

A decorative graphic on the left side of the slide consists of two overlapping parallelograms. The front one is blue and the back one is light green. They are positioned diagonally, with the blue one partially covering the green one.

Induktivno učenje - Primena ID3 Algoritma



Pristupi masinskom ucenju :

- **Induktivno ucenje**
- Analiticko ucenje
- Case-Based learning sistem
- Neuronske mreze
- Genetski algoritam
- Hibridni model

Indutivno ucenje zasniva se na kreiranju stabla odlucivanja algoritmom ucenja. Potreno je za definisane vrednosti atributa u takvim sistemima odrediti klasifikaciju na osnovu pravila odluke.



Sta je stablo odluke ?

Stablo odluke predstavlja struktura tipa stabla gde:

Unutrasnji cvorovi - odgovaraju atributima uzoraka i predstavljaju izbor izmedju vise alternativnih mogucnosti.

Grane - Odgovaraju vrednostima atributa

Listovi - Predstavljaju odluke odnosno klase kojima pripadaju vrednosti atributa

Dve faze u kreiranju stabla odluke:

Faza izgradnje stabla: (Top - Down)

Inicijalno se krece od korena stabla gde su smesteni svi primeri. Odabirom atributa se rekursivno vrši particionisanje i dolazenje do novih cvorova odluke.

Faza odsecanja stabla: (Bottom - Up)

Uklanjanje podstabla ili grana radi unapredjenja tacnosti modela



ID3 Algoritam

- Kriterijum za selekciju atributa kod ovog algoritma je **Informacijska dobit**. Pretpostavka je da su svi atributi kategoricki.
- Racuna se kao razlika **Mere neizvesnosti proizvoljne promenljive i Entropije cvora**
- Zaustavlja se kada svi primeri pripadaju istoj klasi ili kad je najbolja informacijska dobit manja ili jednaka 0

Mera neizvesnosti proizvoljne promenljive:

$$I = - \sum_c p(c) \log_2 p(c)$$


$p(c)$ - verovatnoca da proizvoljno izabran primer pripada klasi c

Entropija cvora:

$$I_{res} = - \sum_v p(v) \sum_c p(c|v) \log_2 p(c|v)$$


$p(v)$ - verovatnoca da proizvoljno izabran primer v ima vrednost odabranog atributa

$p(c|v)$ - verovatnoca da primer kojima ima v kao vrednost odredjenog atributa pripada klasi c

- 
- Razlika između entropije za slučaj kada nije poznata vrednost atributa i očekivane količine informacija u slučaju poznate vrednosti atributa predstavlja **informacijsku dobit** kada se posmatrani atribut koristi kao kriterijum za razvrstavanje primera.

$$Gain(A) = I - I_{res}(A)$$

- ID3 određuje atribut sa najvećim dobitkom, tj. preferira atribut koji nosi najviše informacija za ceo skup primera
- Iz cvora koji je obeležen izabranim atributom postoji onoliko grana koliko vrednosti ima izabrani atribut i originalni skup primera deli se u disjunktne podskupove prema vrednostima tog atributa. Proces se ponavlja rekurzivno, sve dok svi primeri u posmatranom podskupu ne pripadaju istoj klasi.

- 
- ID3 algoritam se koristi za indukovanje stabla odluke na osnovu primer tipa:

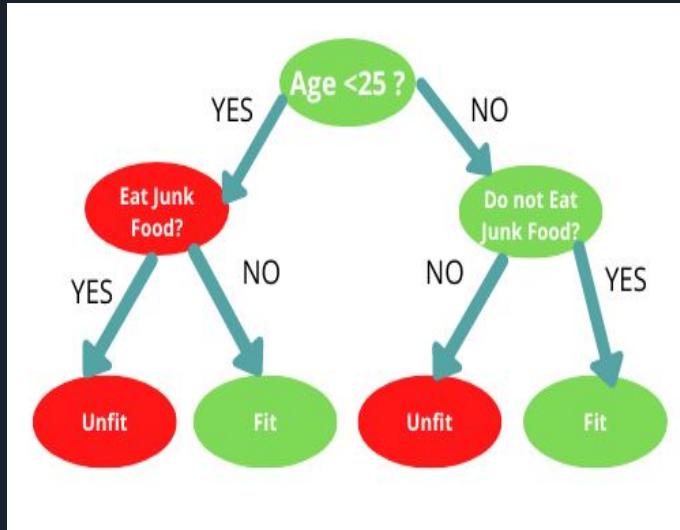
(v_atribut1, v_atriut2,...v_atributn, klasa)

- Nakon toga dobijeno stablo odluke se koristi za klasifikaciju novih uzoraka. Potrebno je definisati funkciju koja ce biti mera za izbor atributa. Takva funkcija je metrika **Informacijska dobit**.
- Da bi definisali pojam informacijska dobit, neophodna nam je definicija entropije.
- Pretpostavka: stablo odluke koje treba da dobijemo klasifikuje instance u dve kateorije: **P(positive)** i **N(negative)**
- Za zadati skup S, koji sadrži takve pozitivne i negativne klase, entropija za S u odnosu na takvu Bulovu klasifikaciju je:

$$H(S) = - P(\text{positive}) \log_2 P(\text{positive}) - P(\text{negative}) \log_2 P(\text{negative})$$

P(positive): broj pozitivnih primera u S , P(negative): broj negativnih primera u S

ID3 Algoritam



1. Ako svi primeri pripadaju istoj klasi kreiraj list sa vrednošću koja odgovara toj klasi.
2. U suprotnom:
 - a) Nađi atribut sa najvećom dobiti.
 - b) Dodaj granu za svaku vrednost tog atributa.
 - c) Rasporedi primere u odgovarajuće podskupove.
 - d) Za svaki podskup ponovi algoritam.



Zakljucak

Prednosti

- Stabla odluke se lako prate ukoliko su kompaktna, mogu se predstaviti preko skupa pravila pa se smatraju lako razumljivim
- Koriste nominalne(kategoricke) i numericke(kontinualne) attribute
- Mogu raditi sa skupovima podataka koji poseduju greske

Mane

- ID3 algoritam zahteva attribute za diskretnim vrednostia
- Pokazuju se lose kada postoji mnogo kompleksnih interackija a dobro kada postoji mali broj visoko relevantnih atributa