기계학습 3주차 과제

12191656 이채연

HomeWorks #1

머신러닝에서 Overfitting 문제에 대해 설명하고, 이를 완화하기 위한 방법을 제시하시오.

Overfitting이라는 것은 fitting이 지나치게 많이(over) 됐다는 것이다.
이러한 overfitting이 발생이 되는 경우, input training data에 대해서 모든 minor variation을 modeling하기 위해서 model을 학습했기 때문에 이런 문제가 발생이 된다. Minor한 작은 변화까지 다 catch해서 fitting을 하게 되면 그러한 모델은 training data에서는 error rate이 굉장히 낮지만, 실제 test환경에서 이 모델을 가지고 성능을 측정해보면(test 해보면) 성능이 그렇게 좋지 않다. 이러한 이유는 minor variation이 noise일 가능성이 높은데, true signal보다 noise에 대해서 noise pattern까지도 다 학습을 하려고 했기 때문에 이러한 문제가 발생이 된다.
그리고 high degree polynomial result가 이런 경우에 extreme한 oscillation이 발생이 된다. 그러한 이유는 polynomial regression을 할 때 degree가 높으면 높을수록 function의 oscillation이 굉장히 많다. 그렇게 때문에 high degree polynomial, 즉 지나치게 복잡한 모델을 사용할 경우에 이러한 현상이 발생이 된다. 그래서 그 결과로 future output에 대해서 inaccurate한 prediction이 발생이 된다.
따라서 overfitting을 완화하기 위한 방법 중 하나는 모델을 simple하게 만드는 것이다. 예를 들어 polynomial regression을 할 때는 degree를 낮추면서 test error rate이 낮아지는 적절한 degree을 찾아 degree값을 바꾸면 overfitting 문제를 완화할 수 있다. 또 다른 예로 non-parametric 모델인 KNN classifier를 사용할 때 k값에 따라서 모델의 complexity가 다르다. K가 작을 때 모델의 complexity가 높아져 overfitting이 발생되는 경우가 많다. 따라서 이럴 때 K값을 높여 모델을 단순하게 만들어야 한다. K값을 높이면서 test error rate이 낮아지는 적절한 K값으로 바꾸면 overfitting문제를 완화할 수 있다.

HomeWorks #2

Conditional independence(조건부 독립)에 대해 설명하시오. p(X|Y,Z) = p(X|Z)임을 증명하시오.

Conditional independence(조건부 독립)은 어떤 조건 하에 X와 Y가 independent하다는 뜻이다. 즉, 어떤 다른 random variable Z라는 event가 주어졌을 때 X와 Y는 independent 하다는 것이다.

Conditional independence의 의미는 X라는 random variable의 확률 값을 추론하는데 Y는 알 필요가 없고 Z만 알면 된다는 것이다. 왜냐하면 Z라는 event가 발생이 되면 Y가 당연히 발생이 되기 때문에 굳이 다른 variable을 참고해서 X값을 구할 필요가 없다. 따라서 $p(X|Y,Z)$는 Z가 발생되면 Y는 알 필요가 없으므로(당연히 발생이 되므로), $p(X|Z)$와 같다. 즉 $p(X|Y,Z) = p(X|Z)$이다.