AI Final Project

Korea University AI Class (Cose-361, Section Number 2, Spring, 2018) Final Project Report,

Analyze two public domain datasets (Student Performance, Breast Cancer) using four or more classifiers.

Author: Jinheon Baek

Computer Science and Engineering, Korea University, Seoul, Korea

Jinheon.Baek@outlook.kr

Friday, Jun, 22, 2018

Abstract

해당 레포트는 인공지능 수업에서 배운 다양한 Classification 학습 모델을 실 데이터에 적용해보는 과정을 담고 있습니다. 모델의 학습과 평가를 위해 UCI Machine Learning Repository에 존재하는 두 개의 Dataset을 선정하였으며, 두 데이터는 아래와 같습니다. 1) Student Performance Data Set, 2) Breast Cancer Coimbra Data Set. 두 데이터는 Class Label 이 Multivariate 한 특성이 있고, 모두 적지 않은 수의 features 를 담고 있습니다. 전 해당 Classification 모델을 개발하기에 앞서, 필요한 부분에 대해서는 One-hot encoding을 비롯한 다양한 Pre-processing 작업들을 해주었으며, 모델의 Baseline 으로 ZeroR, OneR 모델을 사용하였습니다. 또한 Python - sklearn에 담겨있는 총 8개의 Classification 모델을 사용하였으며, 사용한 모델은 수업시간에 배운 모델과, 해당 모델을 응용한 모델로 다음과 같습니다. 1) K-Nearest Neighbors, 2) Linear SVM, 3) RBF SVM, 4) Decision Tree, 5) Random Forest, 6) Neural Network, 7) AdaBoost (Ensemble), 8) Gaussian Naïve Bayesian. Student Performance Data Set에 대해서는 Linear SVM(Accuracy: 0.855) 이 가장 좋은 성능을 보였으며, Breast Cancer Coimbra Data Set에 대해서는 Neural Network(0.747) 이 가장 좋은 성능을 보였습니다.

[Student Performance Classification]

1. Introduction (데이터셋을 선택한 이유)

성적은 학생을 평가하는 지표 중 주요한 요소로 사용되고 있습니다. 학생의 성적을 매길 때 대부분의 학교에서 주요하게 사용되는 요소 중 하나는 공부한 내용을 바탕으로 출제되는 필기 시험 성적일 것입니다. 하지만 시험의 경우 학생의 그 날 컨디션과, 운에 따라 좌우될 수 있는 여지가 남아있습니다.

학생이 보인 한 학기동안의 지표 뿐만 아니라 학생의 개인 정보, 직전학기 성적을 바탕으로 학생의 이번학기 성적을 예측하는 모델을 만들 수 있으며, 해당 모델을 사용한다면 학생의 필기 시험 성적에 더해 학생을 평가하는 더 좋은 지표로서 활용될 수 있습니다.

전 이전부터 운과 컨디션에 좌우될 수 있는 필기시험과, 해당 필기시험이 전체 성적 평가의 많은 비중을 차지한다는 것에 아쉬움을 느꼈고, 필기시험 뿐만 아니라 학생이 가지고 있는 정보를 바탕으로 성적을 측정할 수 있는 모델을 만든다면 더욱 명확한 성적 산정 기준이 만들어질 수 있을 것이라고 생각하여, 해당 데이터셋을 사용하여 아래와 같이 분류 모델을 개발하였습니다.

2. Explore Data

UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets/Student+Performance

Student Performance Data Set 은 총 33개의 Attributes 로 구성되어 있습니다. 해당 데이터는 포르투칼에 있는 두 학교 학생들에 대해, 학생들의 개인 정보와 수학, 포르투갈 언어의 성적을 기록해 놨으며, 학생들의 개인 정보와 이전의 성적 (1^{st}, 2^{nd} period grades) 정보를 바탕으로 최종 성적(3^{rd} period)을 예측하는 문제입니다.

최종 성적을 위해 주어진 Attribute 정보는 다음과 같습니다.

1) school - student's school

   (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)

2) sex - student's sex

   (binary: 'F' - female or 'M' - male)

3) age - student's age

(numeric: from 15 to 22)

4) address - student's home address type

(binary: 'U' - urban or 'R' - rural)

5) famsize - family size

(binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)

6) Pstatus - parent's cohabitation status

(binary: 'T' - living together or 'A' - apart)

7) Medu - mother's education

(numeric: 0 - none, 1 - primary education (4th grade),

2 - 5th to 9th grade, 3 - secondary education, or 4 - higher education)

8) Fedu - father's education

(numeric: 0 - none, 1 - primary education (4th grade),

2 - 5th to 9th grade, 3 - secondary education, or 4 - higher education)

9) Mjob - mother's job

(nominal: 'teacher', 'health': care related, civil: 'services', 'at_home' or 'other')

10) Fjob - father's job

(nominal: 'teacher', 'health': care related, civil: 'services', 'at_home' or 'other')

11) reason - reason to choose this school

(nominal: close to 'home', school 'reputation', 'course' preference or 'other')

12) guardian - student's guardian

(nominal: 'mother', 'father' or 'other')

13) traveltime - home to school travel time

(numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)

14) studytime - weekly study time

(numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

15) failures - number of past class failures

(numeric: n if 1<=n<3, else 4)

16) schoolsup – extra educational support

(binary: yes or no)

17) famsup – family educational support

(binary: yes or no)

18) paid – extra paid classes within the course subject (Math or Portuguese)

(binary: yes or no)

19) activities – extra-curricular activities

(binary: yes or no)

20) nursery – attended nursery school

(binary: yes or no)

21) higher – wants to take higher education

(binary: yes or no)

22) internet – Internet access at home

(binary: yes or no)

23) romantic – with a romantic relationship

(binary: yes or no)

24) famrel – quality of family relationships

(numeric: from 1 – very bad to 5 – excellent)

25) freetime – free time after school

(numeric: from 1 – very low to 5 – very high)

26) goout – going out with friends

(numeric: from 1 – very low to 5 – very high)

27) Dalc – workday alcohol consumption

(numeric: from 1 – very low to 5 – very high)

28) Walc – weekend alcohol consumption

(numeric: from 1 – very low to 5 – very high)

29) health – current health status

(numeric: from 1 – very bad to 5 – very good)

30) absences - number of school absences

(numeric: from 0 to 93)

31) G1 - first period grade

(numeric: from 0 to 20)

32) G2 - second period grade

(numeric: from 0 to 20)

33) G3 - final grade (target)

(numeric: from 0 to 20, output target)

위와 같이 해당 Dataset 에는 Target Feature - G3을 포함해서 총 33개의 Attributes 로 구성되어 있으며, 해당 Attributes를 구성하는 데이터 개수는 395개입니다(Rows 개수). 또한 수학 과목과 포르투갈 언어 각각에 대해 데이터가 기록되어 있습니다.

전 해당 데이터셋 중 "수학" 하나를 선택하였으며, 32개의 Attributes 를 가지고 이 학생의 3rd period 수학 성적을 예측할 것입니다.

3. Pre-processing

해당 데이터는 Numeric한 Attributes 뿐만 아니라, Attributes 내 값들이 String 인 경우도 있습니다. Ex) sex: M (Male) or F (Female). 따라서 해당 Attributes 들은 숫자로 표현되지 않으므로 Machine Learning 모델을 만들 때 학습이 원활하게 이루어지지 않을 수 있으며, 이러한 String Attributes에 대해 Label Encoding을 사용하여 표현할 경우 Weights 가 잘못 학습될 위험이 있습니다.

따라서 전 해당 String Attributes에 대해 One-hot encoding을 사용하였으며, One-hot encoding의 결과인 sparse matrix 를 통해 해당 Features 들을 학습하였습니다.

뿐만 아니라 해당 dataset의 data 개수가 355개 밖에 되지 않지만, 맞춰야 하는 G3 column의 데이터 범위가 0에서 20 사이의 정수라는 것을 보고 학습이 쉽지 만은 않을 것 같다고 느꼈습니다.

따라서 전 0부터 20 사이의 범위를 가지고 있는 G3 column에 대해 5로 나눈 몫을 사용하여 구간을 좁혔으며, 따라서 제가 학습한 모델은 0부터 20 사이의

Class를 예측하는 Classification Model이 아닌, 0부터 5 사이의 Class를 예측하는 Classification Model이 됩니다.

마지막으로 데이터를 import 할 때부터 정렬되어 있는 데이터의 index 순서를 바꿔 학습해야 전체 domain의 데이터를 보다 잘 표현하여 학습할 수 있다고 생각하였고(만약 import 시점의 데이터 index 순서가 의미가 있었다면), 따라서 random permutation을 통해 데이터를 re-indexing 해주었습니다.

4. Classification Model

전 3rd period 수학 성적을 예측하는 모델 개발을 위해 Baseline 으로 OneR, ZeroR 모델을 택했습니다. 뿐만 아니라 수업시간에 언급하고 넘어갔던 아래와 같은 8가지 모델을 사용하였습니다.

각 모델의 이름과 설정한 Hyper-parameter 은 다음과 같습니다.

1) K-Nearest Neighbors

   K: 3

2) Linear SVM (Kernel: Linear)

   Kernel: linear, C (penalty parameter of the error term): 0.025

3) RBF SVM (Kernel: RBF)

   Kernel: RBF, C (penalty parameter of the error term): 0.025

4) Decision Tree

   Max-depth: 10

5) Random Forest

   Max-depth: 10, number of estimators: 300

6) Neural Network

   Alpha (L2 Regularization term): 1, hidden-layer: (100, 50, 30)

7) Ada Boost (Ensemble based on the Decision Tree)

   Number of estimators: 500

8) Gaussian Naïve Bayes

Default.

해당 Hyper-parameters는 각 모델에서 사용되는 기본 Hyper-parameters에서 성능 향상을 도모하기 위해 조금씩 Tuning 해준 값들을 사용하고 있습니다. 물론 하단에 기술되어 있는 Cross-Validation을 통해 각 모델의 Hyper-parameters 값들을 세밀하게 조정해주는 과정을 반복해서 거쳐 성능이 가장 우수한 한 모델을 선택할 수 있습니다.

하지만 Hyper-Parameter을 세밀하게 맞춰가며 성능이 가장 우수한 한 모델을 선발하는 것 보다, 다양한 모델들을 직접 돌려보고 성능을 비교하는 것에 더 큰 의미를 두었기에 위와 같이 비교적 간단하게, 한 모델에 대해 한 Hyper-parameter을 두어 Classification 모델을 학습하였습니다.

5. Evaluation

ZeroR, OneR을 포함한 10개의 Classification Model에 대한 성능 비교를 위해 Cross Validation을 사용하였습니다. 전 K-fold Cross Validation에서 K 값을 10으로 두어, 전체 데이터의 90%는 Training에, 10%는 Test에 사용되게 하였고, 각각의 모델을 평가하는 주요 척도로 Accuracy를 사용하였습니다.

뿐만 아니라 각 모델에서 나온 예측 값을 바탕으로 Precision, Recall, F1 score 역시 계산하였으며, 해당 결과를 Model의 Accuracy와 함께 출력되게 만들어, 해당 척도들 역시 모델을 평가할 때 사용될 수 있게끔 개발하였습니다.

Accuracy가 높은 순서로 각 모델의 이름과 Accuracy, Precision, Recall, F1 score에 대해 표로 정리하자면 다음과 같습니다. 해당 표에서 사용된 모델의 Hyper-parameter는 4. Classification Model에서 서술한 Hyper-parameter와 동일합니다. 또한 각 Model에서 보이는 Score 들은 소수점 아래 4자리에서 반올림 한 수치를 사용하고 있습니다.

Model	Accuracy	Precision	Recall	F1 Score
Linear SVM	0.855	0.854	0.856	0.855
AdaBoost	0.826	0.862	0.828	0.824
Neural Network	0.822	0.819	0.820	0.819
Decision Tree	0.815	0.817	0.808	0.809
Random Forest	0.812	0.816	0.818	0.815
Nearest Neighbors	0.749	0.752	0.749	0.747
OneR	0.648	0.460	0.648	0.528
RBF SVM	0.486	0.236	0.648	0.528
Zero R	0.486	-	-	-
Naïve Bayes	0.436	0.564	0.435	0.407

위 결과를 보면 Linear SVM 이 가장 좋은 성능을 보이고 있습니다. SVM의 경우 각 class에 대해 maximum margin을 갖는 Linear 선을 찾는 학습 모델이며, Kernel로 Linear Kernel을 사용하였습니다. 또한 수업시간에 배운 AdaBoost (Ensemble), Neural Network, Decision Tree, Random Forest 기법이 모두 Baseline으로 삼았던 OneR 모델에 비해 좋은 성능을 보였습니다.

반면 RBF SVM의 경우 OneR 보다 좋은 성능을 보이지 못하는데, 이는 Kernel Function 으로 사용되는 RBF가 exponential 한 의미를 담고 있지만, 내 Attributes 들은 exponential 한 의미를 담고 있지 않아 좋은 성능을 보이지 않는 것이라고 추측할 수 있습니다.

마지막으로 Baseline 으로 삼았던 ZeroR에 비해 Gaussian Naïve Bayesian 은 더 좋지 못한 성능을 보입니다. Gaussian Naïve Bayesian 은 Naïve Bayesian에

서 각 Attributes 들의 Naïve Bayes 값을 계산할 때 Gaussian 분포를 이용한 것으로, 각 Features 내에 속한 Data의 분포가 Bell-shaped을 따르지 않을 뿐만 아니라, One-hot encoding 으로 표현했던 String 데이터들에 대해 낮은 확률 값을 보이는 경우가 많아 높은 성능을 보이지 않은 것이라고 추측할 수 있습니다.

6. Result

포르투갈 내 두 학교에 재학중인 학생의 개인 정보와 이전 학기의 수학 성적들을 바탕으로 3rd period의 수학 성적을 예측하는 Classification 모델을 위처럼 만들어 보았습니다. 학습을 위해 Dataset 내 총 395개의 Data를 사용하였고, 각 Data 들은 Target variable 'G3'을 포함해서, 총 33개의 Attributes 로 구성되어 있습니다.

전 String 데이터에 대한 One-hot encoding 뿐만 아니라, Classify 해야 하는 Target variable의 범위를 줄여 Classification이 보다 원활하게 이루어지게끔 만들었으며, baseline 으로 사용한 ZeroR, OneR 모델 뿐만 아니라 수업시간에 다룬 모델을 포함하여 총 10개의 학습 모델을 사용하였습니다.

ZeroR을 제외한 각 모델들은 모두 10-Fold Cross Validation을 사용하여 평가하였으며, 평가 핵심 지표로 Accuracy를 두었고, Precision, Recall, F1 score 역시 함께 볼 수 있게끔 표로 정리하여 보였습니다.

정확도 기준, 10개의 모델들을 돌려 본 결과 나타난 가장 우수한 모델은 Linear Kernel Function을 사용하는 SVM 모델입니다. 해당 모델의 정확도는 0.855 로 5개의 Target Values 를 적절하게 Classify 해야 하는 모델임에도 불구하고 비교적 높은 정확도를 보였습니다. 뿐만 아니라 해당 모델이 다른 9개의 모델보다 우수한 F1 score를 보이기도 했습니다.

물론 G3 와 Correlation 이 매우 높은 G1, G2 라는 이전학기 수학 성적들을 사용하였기에 높은 정확도를 보이는 것일 수도 있지만, 학생의 개인 정보와 이전학기 수학 성적들을 바탕으로 3rd period의 수학 성적을 꽤 정확하게 예측하는 것으로 보아 추후 성적평가 방식에 필기고사 뿐만 아니라 다양한 지표를 활용한 모델을 사용할 수 있을 것이라고 생각합니다. 하지만 추후 G1, G2 Features 를 제외한 Classification 모델에서도 성능이 잘 나오는지 검증해 볼 필요는 있습니다.

[Breast Cancer Coimbra Classification]

1. Introduction (데이터셋을 선택한 이유)

암은 과거부터 지금까지 각 나라별 사망률에 높은 비율을 차지하는 질병 중 하나로 알려져 있습니다. 특히 유방암의 경우 모든 암 중에서 가장 연구가 많이 된 편임에도 불구하고 환경적 요인과 유전적 요인 두 가지에 의해 발생한다는 명확하지 않은 지식만이 있으며, 아직 유방암의 발생 원인에 관해서는 확립된 정설이 없습니다.

전 평소 바이오 문제에 관심이 많은 학생이고, 특히 기계학습을 이용하여 산재되어 있는 바이오 도메인의 문제를 해결하는 것에 큰 흥미를 가지고 있습니다. 따라서 암의 일종인 유방암을 예측할 수 있는 한 방법으로, 인체측정학 및 혈액에서 나온 호르몬, 단백질, 유전자 표현 등의 데이터를 바탕으로 유방암 진단 마커 모델을 개발해보는 해당 과제를 선택하게 되었습니다.

2. Explore Data

UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra

Breast Cancer Coimbra Data Set 은 총 10개의 Attributes 로 구성되어 있습니다. 해당 데이터는 포르투갈에 있는 사람에 대해 환자의 의료 정보를 바탕으로 해당 환자에게 유방암이 존재하는지 아닌지를 기록해 놨으며, 환자 의료 정보를 바탕으로 환자의 유방암 유무를 예측하는 문제입니다.

유방암 예측을 위해 주어진 Attribute 정보는 다음과 같습니다.

[Target Features을 제외한 모든 Attributes는 모두 Quantitative 합니다.]

1) Age (years)
2) BMI (kg/m2)
3) Glucose (mg/dL)

   (체내 포도당 비율)
4) Insulin (μU/mL)

   (체내 인슐린 비율)

5) HOMA

(체내 인슐린 저항성과 인슐린을 만드는 세포인 B-cell의 활성 정도를 통해 측정한 값)

6) Leptin (ng/mL)

(식욕, 배고픔, 물질대사 등을 포함한 에너지 섭취, 소비 조절에 중요한 역할을 하는 호르몬의 체내 양)

7) Adiponectin (μg/mL)

(지방 세포에서 분비되는 단백질의 일종으로, 인슐린 저항성을 발생시키는 결정적 요소의 체내 양)

8) Resistin (ng/mL)

(RETN 유전자에 의해 인코딩 되어있는 호르몬의 체내 양으로 심장 병의 위험을 높이는 요소, 콜레스테롤과 관련이 있음)

9) MCP-1(pg/dL)

(단백질의 일종으로, 단핵 세포 내 식세포의 이동과 침윤을 조절하는 핵심 단백질 분자의 체내 양)

10) Target

(1: Healthy, 2: Patient)

위와 같이 해당 Dataset 에는 Target Feature을 포함해서 총 10개의 Attributes 로 구성되어 있으며, 해당 Attributes를 구성하는 데이터 개수는 116개입니다(Rows 개수).

전 해당 데이터셋에 있는 9개의 Attributes 를 가지고 해당 환자의 유방암 진단 분류 모델을 개발할 것입니다.

3. Pre-processing

앞서 분석한 Student Performance Dataset과 다르게 해당 데이터를 이루는 Attributes의 Data type은 모두 Numeric 하게 표현되어 있습니다. 따라서 데이터를 학습시키기에 앞서 작업했던 String Attributes에 대한 One-hot encoding을 해주지 않아도 됩니다.

해당 데이터셋을 이루는 Target Attribute를 제외한 Features 들은 모두 Numeric 한 값으로 구성되어 있어, 해당 데이터셋에 Gaussian 분포 등을 취해 각 Features 들을 Normalize 해줄 수 있습니다. 또한 Normalize 를 통해 모델

을 학습한다면 학습 과정 내에서 “거리 정보”을 이용하는 모델 뿐만 아니라 다른 모델들에 대해 모두 좋은 성능을 보일 수 있고, 따라서 전 Target Attribute을 제외한 모든 Features 들에 대해 Gaussian 분포를 취해 Normalize 해주었습니다.

마지막으로 앞서 사용한 방법처럼, 데이터를 import 할 때부터 정렬되어 있는 데이터의 index 순서를 바꿔 학습해야 전체 domain의 데이터를 보다 잘 표현하여 학습할 수 있다고 생각하였고(만약 import 시점의 데이터 index 순서가 의미가 있었다면), 따라서 random permutation을 통해 데이터를 re-indexing 해주었습니다.

4. Classification Model

앞서 언급한 방식처럼 전 유방암의 발병 여부를 예측하는 모델 개발을 위해 Baseline 으로 OneR, ZeroR 모델을 택했습니다. 뿐만 아니라 수업시간에 언급하고 넘어갔던 아래와 같은 8가지 모델을 사용하였습니다.

각 모델의 이름과 설정한 Hyper-parameter 은 다음과 같습니다.

1) K-Nearest Neighbors

   K: 3

2) Linear SVM (Kernel: Linear)

   Kernel: linear, C (penalty parameter of the error term): 0.025

3) RBF SVM (Kernel: RBF)

   Kernel: RBF, C (penalty parameter of the error term): 0.025

4) Decision Tree

   Max-depth: 8

5) Random Forest

   Max-depth: 8, number of estimators: 50

6) Neural Network

   Alpha (L2 Regularization term): 1, hidden-layer: (50, 30)

7) Ada Boost (Ensemble based on the Decision Tree)

Number of estimators: 100

8) Gaussian Naïve Bayes

Default.

해당 Hyper-parameters는 각 모델에서 사용되는 기본 Hyper-parameters에서 성능 향상을 도모하기 위해 조금씩 Tuning 해준 값들을 사용하고 있습니다. 다만 위에서 사용했던 Student Performance 데이터와는 다르게 해당 데이터셋을 이루고 있는 데이터의 개수는 116개 밖에 되지 않으며, 따라서 Capacity가 큰 모델을 개발한다면 Overfitting 되기 쉬운 상황이 발생하기에 Model의 Complexity를 줄이는 방향으로 Hyper-parameter을 설정하였습니다.

또한 앞서 언급했던 것처럼, 하단에 기술되어 있는 Cross-Validation을 통해 각 모델의 Hyper-parameters 값들을 세밀하게 조정해주는 과정을 반복해서 거쳐 성능이 가장 우수한 한 모델을 선택할 수 있습니다.

하지만 Hyper-Parameter을 세밀하게 맞춰가며 성능이 가장 우수한 한 모델을 선발하는 것 보다, 다양한 모델들을 직접 돌려보고 성능을 비교하는 것에 더 큰 의미를 두었기에 위와 같이 비교적 간단하게, 한 모델에 대해 한 Hyper-parameter을 두어 Classification 모델을 학습하였습니다.

5. Evaluation

ZeroR, OneR을 포함한 10개의 Classification Model에 대한 성능 비교를 위해 위처럼 Cross Validation을 사용하였습니다. 전 K-fold Cross Validation에서 K 값을 10으로 두어, 전체 데이터의 90%는 Training에, 10%는 Test에 사용되게 하였고, 각각의 모델을 평가하는 주요 척도로 Accuracy를 사용하였습니다.

뿐만 아니라 각 모델에서 나온 예측 값을 바탕으로 Precision, Recall, F1 score 역시 계산하였으며, 해당 결과를 Model의 Accuracy와 함께 출력되게 만들어, 해당 척도들 역시 모델을 평가할 때 사용될 수 있게끔 개발하였습니다.

Accuracy가 높은 순서로 각 모델의 이름과 Accuracy, Precision, Recall, F1 score에 대해 표로 정리하자면 다음과 같습니다. 해당 표에서 사용된 모델의 Hyper-parameter는 4. Classification Model에서 서술한 Hyper-parameter와 동일합니다. 또한 각 Model에서 보이는 Score 들은 소수점 아래 4자리에서 반올

림 한 수치를 사용하고 있습니다.

Model	Accuracy	Precision	Recall	F1 Score
Neural Network	0.747	0.752	0.75	0.751
Linear SVM	0.720	0.725	0.724	0.725
Random Forest	0.706	0.715	0.716	0.714
Decision Tree	0.695	0.698	0.698	0.698
OneR	0.689	0.691	0.690	0.690
Nearest Neighbors	0.682	0.691	0.681	0.682
Ada Boost	0.668	0.678	0.672	0.673
Naïve Bayes	0.619	0.694	0.621	0.602
ZeroR	0.552	-	-	-
RBF SVM	0.552	0.304	0.552	0.392

위 결과를 보면 Neural Network(Multi-layer Perceptron) 이 가장 좋은 성능을 보이고 있습니다. Neural Network의 경우 각 Perceptron 마다 위치해 있는 Weight 값을 이용하여 Target Variable을 예측하는 모델입니다. 해당 모델을 Feedforward 로 계산된 값과 실제 값의 에러 차이를 갖고, Back-propagation을 이용하여 각 Layer의 Perceptron 들을 학습하게 됩니다. 또한 수업시간에 배운 Linear SVM, Random Forest, Decision Tree 기법이 모두 Baseline으로 삼았던 OneR 모델에 비해 좋은 성능을 보였습니다.

반면 Nearest Neighbors, AdaBoost (Ensemble), Naïve Bayes 경우들에 대해서는 OneR 보다 좋은 성능을 보이고 있습니다. 각각의 이유에 대해서 생각해보자

면, Nearest Neighbors는 모든 Features의 거리를 동일한 Weights을 두고 계산한다는 점에서 중요한 Features를 기준으로 학습하지 못하는 한계점이 있습니다. 또한 Ensemble을 이용한 AdaBoost의 경우 전체 데이터가 116개 밖에 되지 않지만, Hyper-parameter 로 준 Number of estimators 값이 100이나 되는 큰 수이기에 잘못 분류된 Signal 뿐만 아니라 Noise까지 학습하여 좋지 못하는 성능을 보인다고 추측할 수 있습니다. 마지막으로 Naïve Bayes 역시 Nearest Neighbors와 비슷한 이유 때문에 중요한 Features를 기준으로 학습하지 못하는 한계점이 존재해 높은 성능을 보이지 못하는 것이라고 예상할 수 있습니다.

마지막으로 Baseline 으로 삼았던 ZeroR에 비해 RBF SVM은 더 좋지 못한 성능을 보이고 있습니다. 이는 Kernel Function 으로 사용되는 RBF가 exponential 한 의미를 담고 있지만, 내 Attributes 들은 exponential 한 의미를 담고 있지 않아 좋은 성능을 보이지 않는 것이라고 추측할 수 있습니다.

6. Result

각 환자들이 보이는 인체측정 및 혈액 데이터를 바탕으로 환자의 유방암 유무를 예측해보는 분류 모델을 다음과 같이 만들어 보았습니다. 학습을 위해 Dataset 내 총 116개의 데이터를 사용하였고, 각 Data 들은 Target variable을 포함해서, 총 10개의 Attributes 로 구성되어 있습니다.

전 성능 향상을 도모하기 위해 Numeric 하게 표현된 각 Features에 대해 Gaussian 분포를 취해 모두 Normalize 해주었습니다. 또한 Baseline 으로 사용한 ZeroR, OneR 모델 뿐만 아니라 수업시간에 다룬 모델을 포함하여 총 10개의 학습 모델을 사용하였습니다.

ZeroR을 제외한 각 모델들은 모두 10-Fold Cross Validation을 사용하여 평가하였으며, 평가 핵심 지표로 Accuracy를 두었고, Precision, Recall, F1 score 역시 함께 볼 수 있게끔 표로 정리하여 보였습니다.

정확도 기준, 10개의 모델들을 돌려 본 결과 나타난 가장 우수한 모델은 Neural Network를 이용한 모델입니다. 해당 모델의 정확도는 0.747 로 간단한 인체 측정 데이터와 혈액 데이터를 바탕으로 유방암을 적정하게 분류해주는 정확도를 보였습니다. 뿐만 아니라 해당 모델이 다른 9개의 모델보다 우수한 F1 score를 보이기도 했습니다.

해당 과제는 Numeric 하게 계산될 수 있는 Features 들을 중심으로 유방암 진

단 마커인, 분류 모델을 개발해 보았습니다. 하지만 최근 이미지 데이터를 활용하여 유방암의 유무를 확인하려는 task 들이 증가하고 있으며, 따라서 해당 과제도 단순한 수치 데이터 뿐만 아니라 Image 데이터를 함께 결합하여 유방암 진단 모델을 개발할 필요가 있습니다.

7. 기타

제가 개발한 모델의 데이터와 코드는 모두 제 개인 Github Repositories https://github.com/JinheonBaek/AI-Final-Project 에서 확인할 수 있습니다. 해당 Repositories의 경우 과제 제출 직전까지 Private 상태로 있어 외부인이 코드 등을 볼 수 없으며, 과제 제출과 동시에 공개될 것입니다.

데이터에 대한 설명과 기술은 모두 UCI Machine Learning Repository의 Data Set Description을 참고하였습니다. 뿐만 아니라 각 의료 데이터에 대해서는 Wikipedia에 기술된 정보와, 논문(https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2755091/)을 참조하였습니다.