

Класификационно дърво(Алгоритъм ID3)

Вход: обучаващи примери S описани чрез номинални атрибути $A_1 \dots A_n$

1. Изчисли ентропията за всеки един от атрибутите

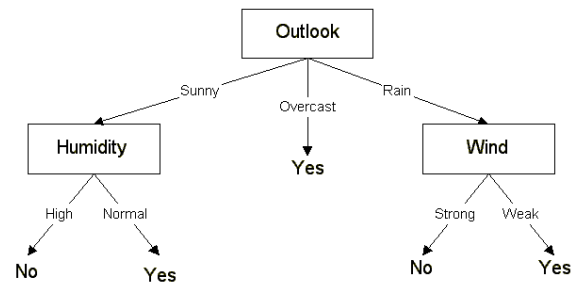
2. Раздели множеството S на подмножества използвайки атрибута с най-добра информационна печалба(Information Gain)

3. Направи възел в дървото съдържащ този атрибут

4. Повтори за всяко от подмножествата използвайки останалите атрибути

5. Обозначи листата на дървото спрямо преобладаващият клас на примерите попадащи в листо.

Критерият за спиране на разделянето на под множество може да е, че всички примери са от един клас или намаляне на бройката на примерите в множеството под определен минимум.



Задача 8: Реализирайте алгоритма за класификационно дърво ID3 и го приложете върху данните breast-cancer.arff(<http://repository.seasr.org/Datasets/UCI/arff/breast-cancer.arff>). Използвайте кросвалидация за изчисляване на точността на модела върху обучаващото множество.

За избягване на преспецифициране(overfitting) на дървотоизползвайте константа K дефинираща минимален брой на обучаващи примери в множеството.