

STROJNO UČENJE

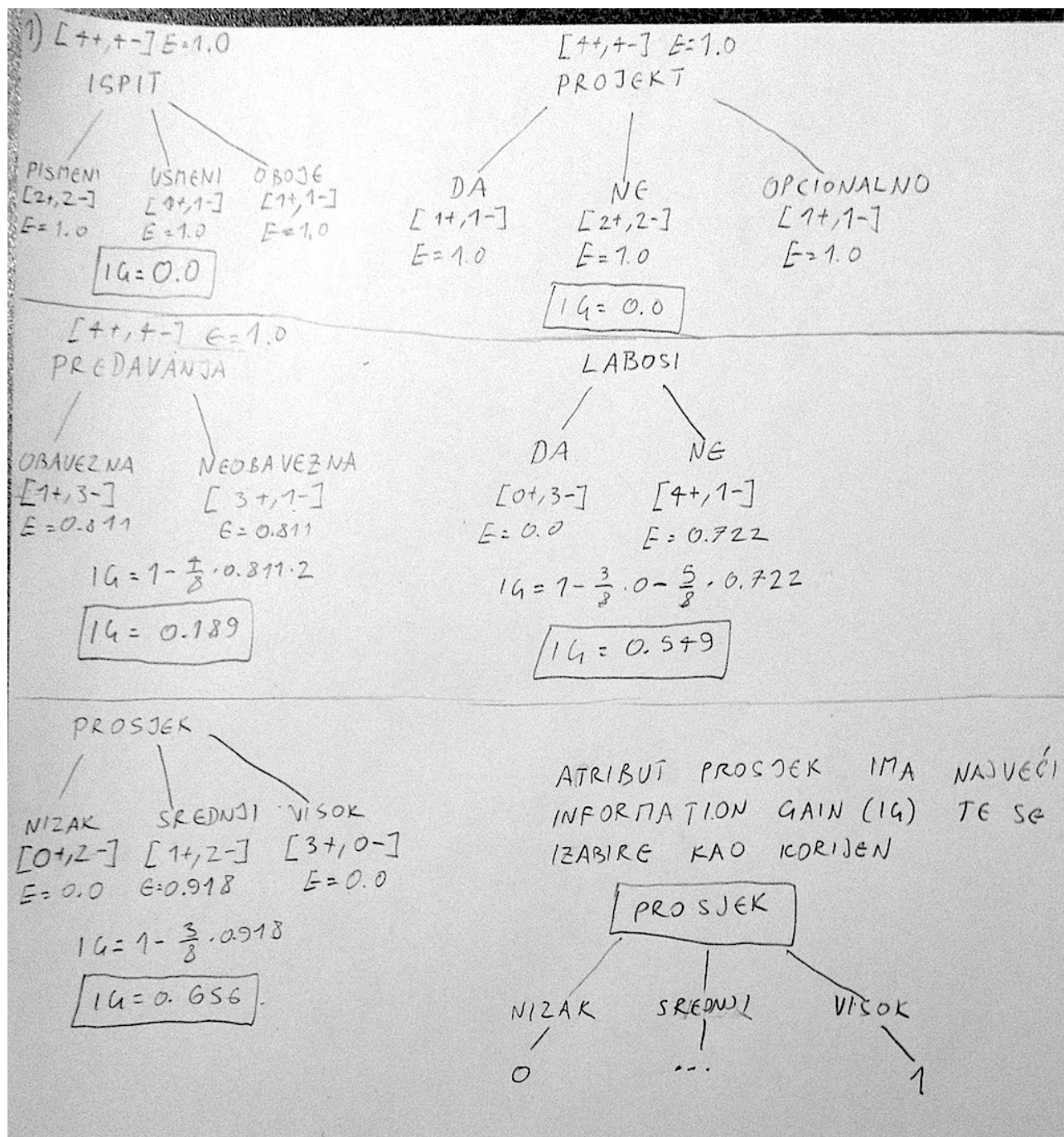
4. Domaća Zadaća

Krešimir Špes

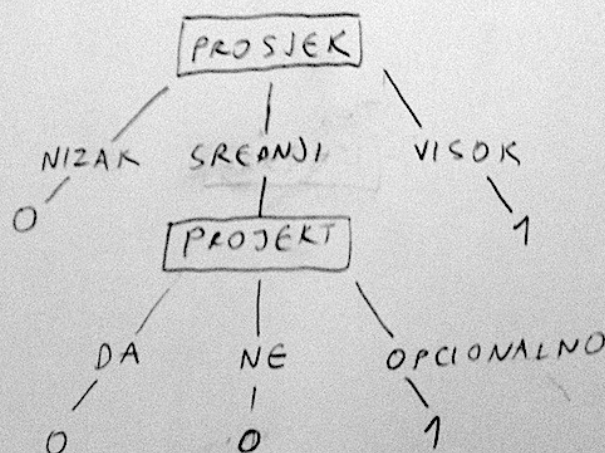
0036419866

ak. god. 2011. / 2012.

a)



2) KORJEN = PROSJEK, GRANA PROSJEK = SREDNJI, $E = 0.918$
 ISPIT: PISMENI [0+1-], USMENI [1+1-], OBOJE [0+0-], $IG = 0.251$
 PROJEKT: DA [0+1-], NE [0+1-], OPCIONALNO [1+0-], $IG = 0.918$
 LABOSI: DA [0+1-], NE [1+1-], $IG = 0.251$
 PREDAVANJA: OBAVEZNA [0+1-], NEOBAVEZNA [1+1-], $IG = 0.251$
 ČVOR: PROJEKT



b) Uzeli bi sve ostale poznate vrijednosti tog atributa te na temelju njih izračunali vjerojatnosti koje će se koristiti za računanje informacijske dobiti.

c) Može se dogoditi. Mogući razlozi su na primjer teacher noise ili razlika u subjektivnosti više označivača.

Problem možemo riješiti tako da zaustavimo rast stabla prije savršene klasifikacije ili možemo izgraditi stablo unatoč šumu u podacima i naknadno ga podrezati. Podrezivanje se je pokazalo u praksi kao boljom metodom.

Podrezivat ćemo tako da za sva podstabla iterativno pokušamo podrezati granu i testirati stablo na skupu primjera za provjeru. Ako dobijemo istu ili bolju klasifikaciju, grana se podrezuje. Postupak se ponavlja dok se ne počinje narušavati točnost. Ovakva metoda se koristi i za rješavanje problema šuma u podacima i ograničavanje složenosti stabla.

d) To znači da model nema fiksni broj parametara koje će optimizirati već složenost klasifikatora raste sa brojem primjera. Konkretno kod stabla odluke parametri modela su čvorovi stabla.

e) Da. Jer Stablo odluke nije ograničeno fiksnim brojem parametara kao što je to pravac. Složenost modela a time i odjeljivanje primjera raste sa brojem čvorova stabla. Možemo jednostavno sve primjere ugraditi u stablo i stablo će ih zbog toga savršeno klasificirati. To naravno ima smisla samo za manje skupove primjera jer inače dobivamo ogromna stabla koja više štete nego koriste. Pa iako za velike skupove primjera možemo napraviti stablo koje sadrži sve primjere, često se to ne radi, tj. Ograničava se kompleksnost stabla do neke razumne razine.

f) ovisnosti o distribuciji primjera, stablo ograničeno na 3 razive ne bi moglo razdjeljivati sve moguće distribucije primjera. Pošto smo ograničili dubinu stabla na 3, time smo zapravo ograničili model, tj. Broj parametara modela što se reflektira na složenost samog modela (ograničenje jezikom), stoga logički slijedi da ograničen model neće uvijek moći dati hipotezu koja je konzistentna s primjerima za učenje.

Ima smisla ograničavati stablo jer općenito ne želimo stablo koje je prekompleksno jer pretjeranom složenošću time ograničavamo moć generalizacije, a može doći i do prenaučenosti modela.

Ograničio bi dubinu stabla ovisno o količini primjera, npr max_dubina je $\log_2(\text{broj_primjera})$.

U navedenom primjeru to ograničenje bi bilo $\log_2(8) == 3$.