**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 = = 0.77

Sensitivity = = 0.05

Specificity = 0.95

Recall =

Precision = = 0.2

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

Accuracy는 0.77로써, 꽤 높은 정확도를 보인다고도 할 수 있다. 하지만, 실제 예측된 결과와 실제 class를 비교해보면, class 1로 예측되는 경우가 굉장히 적으며 대부분의 데이터가 class 0으로 나타난다. 그런데, 정확도가 높게 나오는 이유는 test sample의 80%의 data가 class 0인 data이기 때문이다. 즉, sample data set에 따라서 정확도는 달라질 수 있으므로, accuracy만으로 system의 수행을 판단해서는 안되고 recall, precision, sensitivity, specificity같은 결과를 정확도와 함께 보면서 system의 수행이 적합한지 판단해야 한다.

* 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.

실제 전염병이 걸린 환자의 경우를 class 1이라고 할 때, 위의 system은 굉장히 부적합한 system이라고 할 수 있다. 위의 system을 쓰게 되면 전염병인 환자를 아닌 환자로 잘 못 오인하는 경우가 더욱 더 발생될 것이다. 그렇다면 현재 전 세계의 이슈인 코로나바이러스처럼 심각한 문제가 발생할 수 있다. 이 경우에는 오히려 class 1이 class 0보다 더욱 잘 잡히는 (class 1이 되는 경우가 많은) system을 써야한다. 물론 위의 system 보다 좋다는 말이며, class 0가 더 많이 잡힌다고 좋은 경우는 아니다. 즉, 위의 system은 이 case에 굉장히 부적합하다.

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.

() = 이 식을 통해서 Bo는 Y 절편이고, B1은 기울기이며, yi는 실제 y값이다. 그리고 (Bo + B1xi)는 예측된 y값이다. 이때, simple linear regression의 optimal parameter는 모든 training data의 실제 y값과 예측된 y값의 차이의 제곱을 최소로 하는 Bo 와 B1을 찾는 것이다. 즉, 최적의 parameter B0, B1을 구하는 방법은 위의 2차방정식을 B0와 B1으로 각각 미분한 후, 미분이 0이되는 B0, B1을 찾아서 구할 수 있다.

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

Ridge or Lasso를 쓰게 되면, () = 이 식에다가 추가적인 penalty 값을 추가하게 된다. 이 때, Ridge는 L2 norm을 더하게 되고 Lasso는 L1 norm을 더하게 된다. 그러면 전체적인 식이 () = 이와 같이 되고  각각의 식이 이렇게 된다. 이렇게 되면, penalty와 실제값과 예측값의 차이의 제곱 모두 낮춰야 하는데, 그렇게 되면 error를 낮추기 위해서 coefficient 값들 중에 튀는 값들이 있어서는 안되고, 0근처로 골고루 작게 나와야 한다. 그렇게 되면 coefficient 들이 regularization이 되야 한다. 그래서 그러한 특징을 가진 coefficient 중에 error를 가장 작게 하는 것을 찾게 된다.

하지만 위와 같은 경우는, X variable의 수가 많은 경우에 이점이 있다.

즉, training dataset이 이와 같이 주어지면, 정규화가 의미가 없어진다. 왜냐하면, 각 y값에 대응되는 x 가 하나이기 때문이다. 즉, simple linear reggrssion 모델의 경우에는 ridge와 Lasso를 못 쓰는 것은 아니지만, 필수적인 것은 아니다.