TERM PROJECT

C C A 2 0 2 1

컴퓨터공학부 2017-18538 황선영

컴퓨터공학부 2017-12734 윤종선

CONTENTS

01

Topic: Breast cancer dataset

Breast cancer dataset

569 data (2 class: 212 malignant, 357 benign)

30 numerical features

provided with sklearn.datasets package

Task

Splitting the dataset into the Training set and Test set

Apply dimension reduction into training and testing set for analysis

Classify malignant / benign cancers

02

Method we plan to use

Procedure

1. Split the dataset into the Training set and Test set
2. Apply the PCA function into training and testing set for analysis.
3. Fit Ridge / Lasso / Logistic Regression To the training set
4. Predict the test set result and training set result
5. Visualise the Test set results
6. Compare results according to the regression method

03

Expectation of final result

Mean squared error

```
Mean squared error: 0.041460
Mean squared error: 0.056980
```

[Figure1] Mean squared error of Lasso / Ridge regression

Logistic regression에 대해서도 적절한 MSE 값을 얻을 것으로 예상

Classification

Ridge regression: 변수간 상관관계가 높은 상황에서 높은 예측 성능

Lasso regression: 변수간 상관관계가 높은 상황에서 Ridge보다 낮은 예측 성능

Logistic regression: Sigmoid 함수의 효과로 인해 범위를 벗어나는 경우를 예방하고 정확도가 떨어지는 것을 방지

THANK YOU