기계학습 프로젝트 outline

지금까지 실습 때 배운 내용정리

* 1차시 : zero-R, one-R, Naïve Bayes Classifier, \*object type 정수형으로 전처리
* 2차시 : decision tree, pruning
* 3차시 : linear regression, logistic regression (sigmoid, odds..)
* 4차시 : perceptron, mlpclassifier , 파라비터 조정 (one-hot encoding)
* 5차시 : KNN Classifier, Cross Validation
* 6차시 : (성능평가) confusion matrix, ROC&AUR, Z-score normalization
* 7차시 : (앙상블) random forestclassifier, bagging, boosting, (clustering) k-means, (feature selection)

Mushroom data 사용하자

* (more than 4 classification or regression models at minimum + more than 1 clustering model)

1. Decision tree classifier
2. Linear regression(perceptron)
3. Mlpclassifier
4. Knn classifier
5. Randomforeset
6. K-means

Feature selection 기법 최소 하나

Apply clustering to your dataset and create clusters. Then, compare the resulting clusters and the labels of the classification.

* + The dataset should be divided into train, validation and test set or be divided into train and test set. For the latter case, the validation process should be done through cross-validation. The final model evaluation must be done with test dataset.

**Conditions on Dataset Selection**

* At least more than 10 features (attributes)
* At least more than 5,000 instances

질문사항

* . If there are differences among the model’s performance, analyze the model itself if the model can be analyzed. If the model cannot be analyzed (meaning black-box models), you should provide the analysis by inferring through the dataset features
* 여기에서 analyzed 가능한 모델과 불가능한 모델이 어떤 의미인지 궁금합니다
* Cross validation을 모든 모델에 대해 적용해야하나요
* Feature selection 전후 비교과정을 모든 모델에 대해 진행해야하나요
* Model parameter 변경을 모든 모델에 대해 진행해야하나요

총 attribute 22개 , class 1개

총 instance 8124개

전처리 과정

* Na값 있음. (?로 표시되어 있음), 2480 개
* Missing value 제거 하고 난 instance 수 : 5644
* 모든 feature가 object type이므로 정수형으로 전처리
* #dt['class'].value\_counts()
* #dt['cap-color']
* 이런 명령어로 encoding 된 값이 뭔지 확인 하자
* Train, test로 나눔
  1. Decision tree

진행상황

* Dt, perceptron, mlp, knn, forest, k-means 기본 skitlearn으로 만듬
* 고려해야할 상황
* 굳이 모든 feature를 정수형으로 바꿔야했었나? 그냥 class만 정수형으로 하면 안되나? 특히 decision tree에서
* 성능이 너무 좋은데?
* 현재 값들이 encoding되었는데 0,1인 feature도 있고 0~6인 feature도 있는데 normalized 필요 없나? -> one-hot encoding
* X\_train, y\_trains test 가 기존 dataset에서의 index를 가지고있음. Index 0부터 하고 싶으면 to\_numpy() 사용,
* Kmeans 에서 예측된 label 과 실제 test data label 이 좀 많이 다른듯 하다

Part 1 : 데이터 불러오기 및 데이터 전처리

Part 2 : feature selection 없이 6개의 모델 만들고 분석

Part 3 : feature selection 진행

Part 4: feature selecton 후 part 2 다시 진행

Part2

1. Decision tree
   * + 1. Train data와 test data만을 이용하여 default 모델 만들자
       2. Validation 과정을 통해 최적의