**예측 모델 Homework #4**

**학번 : 2019021168 이름 : 조억**

**분류용 데이터와 예측용 데이터 하나씩 찾고 각 데이터에 대해 다음을 수행하시오.**

<분류용 데이터 = 유방암 데이터>

유방암 데이터는 양성과 악성을 나타내는 diagnosis 변수와 10개의 세포특징들에 대해 평균값과 표준편차, 그리고 이상값을 나타내는 3개 쌍의 값으로 총 30개의 설명 변수를 가지고 있습니다. 이 데이터를 고른 이유는 SVM이 이진 분류를 더 잘할 것 이기 때문에 얼마나 예측이 높은지 파악하게 위해 사용한 이유도 있습니다.

|  |  |
| --- | --- |
| id | ID number |
| Diagnosis | The diagnosis of breast tissues (M = malignant, B = benign) |
| Radius | Mean of distances from center to points on the perimeter |
| Texture | Standard deviation of gray-scale values |
| Perimeter |  |
| Area |  |
| Smoothness | Local variation in radius lengths |
| Compactness | Perimeter^2 / area - 1.0 |
| Concavity | Severity of concave portions of the contour |
| Concave points | Number of concave portions of the contour |
| Symmetry |  |
| Fractal\_dimension | “Coastline approximation" - 1 |

< 예측용 데이터셋 = 보스턴 집값 데이터>

보스턴 집값의 데이터의 경우는 13개의 설명변수와 하나의 타겟변수로 이루어진 데이터셋입니다.

|  |  |
| --- | --- |
| CRIM | 자치시(town) 별 1인당 범죄율 |
| ZN | 25,000 평방피트를 초과하는 거주지역의 비율 |
| INDUS | 비소매상업지역이 점유하고 있는 토지의 비율 |
| CHAS | 찰스강에 대한 더미변수(강의 경계에 위치한 경우는 1, 아니면 0) |
| NOX | 10ppm 당 농축 일산화질소 |
| RM | 주택 1가구당 평균 방의 개수 |
| AGE | 1940년 이전에 건축된 소유주택의 비율 |
| DIS | 5개의 보스턴 직업센터까지의 접근성 지수 |
| RAD | 방사형 도로까지의 접근성 지수 |
| TAX | 10,000 달러 당 재산세율 |
| PTRATIO | 자치시(town)별 학생/교사 비율 |
| B | 1000(Bk-0.63)^2, 여기서 Bk는 자치시별 흑인의 비율을 말함 |
| LSTAT | 모집단의 하위계층의 비율(%) |
| MEDV | 본인 소유의 주택가격(중앙값) (단위: $1,000) |

**(1) SVM (분류), SVR (예측) 모델을 하이퍼파라미터를 다양하게 바꾸어서 구축하시오 (kernel function, C, ...)**

분류를 위한 SVM의 SVC 모델의 커널은 linear, rbf, sigmoid 총 3개의 커널을 지정하였고 각각 설정 가능한 파라메터군이 다르기 때문에 아래와 같이 커널별 하이퍼 파라메터 후보군을 지정하였습니다.

|  |  |
| --- | --- |
| Kernel = Linear | 파라메터 후보군 |
| C | 0.1, 1, 100, 1000 |

|  |  |
| --- | --- |
| Kernel = rbf | 파라메터 후보군 |
| C | 0.1, 1, 100, 1000 |
| Gamma | 0.0001, 005, 0.1, 1, 5 |

|  |  |
| --- | --- |
| Kernel = sigmoid | 파라메터 후보군 |
| C | 0.1, 1, 100, 1000 |
| Gamma | 0.0001, 0.005, 0.1, 1, 5 |
| Coef0 | 0.1, 0.01, 0.0001 |

예측을 위한 SVR 모델의 경우는 마찬가지로 아래와 linear, rbf, sigmoid 총 3개의 커널을 지정하였고 모델을 구축하였다.

|  |  |
| --- | --- |
| Kernel = Linear | 파라메터 후보군 |
| C | 0.1, 1, 100, 1000 |
| epsilon | 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10 |

|  |  |
| --- | --- |
| Kernel = rbf | 파라메터 후보군 |
| C | 0.1, 1, 100, 1000 |
| Epsilon | 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10 |
| Gamma | 0.0001, 0.001, 0.005, 0.1, 1, 3, 5 |

|  |  |
| --- | --- |
| Kernel = sigmoid | 파라메터 후보군 |
| C | 0.1, 1, 100, 1000 |
| Epsilon | 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10 |
| degree | 2, 3, 4 |
| coef0 | 0.1, 0.01, 0.001, 0.0001 |

위와 같이 커널별 다양한 하이퍼파라메터의 조합으로 돌리기 위하여 그리드 서치 돌릴 수 있게끔 모델 구축 완료하였습니다.

**(2) Training set과 Testing set으로 나누고 다양한 평가 척도를 이용하여 예측성능을 평가하시오.**

Training과 Test 데이터셋은 70%과 30%으로 나누었고 평가 척도는 분류와 예측 각각의 성능을 아래와 같은 평가 척도를 사용하였습니다.

분류의 경우, Accuracy, F1\_score, Recall, Precision, ROC\_AUC 스코어를 각각 사용하였으며, 예측의 경우는 MAE, MSE, MSLE(mean\_squared\_log\_error), R2 스코어의 평가지표를 가지고 예측성능을 평가하였습니다.

위의 조합에서 가장 좋게 나온 성능 모델은 커널별로 봤을 때 예측성능은 아래와 같다.

<분류>

Linear 커널의 경우, C 상수가 100일 때 가장 좋게 나왔으며

|  |  |
| --- | --- |
| Accuracy: | 0.9707 |
| F1 Score | 0.9765 |
| Recall | 0.9629 |
| Precision | 0.9904 |
| ROC\_AUC | 0.9735 |

RBF 커널의 경우는 C 상수가 1일 때 그리고 Gamma가 0.0001일 때,

|  |  |
| --- | --- |
| Accuracy: | 0.9707 |
| F1 Score | 0.9771 |
| Recall | 0.9907 |
| Precision | 0.9639 |
| ROC\_AUC | 0.9636 |

마지막으로 Sigmoid 커널의 경우는 C가 1, coef0이 0.1, gamma가 0.0001일 때 가장 높게 나왔다.

|  |  |
| --- | --- |
| Accuracy: | 0.6315 |
| F1 Score | 0.7741 |
| Recall | 1.0 |
| Precision | 0.6315 |
| ROC\_AUC | 0.5 |

결과를 보면 Linear 커널과 RBF 모델의 best score가 유사함을 볼 수 있다.

<예측>

Linear 커널의 경우, C 상수가 1000, Epsilon값이 5일 때 가장 좋게 나왔으며

|  |  |
| --- | --- |
| MAE | 4.5561 |
| MSE | 34.9461 |
| MSLE | 0.0580 |
| R2 | 0.5661 |

Sigmoid 커널의 경우는 C 상수가 0.1, Coef0이 0.1, degree가 2, Epsilon이 10일 때,

|  |  |
| --- | --- |
| MAE | 6.8402 |
| MSE | 81.0575 |
| MSLE | 0.1348 |
| R2 | -0.0062 |

마지막으로 RBF커널의 경우는 C가 1000, Epsilon이 5, gamma가 0.001일 때 가장 높게 나왔다.

|  |  |
| --- | --- |
| MAE | 3.3170 |
| MSE | 19.7766 |
| MSLE | 0.0325 |
| R2 | 0.7545 |

예측의 경우에는 3가지 커널 중에 RBF 커널을 사용한 것이 가장 값이 잘나온 것을 눈으로 확인 할 수 있다.

**(3) 다른 분류모델과 비교하여 SVM만의 특징을 기술하시오.**

SVM과 비교할 분류 모델은 Random Forest(이하 RF)를 사용하였습니다. 마진을 최대화하여 초평면으로 데이터를 구분하는 SVM의 특성상 이진 분류를 잘 할 것 같아서 고른 데이터셋을 가지고 실험한 결과 확실히 아래와 같이 Random Forest모델에서 같은 데이터를 가지고 분류기를 늘려봐도 가장 많은 125개의 트리를 사용했을 때 96.24%를 기록하고 쉽게 우리가 SVM을 통해서 얻을 수 있는 97%이상의 정확도를 얻기는 어려웠다.

|  |  |
| --- | --- |
| Count of Estimators | Accuracy |
| 25 | 0.9342 |
| 50 | 0.9413 |
| 75 | 0.9530 |
| 100 | 0.95539 |
| 125 | 0.9624 |
| 150 | 0.9530 |

보시는 것과 같이, SVM은 태생적으로 2개의 클래스를 구분하는데 특화되어 있다고 볼 수 있습니다. RF의 경우 수치나 분류형 데이터가 혼재 되어 있는 데이터셋도 그대로 사용할 수 있는 점이 있는 반면, SVM의 경우에는 최소/최대화 값을 각각 구해서 Normalization을 해줘야 하며 범주형 데이터의 경우는 원핫 인코딩을 무조건 해줘야 한다. Sparse Model이라고 도 불리는 이유가 있다. 이것은 마진을 최대하는 여백을 만드는데 사용하는 decision boundary를 만드는게 전체 데이터를 쓰지 않기 때문이다. SVM의 경우는 m개 피쳐를 가진 n개의 관측치들이 있을 경우는 n x n 의 복잡도를 가진다. 그래서 관측치가 많을수록 오랜 학습시간이 필요하며, 일정 크기 이상으로 스케일을 키우기가 어렵다고 볼 수 있다. 하지만 n x n의 복잡도를 가지는 이상 많은 피쳐들이 있다고 문제가 되지는 않는다. RF의 경우는 각 범주의 확률을 주는 반면 SVM의 경우에는 거리를 결과값으로 준다. 일반적으로 RF보다 SVM의 성능이 좋다고 알려져 있다. 위의 데이터도 그걸 설명하고 있다. 또한 해석력에서 강점을 가지고 있다고 볼 수 있다. 즉 support vector에 대한 값을 확인함으로써 결과에 대한 해석에 도움이 될 수 있다. 즉 초평면을 지지하는 관측치들의 데이터를 가지고 분류기준에 대한 해석이 가능할 수 있다. 그리고 이런 서포트 벡터들을 만드는 데이터는 다다른 어떤 모델보다 Sparse Model이라고 불리는 SVM만의 특징을 강조할 수 있으며, 이것의 장점은 Sparse Data데이터에서도 더 좋은 성능을 낼 수 있다고 함. 예를 들면 문서 분류의 경우, 선형적으로 데이터가 분리가 안되는 경우, SVM에서는 RBF라는 커널을 사용하여 분류가 가능하며, 고차원의 데이터도 잘 분류된다고 한다.

그리고 회귀 모형(Linear Regression)의 경우에도 SVR하고 비교를 해보았습니다.

|  |  |
| --- | --- |
| MAE | 4.5656 |
| MSE | 34.5823 |
| MSLE | 0.0684 |
| R2 | 0.5707 |

위의 데이터를 보면, 데이터 자체가 보스턴 집값으로 500개 정도의 데이터셋을 가지고 있어서 확실히 많은 데이터는 아니지만, linear regression의 경우 Y 절편을 사용하는 경우와 안 사용하는 경우 중 더 잘 나온 이 경우와 비교를 해봐도 위의 MAE가 3.3170, MSE가 19.7766으로 나와서 회귀모형보다 SVR이 더 잘 나오는 결과를 얻을 수 있어서 SVM은 예측과 분류가 모두 잘 되는 결과를 얻었고 결론적으로 강력한 모델이라는 생각에 이르렀습니다.