한중 쇼핑 데이터 분석
(Decision Tree 부가 설명)

2020 년 05 월 10 일

Pseudo Code : Making Decision_tree

Algorithm Decision_Tree(D)

Input : an attribute-valued dataset D

1: Tree = {}
2: **if** D is "pure" OR other stopping criteria met **then**
3: terminate
4: **end if**
5: **for** all attribute a $\in$ D **do**
6: Computer information-theoretic criteria if we split on a
7: **end for**
8: a_{test} = Best attribute according to above computed criteria
9: Tree = Create a decision node that tests a_{test} in the root
10: D_v = Induced sub-datasets from D based on a_{test}
11: **for all** D_v **do**
12: $Tree_v$ = Decision_Tree(D)
13: Attach $Tree_v$ to the corresponding branch of Tree
14: **end for**
15: **return** Tree

sklearn.tree.DecisionTreeClassifier API

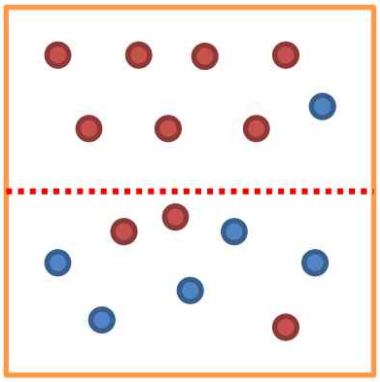
class sklearn.tree.DecisionTreeClassifier(*criterion='gini', splitter='best', max_depth=None , min_samples_split=2, min_samples_leaf=1, min_weight_fraction_leaf=0.0, max_features =None, random_state=None, max_leaf_nodes=None, min_impurity_decrease=0.0, min_im purity_split=None, class_weight=None, presort='deprecated', ccp_alpha=0.0*)

Parameter	Default	Description
Criterion	gini	Tree Split 지표
Splitter	best	각 노드의 split 전략
max_depth	None	Tree의 최대 깊이 제한
min_samples_split	2	Split을 위한 최소 노드 수
min_samples_leaf	1	최소 단말 노드 수
min_weight_fraction_leaf	0	총 가중치 합계의 최소 가중 비율
max_features	None	최대 특징 수
random_state	None	random number generator
max_leaf_nodes	None	최대 단말 노드 수
min_impurity_decrease	0.0	최소 불순도
min_impurity_split	1e-7	Threshold for early stopping in tree growth
class_weight	None	Weights associated with classes in the form {class_label: weight}.
presort	deprecated	##제거 예정
ccp_alpha	0.0	Minimal Cost-Complexity Pruning에 사용되는 복잡도 매개변수

불순도 / 불확실성

Decision Tree는 구분을 거쳐 각 영역의 순도(homogeneity가 증가, 불순도(impurity) 혹은 불확실성(uncertainty)이 최대한 감소하도록 하는 방향으로 학습을 진행하며 이를 정보획득(information gain)이라고 한다.

예시) 빨간공과 파란공의 구분



1) 엔트로피(Entropy)

- 무질서도를 정량화해서 표현한 값으로 엔트로피가 높을수록 집단의 특징을 찾는 것이 어렵다는 것을 의미
- m개의 레코드가 속하는 A영역에 대한 엔트로피는 아래와 같은 식으로 정의

$$Entropy(A) = -\sum_{k=1}^{m} p_k \log_2 (p_k)$$

- 위의 예시에 대한 엔트로피를 계산해 보면

$$Entropy(A) = -\frac{10}{16}\log_2 (\frac{10}{16}) - \frac{6}{16}\log_2 (\frac{6}{16}) \approx 0.95$$

빨간선을 기준으로 두 개의 집합으로 분할한다고 가정해보면

$$Entropy(A) = \sum_{i=1}^{d} R_i \left(-\sum_{k=1}^{m} p_k \log_2 (p_k) \right)$$

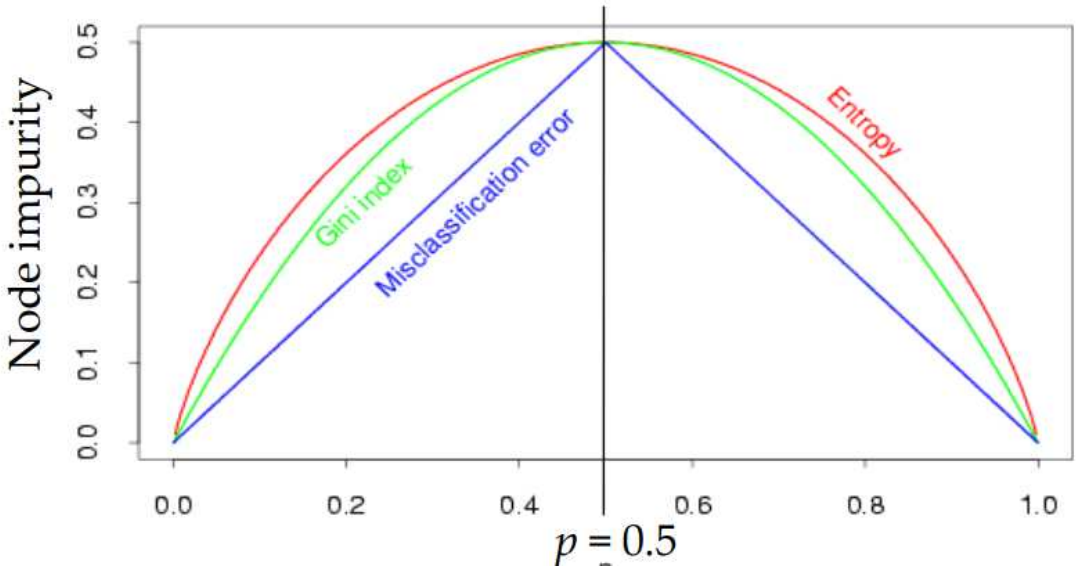
$$Entropy(A) = 0.5 \times \left(-\frac{7}{8}\log_2 (\frac{7}{8}) - \frac{1}{8}\log_2 (\frac{1}{8}) \right) + 0.5 \times \left(-\frac{3}{8}\log_2 (\frac{3}{8}) - \frac{5}{8}\log_2 (\frac{5}{8}) \right) \approx 0.75$$

다음과 같이 불확실성이 감소하는 방향으로 학습을 진행

2) 지니계수(Gini Index)

- 지니지수를 가장 감소시켜주는 예측변수와 그 때의 최적분리에 의해서 자식노드를 선택
- 지니 불순도 지수는 0-(m-1)/m의 사이의 값을 가짐

$$G.I(A) = \sum_{i=1}^{d} \left(R_i \left(1 - \sum_{k=1}^{m} p_{ik}^2 \right) \right)$$



■ scikit-learn에서의 criterion

criterion{*"gini", "entropy"*}, *default="gini"*

The function to measure the quality of a split. Supported criteria are "gini" for the Gini impurity and "entropy" for the information gain.