**인공지능 HW3** -이수원 교수님-

# 20150439 소혜빈

제출 날짜 : 2019.5.19

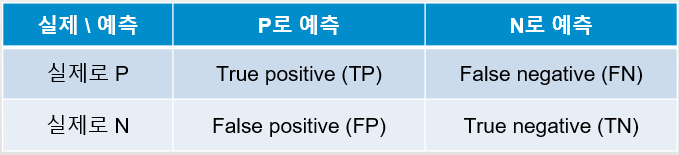
환경: Python3.7.3, Pycharm, 및 sklearn

## 문제

1. 기계학습 모델과 파라미터 별 분류 정확도 비교표 (소수 둘째 자리까지)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Parameter 1** | **Parameter 2** | **Precision** | **Recall** | **F1** |
| Decision Tree | criterion  =‘entropy’ | min\_impurity\_split=0.1 | 0.88,0.77 | 0.88,0.77 | 0.88,0.77 |
| min\_impurity\_split=0.2 | 0.87,0.85 | 0.93,0.73 | 0.90,0.73 |
| min\_impurity\_split=0.3 | 0.89,0.85 | 0.93,0.77 | 0.91,0.81 |
| min\_impurity\_split=0.4 | 0.89,0.85 | 0.93,0.77 | 0.91,0.81 |
| min\_impurity\_split=0.5 | 0.90,0.86 | 0.93,0.80 | 0.92,0.83 |
| criterion  =‘gini’ | min\_impurity\_split=0.1 | 0.90,0.86 | 0.93,0.80 | 0.92,0.83 |
| min\_impurity\_split=0.2 | 0.91,0.83 | 0.91,0.83 | 0.91,0.83 |
| min\_impurity\_split=0.3 | 0.91,0.83 | 0.91,0.83 | 0.91,0.83 |
| min\_impurity\_split=0.4 | 0.95,0.84 | 0.91,0.90 | 0.93,0.87 |
| min\_impurity\_split=0.5 | 0.66,0.00 | 1.00,0.00 | 0.79,0.00 |
| K Nearest Neighbor | weights  =‘uniform’ | k=1 | 0.90,0.80 | 0.90,0.80 | 0.90,0.80 |
| k=5 | 0.91,0.76 | 0.86,0.83 | 0.88,0.79 |
| k=10 | 0.88,0.74 | 0.86,0.77 | 0.87,0.75 |
| k=20 | 0.88,0.79 | 0.90,0.77 | 0.89,0.78 |
| k=30 | 0.88,0.77 | 0.88,0.77 | 0.88,0.77 |
| weights  =‘distance’ | k=1 | 0.90,0.80 | 0.90,0.80 | 0.90,0.80 |
| k=5 | 0.91,0.78 | 0.88,0.83 | 0.89,0.81 |
| k=10 | 0.89,0.75 | 0.86,0.80 | 0.88,0.77 |
| k=20 | 0.88,0.77 | 0.88,0.77 | 0.88,0.77 |
| k=30 | 0.88,0.77 | 0.88,0.77 | 0.88,0.77 |

1. ①의 결과를 해석한다면 ?



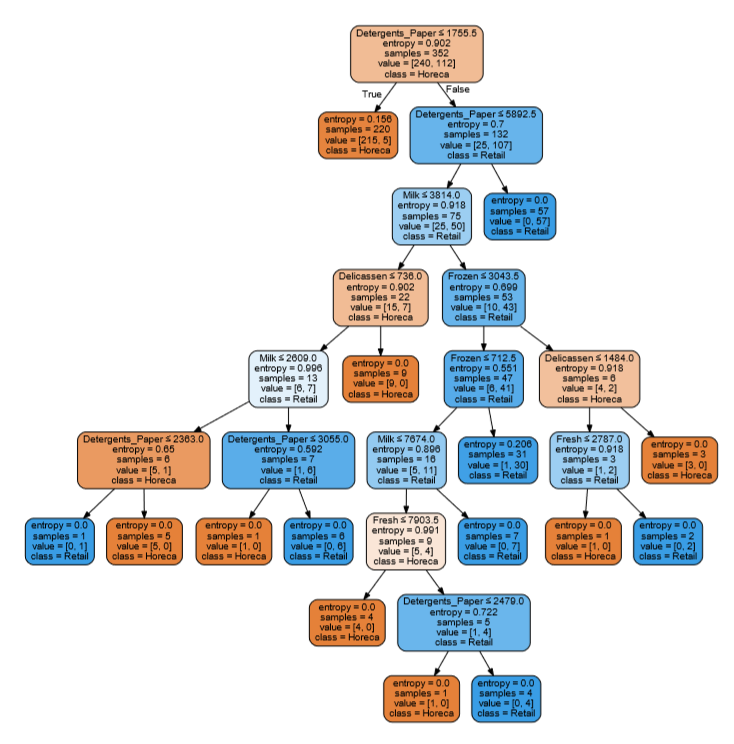
[그림1] 컨퓨전 매트릭스 (출처:이수원 교수님 11강 강의 보조자료 )

①의 표의 수치는 Decision Tree와 K Nearest Neighbor 알고리즘을 이용해 분류한 Horeca(1) ,Retail(2) 각 클래스 별 precision, recall, f1 측정 결과값이다. precision은 예측의 정밀도로, TP/(TP+FP)로 계산한다. recall은 관심대상을 얼마나 잘 찾아내는가에 대한 평가값으로 TP/(RP+FN) 으로 계산한다. f1은 recall과 precision 두가지 지표를 동시에 반영한 지표로써, 두 값의 조화평균 값이다.

Decision tree 를 이용해 train할 경우, 분할 기준을 criterion 파라미터에 ‘entropy’ 또는 ‘gini’ string값을 주어 변경할 수 있다. ‘entropy’일 때, information gain을 이용해 결정에 있어서 중요한 property를 결정하게 되고 자식노드의 impurity가 낮을수록 information gain의 값이 커지며 값이 클수록 중요한 property이다. ‘gini’ 일 때는 gini impurity 를 이용해 중요한 property를 결정하게 되는데, misclassification의 확률을 최소화하는 것이 목표이다. 특히, gini 를 적용했을 때, min\_impurity\_split을 0.5 로 설정하면, Retail class의 precision, recall, f1 측정 결과값이 모두 0이 나오는 것을 확인 할 수 있다.

K Nearest Neighbor 를 이용해 train 할 경우, criterion 파라미터에 ‘uniform’ 또는 ‘distance’ string 값을 주어 가까운 데이터에 가중치를 주는 방법을 변경할 수 있고, ‘k값으로 얼마를 주느냐에 따라 근접 몇 개의 데이터를 기준으로 분류할 것인지 정할 수 있다. ‘uniform’ 방식의 경우 거리에 관계없이 모두 동일한 가중치를 갖게 되며, ‘distance’ 방식의 경우 1/distance, 즉 거리에 반비례 하는 값을 가중치로 갖게 된다.

1. data X의 속성 중 가장 중요한 속성은 ? (Decision Tree 활용)



[그림 2 ] ‘entropy’, min\_impurity\_split(임계값) = 0.3일 때 생성된 decision tree

그림2를 보면 생성된 root노드의 property가 Detergents\_Paper 임을 확인할 수 있다. 그러므로 data X의 속성 중 가장 중요한 속성은 Detergents\_Paper이다.

1. ③의 결과를 해석한다면 ?

Detergents\_Paper 의 수치가 1755.5 이하이면 Horeca로 분류하고 그렇지 않으면, 다시 Detergents\_Paper 의 수치가 5892.5 이하인지 검사한다. 거짓이면, Retail로 분류하고, 참이라면, Milk의 수치가 3814.0 이하인지 검사한다. 다시, 참이면 Delicassen의 수치가 736.0 이하인지 검사하고, 거짓이면 Frozen의 수치가 3043.5이하인지 검사한다. 이런 식으로 각 서브 노드에서 가장 중요하다고 판단되는 속성을 기준으로 검사해, entropy 값이 임계값(위 그래프의 경우 0.3)보다 작아질 때 까지 반복하며 leaf 노드에서 class 를 예측한다.

## Run

>>>

criterion: entropy

impurity split threshold: 0.1

Precision\_score: 0.88,0.77

Recall\_score:0.88,0.77

F1\_score:0.88,0.77

criterion: entropy

impurity split threshold: 0.2

Precision\_score: 0.87,0.85

Recall\_score:0.93,0.73

F1\_score:0.90,0.79

criterion: entropy

impurity split threshold: 0.3

Precision\_score: 0.89,0.85

Recall\_score:0.93,0.77

F1\_score:0.91,0.81

criterion: entropy

impurity split threshold: 0.4

Precision\_score: 0.89,0.85

Recall\_score:0.93,0.77

F1\_score:0.91,0.81

criterion: entropy

impurity split threshold: 0.5

Precision\_score: 0.90,0.86

Recall\_score:0.93,0.80

F1\_score:0.92,0.83

criterion: gini

impurity split threshold: 0.1

Precision\_score: 0.90,0.86

Recall\_score:0.93,0.80

F1\_score:0.92,0.83

criterion: gini

impurity split threshold: 0.2

Precision\_score: 0.91,0.83

Recall\_score:0.91,0.83

F1\_score:0.91,0.83

criterion: gini

impurity split threshold: 0.3

Precision\_score: 0.91,0.83

Recall\_score:0.91,0.83

F1\_score:0.91,0.83

criterion: gini

impurity split threshold: 0.4

Precision\_score: 0.95,0.84

Recall\_score:0.91,0.90

F1\_score:0.93,0.87

criterion: gini

impurity split threshold: 0.5

Precision\_score: 0.66,0.00

Recall\_score:1.00,0.00

F1\_score:0.79,0.00

weights: uniform

k: 1

Precision\_score: 0.90,0.80

Recall\_score:0.90,0.80

F1\_score:0.90,0.80

weights: uniform

k: 5

Precision\_score: 0.91,0.76

Recall\_score:0.86,0.83

F1\_score:0.88,0.79

weights: uniform

k: 10

Precision\_score: 0.88,0.74

Recall\_score:0.86,0.77

F1\_score:0.87,0.75

weights: uniform

k: 20

Precision\_score: 0.88,0.79

Recall\_score:0.90,0.77

F1\_score:0.89,0.78

weights: uniform

k: 30

Precision\_score: 0.88,0.77

Recall\_score:0.88,0.77

F1\_score:0.88,0.77

weights: distance

k: 1

Precision\_score: 0.90,0.80

Recall\_score:0.90,0.80

F1\_score:0.90,0.80

weights: distance

k: 5

Precision\_score: 0.91,0.78

Recall\_score:0.88,0.83

F1\_score:0.89,0.81

weights: distance

k: 10

Precision\_score: 0.89,0.75

Recall\_score:0.86,0.80

F1\_score:0.88,0.77

weights: distance

k: 20

Precision\_score: 0.88,0.77

Recall\_score:0.88,0.77

F1\_score:0.88,0.77

weights: distance

k: 30

Precision\_score: 0.88,0.77

Recall\_score:0.88,0.77

F1\_score:0.88,0.77

