**예측 모델 Homework #4**

**학번 : 2019021168 이름 : 조억**

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

분류용 데이터로는 유방암 데이터를 사용하였고 예측용 데이터로는 보스턴 집값 데이터를 사용하였습니다.

유방암 데이터

|  |  |
| --- | --- |
| id | ID number |
| diagnosis | The diagnosis of breast tissues (M = malignant, B = benign) |
| radius\_mean | mean of distances from center to points on the perimeter |
| texture\_mean | standard deviation of gray-scale values |
| perimeter\_mean | mean size of the core tumor |
| area\_mean |  |
| smoothness\_mean | mean of local variation in radius lengths |
| compactness\_mean | mean of perimeter^2 / area - 1.0 |
| concavity\_mean | mean of severity of concave portions of the contour |
| concave points\_mean | mean for number of concave portions of the contour |
| symmetry\_mean |  |
| fractal\_dimension\_mean | mean for "coastline approximation" - 1 |
| radius\_se | standard error for the mean of distances from center to points on the perimeter |
| texture\_se | standard error for standard deviation of gray-scale values |
| perimeter\_se |  |
| area\_se |  |
| smoothness\_se | standard error for local variation in radius lengths |
| compactness\_se | standard error for perimeter^2 / area - 1.0 |
| concavity\_se | standard error for severity of concave portions of the contour |
| concave points\_se | standard error for number of concave portions of the contour |
| symmetry\_se |  |
| fractal\_dimension\_se | standard error for "coastline approximation" - 1 |
| radius\_worst | "worst" or largest mean value for mean of distances from center to points on the perimeter |
| texture\_worst | "worst" or largest mean value for standard deviation of gray-scale values |
| perimeter\_worst |  |
| area\_worst |  |
| smoothness\_worst | "worst" or largest mean value for local variation in radius lengths |
| compactness\_worst | "worst" or largest mean value for perimeter^2 / area - 1.0 |
| concavity\_worst | "worst" or largest mean value for severity of concave portions of the contour |
| concave points\_worst | "worst" or largest mean value for number of concave portions of the contour |
| symmetry\_worst |  |
| fractal\_dimension\_worst | "worst" or largest mean value for "coastline approximation" - 1 |

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, ...)**

ㅇㅇㅇ

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

ㅇㅇㅇ

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

SVM과 비교할 분류모델은 Random Forest, KNN, Naïve bayes를 사용하였습니다.

SVM을 사용하는 큰 이유는

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.

SVM의 경우는 m개의 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. 그리고 support vector에 대한 값을 확인함으로써 해석에 도움이 될 수 있다. 즉 초평면을 지지하는 관측치들의 데이터를 가지고 분류기준에 대한 해석이 가능할 수 있다. 그리고 의사결정 나무 모델보다 Sparse Data에 대해서 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 분류한다합니다.

kNN의 경우는 데이터가 많을 경우 노이즈가 많은 트레이닝 데이터에 완고한 특징을 가지나 K라는 파라메터를 결정해야 하는 번거로움이 있다. 훈련시간이 오래 걸리는 단점이 있음.

Random Forest는 의사결정나무를 여러 개 사용