**Machine Learning & Data Mining, Spring 2020**

**Homework 5**

Due April 30

1. Suppose we developed a disease screening system and it produced the following results on the test samples (Class 1 means a positive test result).

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Predicted class | |
|  |  | 1 | 0 |
| Actual class | 1 | 10 | 190 |
| 0 | 40 | 760 |

* 1. Compute Accuracy, Sensitivity, Specificity, Recall and Precision of this system with respect to the class 1.

accuracy = 10 + 760 / (10 + 190 + 40 + 760) = 770/1000 = 0.77

sensitivity=10/(10+190)=10/200=0.05

specificity=760/(40+760)=0.95

recall=0.05(equal to sensitivity)

precision=10/(10+40)=0.2

* 1. By using the result in (a), briefly explain why Accuracy alone is not enough to show real performance of this system.

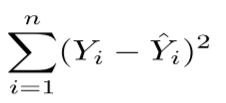
class 0은 200개가 있고, class 1은 800개가 있습니다.

이는 class imbalance상황이고, 항상 0으로 예측하는 예측기는 80%의 성능을 보이기에 정확도만으로 보기에는 잘못된 예측기도 높은 성능을 갖는 것처럼 보입니다. 이것이 정확도만으로 측정하는 것의 문제입니다.

* 1. Suppose that the cost of missing a real patient is really high, as in the case of an epidemic disease. Discuss whether the above system is suitable for use in such cases.

놓치면 가장 큰 문제가 되기 때문에 위 문제의 경우 recall-precision pair가 중요한 성능지표로써 사용되야 하는 것이 맞다고 판단됩니다. 왜냐하면, recall의 경우 실제 병에 걸린 환자를 정확히 판단하는 척도로써, 낮다면, 실제 병에 걸린 환자를 놓칠 가능성이 높은 것입니다. 왜냐하면, 실제병에 걸렸음에도 병에 걸리지 않았다고 판단하는 FN이 포함되는 지표이기 때문입니다. precision의 경우 병에 걸렸다고 예측했지만, 실제로 병에 걸리지 않은 환자들도 포함되니까(FP), 이 지표가 낮다면, 병에 걸렸다고 판단하는 성능이 좋지 않음을 의미합니다. 1번 문제의 경우 precision과 recall이 낮기 때문에 병을 놓치는 것이 cost가 큰 상황에서는 알고리즘을 바꿔야 할 것으로 판단됩니다.

1. Suppose you are given a training dataset of n points to fit a regression model.
2. In a simple linear regression, explain how optimal parameters are defined. You don’t need to derive the equations, but explain what property makes the parameters optimal.

실제 y의 값은 예측이 거의 불가능 하기 때문에, 예측 y값과 오차가 날 수밖에 없는데 이것을 에러라고 표현합니다 실제 y값과 예측을 통해 나온 y값의 차이 즉, 에러 제곱의 합을 SSE(Sum of squared error)라고 하는데 이 값을 최소로 하는 을 학습해 내는 것이 simple regression의 목표입니다. 최적화 값은 (SSE)라는 2차식의 미분계수가 0인 지점에서의 coefficient값 입니다.

1. Discuss whether it is necessary to use Ridge or Lasso regression in this case.

Ridge는 regularize의 한 방법으로 feature selection에 쓰이는데, 람다라는 하이퍼 파라미터와 함께 두어 각 coefficient의 벡터의 크기를 L2 norm으로 측정한 penalty를 두어 각 계수들이 튀는 것을 방지합니다, 데이터의 분산이 크다는 것은 해당 계수가 불안정하다는 뜻인데, 계수는 기울기를 의미하고 기울기는 변화량을 의미하므로 변화량이 왔다 갔다 한다는 뜻으로도 해석될 수 있기 때문입니다. 결론적으로 안정적이지 않다는 것으로 간주됩니다.

이러한 경우 각 계수들을 0에가깝게 보내서 대응되는 feature들의 크기에 중점을 두게 되어(계수는 모두 비슷비슷 하므로) feature의 중요도를 파악하는 방법입니다. 계수가 0으로 수렴은 하지만 0이 되지는 않아서 최종적으로 형성되는 모델은 모든 feature를 포함하게 됩니다. 때문에 해석에 어려움을 겪습니다. 하지만 LASSO는 Ridge와 비슷한 방식이나, penalty term을 달리하여, 필요 없다고 간주되는 feature의 계수를 정확히 0이 되게 만들어서 variable selection의 효과와 regression을 동시에 가능하게 합니다.

결론적으로 Ridge는 모든 feature를 고려해야 하는 상황에서 데이터의 variance가 높을 때 적합하고, LASSO는 전체 dataset에서 중요시되는 feature를 선택해서 학습하여 최적의 효과를 보고싶을 때 사용하면 좋을 듯 싶습니다.