Elektronski fakultet Niš

Induktivno učenje, Primena ID3 algoritma

Student: Miljana Randjelovic Br. Indeksa: 1424

Uvod

Stabla odlučivanja su među najmoćnijim alatkama za mašinsko učenje koje su danas dostupne i koriste se u širokom spektru aplikacija u stvarnom svetu, od predviđanja klikova na oglas na Facebook-u[1] do rangiranja Airbnb iskustava. Ipak, oni su intuitivni, laki za tumačenje — i laki za implementaciju. Učenje putem primera – indukcija je proces izvođenja opštih pravila iz znanja koja se sadrže u konačnom skupu primera. Induktivno učenje se može posmatrati kao pretraživanje prostora problema radi nalaženja rešenja.

U nastavku će biti opisan algoritam ID3 iz grupe algoritama za induktivno učenje i njegova implementacija uz pomoć programskog jezika Python.

Stabla odlučivanja

Stabla odlučivanja se mogu koristiti za regresiju (kontinuirani izlaz realne vrednosti, na primer predviđanje cene kuće) ili klasifikacija (kategorički izlaz, na primer. predviđanje neželjene e-pošte), ali ovde će biti fokusi na klasifikaciji. Klasifikator stabla odluka je binarno stablo gde se predviđanja prave prelaskom stabla od korena do lista — u svakom čvoru idemo levo ako je karakteristika manja od praga, u suprotnom desno. Konačno, svaki list je povezan sa klasom, što je izlaz prediktora.

ID3 algoritam

ID3 (Iterative Dichotomiser 3 – Ross Quinlan) je induktivni algoritam koji se komercijalno koristio. ID3 uzima skup primera problema i generiše stablo odlučivanja. On izgradjuje rešenje sa što manje pretpostavki, odnosno cilj je izgraditi što manje stablo. Primer je kombinacija:

– faktora odlučivanja,

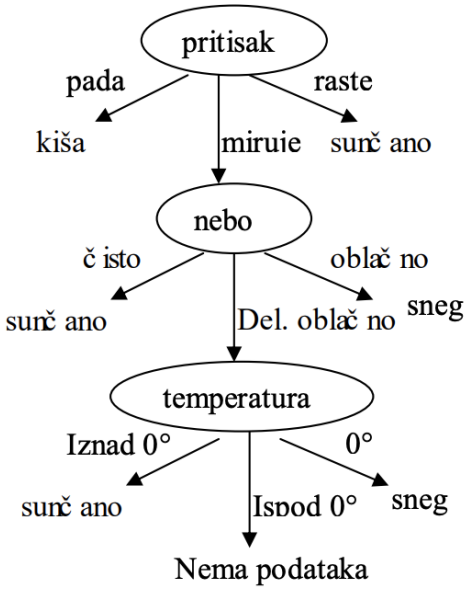
– vrednosti faktora odlučivanja,

– akcije specifične za taj primer.

Primena ID3 algoritma

Faktori odluč ivanja				Rezultat
Temperatura	**Vetar**	**Nebo**	**Pritisak**	**Prognoza**
Iznad 0°	Zapadni	Oblač no	Pada	Kiša
Ispod 0°	*	Oblač no	Miruje	Sneg
Iznad 0°	Istoč ni	Oblač no	Raste	Sunč ano
Iznad 0°	*	Delimič no oblač no	Miruje	Sunč ano
*	*	Čisto nebo	Miruje	Sunč ano
Iznad 0°	Juž ni	Čisto nebo	Pada	Kiša
0°	Severni	Delimič no oblač no	Miruje	Sneg

* - označ ava da taj faktor u tom primeru nije bitan.



Izbor najpre najvažnijeg elementa – prva odluka je izbor atributa barometarski pritisak kao korena stabla. Ispitivanje prvo ovog faktora je najefikasnije, pošto dva od tri moguća odgovora vode neposrednom predviđanju, bez potrebe da se ispituju faktori nebo ili temperatura. Dakle, ID3 smešta najvažnije faktore odlučivanja blizu korena stabla. • Nema podataka – povremeno se dešavaju slučajevi da se ne može dobiti rezultat. To označava situaciju za koju ne postoje primeri na osnovu kojih se može izvući zaključak. Ako se naiđe na ovakve situacije, to je obično znak da je skup primera nedovoljan i da ga je potrebno upotpuniti novim primerima koji bi podržali te mogućnosti. • Isključivanje nevažnih faktora – još jedna važna odlika ID3 je da ustanovi da originalni skup sadrži atribut smer vetra, ali ovaj faktor se ne pojavljuje nigde u stablu odlučivanja. ID3 je odlučio, na osnovu skupa primera, da je ovaj faktor nebitan za predviđanje vremena. Čovek koji je ekspert za predviđanje vremena može da bude nesvestan da je ovaj faktor nevažan.

Koraci koriscenja ID3 algoritma:

1. Izračunati entropiju svakog atributa **a** skupa podataka **S**.
2. Particionisati („podeliti“) skup **S** na podskupove koristeći atribut za koji je rezultujuća entropija posle cepanja minimizirana; ili, ekvivalentno, dobijanje informacija je maksimalno.
3. Napraviti čvor stabla odluka koji sadrži taj atribut.
4. Povratak na podskupove koristeći preostale atribute.

ID3 ne garantuje optimalno rešenje. Može se približiti lokalnom optimumu. Koristi pohlepnu strategiju izborom lokalno najboljeg atributa da podeli skup podataka na svakoj iteraciji. Optimalnost algoritma se može poboljšati korišćenjem vraćanja unazad tokom potrage za optimalnim stablom odlučivanja po cenu mogućeg dužeg trajanja.

ID3 može prepuniti podatke o obuci. Da bi se izbeglo prekomerno prilagođavanje, manja stabla odlučivanja bi trebalo da imaju prednost u odnosu na veća. Ovaj algoritam obično proizvodi mala stabla, ali ne proizvodi uvek najmanje moguće stablo odlučivanja.

ID3 je teže koristiti na kontinuiranim podacima nego na faktorisanim podacima (faktorisani podaci imaju diskretni broj mogućih vrednosti, čime se smanjuju moguće tačke grananja). Ako su vrednosti bilo kog datog atributa neprekidne, onda postoji mnogo više mesta za podelu podataka na ovom atributu, a traženje najbolje vrednosti za podelu može biti dugotrajno.

Upotreba

ID3 algoritam se koristi za obuku na skupu podataka **S** za proizvodnju stabla odlučivanja koje se čuva u memoriji. Tokom rada, ovo stablo odlučivanja se koristi za klasifikaciju novih test slučajeva (vektora karakteristika) prelaskom kroz stablo odlučivanja koristeći karakteristike datuma da bi se došlo do lisnog čvora.

Racunanje

Entropija H(S) je mera količine nesigurnosti u skupu podataka **S.**

$$\mathrm{H}(S) = \sum_{x \in X} -p(x) \log_2 p(x)$$

Gde su:

S – Trenutni skup podataka za koji se izračunava entropija. Ovo se menja u svakom koraku ID3 algoritma, bilo na podskup prethodnog skupa u slučaju razdvajanja na atributu ili na particiju "sestre" roditelja u slučaju da je rekurzija prethodno prekinuta.

Ks – Skup klasa u S

p(k) – Proporcija broja elemenata u klasi k prema broju elemenata u skupu S. Kada je H(S)=0, skup S je savršeno klasifikovan (tj. svi elementi u S su iste klase).

U ID3, entropija se izračunava za svaki preostali atribut. Atribut sa najmanjom entropijom se koristi za podelu skupa S na ovoj iteraciji. Entropija u teoriji informacija meri koliko informacija se očekuje da se dobije merenjem slučajne promenljive; kao takav, može se koristiti i za kvantifikaciju količine do koje je distribucija vrednosti veličine nepoznata. Konstantna količina ima nultu entropiju, pošto je njena distribucija savršeno poznata. Nasuprot tome, ravnomerno raspoređena slučajna promenljiva (diskretno ili kontinuirano uniformna) maksimizira entropiju. Stoga, što je veća entropija u čvoru, manje je informacija poznato o klasifikaciji podataka u ovoj fazi stabla; i stoga, veći je potencijal da se ovde poboljša klasifikacija.

Kao takav, ID3 je pohlepna heuristika koja vrši pretragu u prvom redu za lokalno optimalne vrednosti entropije. Njegova tačnost se može poboljšati prethodnom obradom podataka.

Dobitak informacija

Dobitak informacija G(A) je mera razlike u entropiji od pre do posle skupa S koji je podeljen na atribut A. Drugim rečima, kolika je nesigurnost u S smanjena nakon podele skupa S na atribut A.

Implementacija ID3 uz pomoc PySwarms

PySwarms je alatka zasnovana na Python-u za predvidjanje vremena na osnovu postojeceg seta podataka. PySwarms implementira tehnike optimizacije roja sa više čestica na visokom nivou. Kao rezultat toga, teži da bude lak za upotrebu i prilagodljiv. Pomoćni moduli se takođe mogu koristiti da vam pomognu sa konkretnim problemom optimizacije. Primer implementacija ID3 optimizacije dodat je na git repozitorijum.

Literatura

1. https://www.cise.ufl.edu/~ddd/cap6635/Fall-97/Short-papers/2.htm
2. https://towardsdatascience.com/decision-trees-for-classification-id3-algorithm-explained-89df76e72df1
3. https://en.wikipedia.org/wiki/ID3_algorithm
4. https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/id3.html