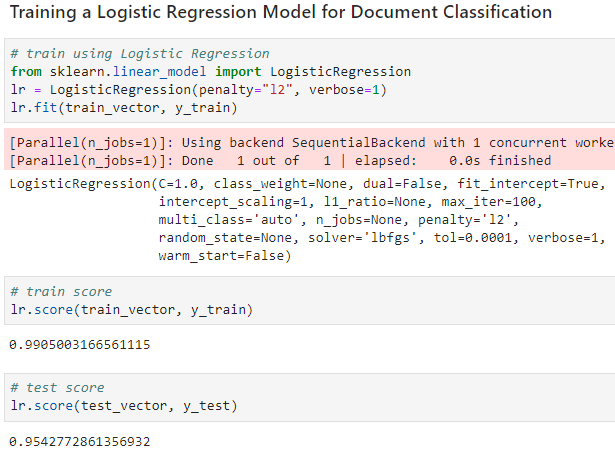
- Drop terms occurred in less than 10 docs

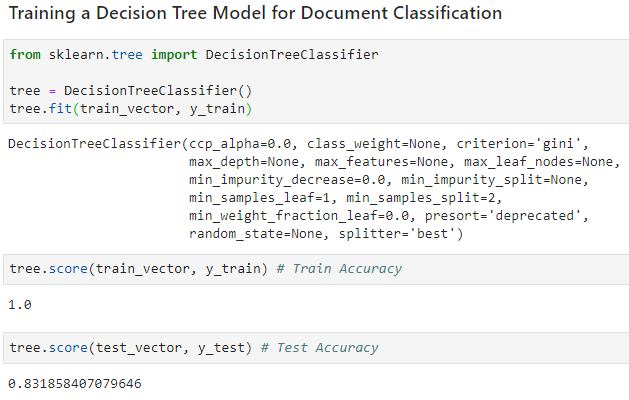
4) Train machine learning models(Logistic Regression, Decision Tree) using TF-IDF vectors.

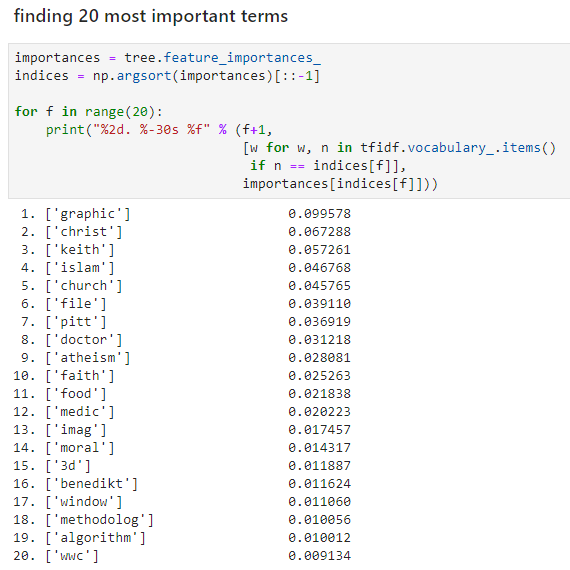
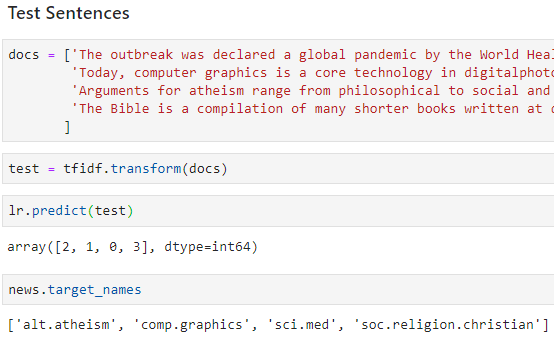
5) Predict the categories of following 4 sentences

|  |
| --- |
| 'The outbreak was declared a global pandemic by the World Health Organization (WHO) on 11 March.',  'Today, computer graphics is a core technology in digitalphotography, film, video games, cell phone and computer displays,and many specialized applications.',  'Arguments for atheism range from philosophical to social and historical approaches.',  'The Bible is a compilation of many shorter books written at different times by a variety of authors, and later assembled into the biblical canon.' |

Expected Output





|  |
| --- |
| Code |
| hw3-6.ipynb 참고 |
| Result(Captured images) |
|  |
| Description |
| X\_train, X\_test, y\_train, y\_test를 생성하고 re 모듈을 사용해 preprocessor함수를 정의하였다. 그리고 nltk.stem.porter에서 PorterStemmer를 이용해 tokenizer\_porter함수도 정의하였고 nltk모듈을 활용해 stopword또한 생성하였다  그 후, TfidVectorizer객체를 생성하여 X\_train, X\_test를 transform하였다. transform된 X\_train\_vector를 기반으로 로지스틱회귀모델과 결정트리 모델을 학습하고 이 모델로 docs에 대한 score를 측정하였다 |

**Note**

1. Summit the file to e-class as pdf.

2. Specify your file name as “hw3\_<StudentID>\_<Name>.pdf”