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

**Homework 5**

Due April 30

**이름: 김도운**

**학과: 사이버보안학과**

**학번: 201520893**

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:
* Sensitivity:
* Specificity:
* Recall:
* Precision:

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

* 전체적인 정확도는 77%로 높게 나왔지만, 클래스1과 클래스0의 크기의 차이가 크고, 클래스 1에 대해서는 제대로 예측하지 못했으며, 사이즈가 큰 클래스 0에 대해서만 거의 완벽하게 분석했기 때문에 전체 성능이 77%로 측정된 것이므로, 이는 왜곡된 결과라고 할 수 있다. 따라서 accuracy만을 보고 실제 성능을 정확하게 표현했다고는 할 수 없다.

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 환자를 놓치는 경우의 cost가 매우 높기 때문에 이 system은 절대로 적합하지 않다. 그 이유로, 실제로 음성인 환자들에 대해서는 음성으로 판정해 정확성을 보이지만, 실제 양성인 환자의 대부분을 음성으로 판정했기 때문에 cost측면에서도 엄청난 손실이 발생하며, 정확성도 좋지 않기 때문에 이러한 시스템은 적합하지 않다.
* 예를 들어 현재 코로나 바이러스와 같은 경우에도, 실제 감염된 사람임에도 불구하고 감염되지 않았다고 판정하는 경우, 해당 감염자는 격리되지 않기 때문에, 다른 사람까지도 감염될 수 있다.

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.**

* 단순 선형 회귀에서는 예측 값과 실제 값의 차이의 제곱의 합, 즉 오차의 제곱의 합을 최소화하는 것이 목표이며, 이 오차 제곱의 합을 최소화하는 coefficient값(simple linear regression 에서는)이 최적의 파라미터가 된다. 이러한 최적의 coefficient 값들을 구하기 위해서는 SSE(Sum of Squared Error)가 최소가 되는 값들로 선정하면 되는데, 이는 각 coefficient값들로 미분을 했을 때 각 항이 0이 되도록 하는 값이 된다.

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

* Ridge나 Lasso등의 shrinkage method들은 Y값을 예측할 때 필요한 X변수들이 많은 경우, 모든 변수들을 적용시키면 model이 너무 커지므로(즉, complexity가 커짐) overfit의 가능성이 있기 때문에, 예측하는데 영향이 적은 변수들의 값을 0에 수렴시키거나(Ridge) 0으로 만들어 버림(Lasso)으로써, 해당 모델의 성능을 향상시키고, 복잡성을 감소시킴으로써 모델의 가독성도 높아진다. 예를 들어 어떤 data의 sample수와 feature수보다 적거나 비슷할 때, 어떤 penalty도 적용하지 않은 OLS(Ordinary Least Squares)를 사용하면 측정된 값이 불안정하고 variance도 높은 경향이 있게 되는데, 이와 같은 경우 L2 penalty를 더해줌으로써(Ridge regression) coefficient값들을 0에 수렴시키며, bias측면에서는 약간의 손해를 보더라도, variance를 굉장히 줄여주므로 성능이 향상하게 된다. 다만 Ridge regression의 경우 coefficient값들을 0에 수렴시킬 뿐 0으로 만드는 것이 아니므로 모든 feature를 고려해야하기 때문에 해석하기에 힘든 점이 있지만, L1 penalty를 더해주는 Lasso regression의 경우 불필요한 coefficient값들을 완전 0으로 만듦으로써 Ridge 모델보다 더 간단하고 강력한 성능으로 향상시킬 수 있다.
* 하지만 위와 같이 simple linear regression의 경우 줄어드는 값은 이므로, 여기서 y-intercept값인 는 줄어들지 않는다. 따라서 Ridge나 Lasso를 적용한다면 값으로만 예측하고, 해석하게 되는데, 이는 거의 무의미한 행위이므로 simple linear regression에서는 굳이 사용할 필요가 없을 것 같다.