**Decision Tree**

Decision Tree는 어떤 데이터의 feature를 보고 단계별로 유익한 feature를 확인하면서 최종 단계에 결정을 내리는 모델이다.

**Construct Procedure**

Decision Tree는 Select, Split, Repeat, Stop과정을 거쳐서 만들어진다.

**Select**에는 가장 유익한 feature를 고르고, **Split**에는 feature이 어떤 상태이냐에 따라 instance들을 여러 집합으로 나눈다. **Repeat**에서는 위 과정을 반복하고, 집합의 instance가 모두 같은 class가 되면 **Stop**을 한다.

**Aim**

우리는 이러한 Decision Tree를 같은 성능이면서 가장 작게 만들어야 한다.

이를 위해서, Select 단계에서 ‘**가장 유익한 feature**’를 잘 골라내야 하고, 이를 ‘**information gain**’을 측정하여 적용함으로써 해결할 수 있다.

**Information Gain**

텍스트, 손목시계, 시계이(가) 표시된 사진

자동 생성된 설명

텍스트, 시계, 손목시계이(가) 표시된 사진

자동 생성된 설명



**Example**

테이블이(가) 표시된 사진

자동 생성된 설명



이 때,

텍스트, 손목시계, 시계이(가) 표시된 사진

자동 생성된 설명

를 통해,

,

텍스트, 시계, 손목시계이(가) 표시된 사진

자동 생성된 설명

를 통해,

텍스트이(가) 표시된 사진

자동 생성된 설명



를 통해,



age 외에, 나머지 feature에 대해서도 gain을 구하면,

텍스트이(가) 표시된 사진

자동 생성된 설명

Gain이 가장 높은 age feature를 Select하여 Split 진행하면 된다.

feature가 연속적인 값을 가지고 있다면, 값들의 중간값을 중심으로 Gain 구하는 방식이 쓰인다.

Information Gain외에 Gini Impurity를 사용하여 Split을 시행할 수도 있다.

**테이블이(가) 표시된 사진

자동 생성된 설명**

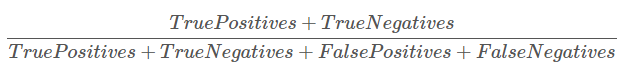
True, False는 분류 결과 맞았는지 틀렸는지를 나타내고,

Positive, Negative는 어떤 현상이 일어났는지 안 일어났는지(실제 정답)를 나타낸다.

(ex. 눈이 왔다 : Positive, 눈이 오지 않았다 : Negative)

**Accuracy(정확도)**

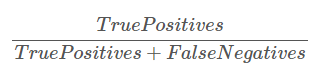
정확도는 모든 테스트에서 올바르게 예측된 데이터에 대한 비율을 말한다.



**Recall (재현율)**

날씨를 생각해보자. 눈이 내리는 날은 그렇지 않은 날에 비해 매우 적다. 그렇다고 해서 정확도(Accuracy)를 높이고자 모든 날의 날씨를 눈이 내리지 않는다고 판단하면 안 될 것이다.

이때, Recall (재현율)이 도움을 줄 수 있다. Recall을 구하는 공식은 다음과 같다.



분모는 눈이 내렸다고 예측한 날로,

* 눈이 내린날(Positive) 눈이 내린다고 예측(True)
* 눈이 내리지 않은 날(Negative), 눈이 내렸다고 예측(False)

2 가지가 있다.

눈이 내렸다고 예측한 날 중 실제로 눈이 내린 날의 비율을 보면 Accuracy의 단점을 보완할 수 있다.

“역방향 느낌” (예측 -> 예측 결과를 돌아봄)

**Precision (정밀도)**

이번에는 위의 날씨 예에서 모든 날을 눈이 내린다고 판단했다고 해보자. Accuracy는 낮더라도 Recall은 높게 가져가게 될 것이다.

이때, Precision (정밀도)가 도움을 줄 수 있다. Precision 공식을 먼저 확인해보자.

텍스트이(가) 표시된 사진

자동 생성된 설명

분모는 실제로 눈이 내린 날로,

* 눈이 내린 날(Positive), 눈이 내렸다고 예측(True)
* 눈이 내린 날(Positive), 눈이 내리지 않았다고 예측(False)

2가지가 있다.

눈이 실제로 내린 날 중, 눈이 내렸다고 예측한 비율을 보면,

모델이 눈이 내리는 날에 대하여 예측하는 능력을 확인할 수 있다.

“정방향 느낌” (눈이 내림 -> 예측)

**Note**

Recall(재현율)과 Precision(정밀도)를 보면, 둘은 서로 trade-off(하나가 증가하면 다른 하나는 감소한다는 뜻)되는 관계가 있다는 것을 알 수 있다.

**F1 Score**

F1 Score은 precision과 recall의 조화평균이다. 구하는 식은 아래와 같다.

텍스트이(가) 표시된 사진

자동 생성된 설명

이를 통해 precision과 recall을 조합하여 하나의 통계치를 얻을 수 있다. 1에 가까울수록 좋은 모델이며, 둘 중 하나의 값이 0과 가까워지면 그냥 평균보다 더욱 나쁜 값을 갖게 한다.