Pretprocesiranje podataka

Pretprocesiranje podataka (eng. preprocessing) je široka oblast koja obuhvata različite tehnike i strategije koje se koriste zarad dobijanja podataka pogodnijih za rad. Te tehnike i strategije spadaju ili u izbor atributa za analizu ili u kreiranje novih atributa/menjanje postojećih atributa i ne postoji generalno pravilo kojim redom će se one primenjivati. Na izbor tehnika i strategija utiče priroda podataka sa kojima se radi.

Transformacija numeričkih atributa

Termin transformacija atributa označava transformaciju koja se primenjuje na vrednosti atributa. To znači da će vrednost atributa svake instance biti transformisana. Postoje različiti tipovi transformacija, a koji tip će biti korišćen zavisi od prirode podataka. U nekim situacijama je dovoljno primeniti jednostavnu matematičku funkciju na vrednosti atributa x (npr. x^2, $logx$, $\sqrt{x}$). U nekim drugim situacijama potrebno je izvršiti standardizaciju, o čemu će biti reči kasnije.

Izbor atributa

Skupovi podataka mogu da sadrže podatke sa velikim brojem atributa tj. podatke koji imaju visoku dimenzionalnost. Možemo imati više hiljada atributa, jedan takav primer je skup podataka koji sadrži broj pojavljivanja svake reči u nekom dokumentu.
Postoje brojne prednosti smanjenja dimenzionalnosti: algoritam mašinskog učenja zahteva manje vremena i memorije, a takodje se i rešava problem prokletstva dimenzionalnosti (prokletstvo dimenzionalnosti je situacija da broj instanci podataka potrebnih za trening modela raste eksponencijalno sa porastom broja atributa), dobijeni model je lakši za tumačenje ako je manji broj atributa uključen i vizualizacija podataka je olakšana ako postoji manji broj atributa.
Smanjenje dimenzionalnosti često se odnosi na tehnike koje do smanjenja vode pravljenjem novih atributa kombinovanjem postojećih. Primer algoritma za smanjenja dimenzija koji ste videli na kursu LSM je Analiza glavnih komponenti (PCA). Algoritmi koji se oslanjaju na grafovsku analizu podataka su dosta bolji od PCA. Primeri takvih algoritama su t-SNE i trenutno najbolji takav algoritam - UMAP.

Standardizacija podataka

Standardizacija se sastoji u tome da izračunamo prosek i standardnu devijaciju nekog atributa na trening skupu i da se od svake vrednosti tog atributa na sva tri skupa (trening, validacioni i test skup) oduzme prosek, pa da se potom svaka vrednost tog atributa podeli standardnom devijacijom. Time se obezbedjuje da svaki atribut ima

prosek 0 i standardnu devijaciju 1.
Kada je potrebno uraditi standardizaciju?
Najbitnija situacija kad je suštinski važno uraditi standardizaciju jeste kada model koristi euklidsku metriku (npr. KNN), jer je tada vrlo bitno da svi atributi budu na istoj skali. Na primer ako je neki atribut reda veličine nekoliko miliona, a neki drugi atribut reda veličine 10, onda bi bez standardizacije drugi atribut bio zanemarljiv u odnosu na prvi u kontekstu računanja metrike, iako je možda značajniji u kontekstu predikcije. Dakle i ako koristimo l_1 ili l_2 regularizaciju neophodno je da prvo uradimo standardizaciju. Drugi razlog za korišćenje standardizacije je interpretabilnost npr. kod linearne regresije. Ako ne uradimo standardizaciju možemo dobiti da je neki koeficijent ogroman, iako je beznačajan. Takodje, pokazalo se da neki metodi (npr. neuronske mreže) brže konvergiraju ako im se proslede standardizovani podaci.

Kategorički prediktori

Pomenuli smo da se u zadacima mašinskog učenja srećemo sa raznim tipovima podataka. Većina modela koje smo do sada koristili uspešno je manipulisala podacima numeričkog tipa, dok kada je bilo reči o podacima koji po svojoj prirodi nisu neke brojevne vrednosti, morali smo da razmislimo na koji način ih možemo kodirati tako da model može uspešno da ih koristi. Takav tip podataka zovemo kategoričkim tipom i sada ćemo mu posvetiti malo pažnje.
Podatke kategoričkog tipa (ili kraće: kategoričke podatke) možemo podeliti na nominalne i ordinalne. Ordinalni kategorički podaci se na neki način mogu sortirati, dok sa nominalnim to nije slučaj. Npr. ordinalni tip podataka bi bio stepen stručne spreme, a nominalni, recimo, boja očiju. U zavisnosti od toga da li kategoričke podatke vidimo kao nominalne ili ordinalne, razlikovaće se i sam način na koji ih obradjujemo.
Dakle, glavni izazov u radu s kategoričkim podacima jeste njihovo kodiranje. Jasno je da kada se bavimo klasifikacijom, naša ciljna promenljiva je kategoričkog tipa i nju nije potrebno na neki specijalan način kodirati, pa ćemo se posvetiti samo kategoričkim prediktorima. Neki modeli poput random forest-a ne zahtevaju posebno zapisivanje kategoričkih prediktora, što je njihova velika prednost. S druge strane, linearni i logistički regresioni modeli su podrazumevali prediktore predstavljene brojevima, što znači da mogu koristiti kategoričke prediktore samo ako se oni predstave na takav način.
Drugi problem koji se javlja u radu s kategoričkim prediktorima je izbor adekvatne metrike za rad sa njima. Npr, modeli KNN i k-means najčešće koriste euklidsku metriku kako bi izmerili sličnost medju podacima, što znači da ako želimo da koristimo i kategoričke prediktore, neophodno je da ih kodiramo kao numeričke ili da koristimo neku drugu metriku.
Kodiranje kategoričkih vrednosti brojevima u ovom slučaju mora biti takvo da zaista izrazi sličnosti i razlike tačaka kako bi ovi modeli funkcionisali na smislen način.
Navedimo sada nekoliko načina na koje možemo kodirati kategoričke prediktore. Neka prediktor koji posmatramo uzima k mogućih kategorija.
1. label encoding

Ovaj način kodiranja koristimo najčešće kod ordinalnih podataka, kako bismo očuvali uredjenost medju njima. Ideja je da kategorije kodiramo rastućim nizom brojeva, obično vrednostima od 0 do $k-1$ u odgovarajućem smeru:

2. one hot encoding

Ovo kodiranje podrazumeva kreiranje k novih indikatorskih prediktora za svaku kategoriju, od kojih tačno jedan uzima vrednost 1, a ostali 0. Češće se koristi kod nominalnih prediktora.

3. dummy encoding

Dummy encoding je tehnika kodiranja slična prethodnoj, s tim što ona konstruiše jedan prediktor manje. Dakle, jedna kategorija se bira kao referentna i svakoj od preostalih $k-1$ kategorija se dodeljuje novi indikatorski prediktor. Prednost dummy encodinga u odnosu na one hot encoding je što dobijamo prediktore koji su linearno nezavisni. Zaista, kod one hot encodinga je zbir svih k kolona jednak koloni koja ima sve jedinice, tj. jednak je slobodnom članu, što predstavlja problem kod npr. linearne i logističke regresije, jer odgovarajuća matrica neće biti invertibilna.

Mana one hot i dummy kodiranja je ta što kada imamo puno kategorija, konstruiše se puno novih prediktora. Postojanje velikog broja prediktora može negativno da se odrazi na kvalitet modela iz više razloga. Kada imamo veliki broj kategorija nekog atributa, najčešže se koriste frequency encoding (kategorija se zamenjuje učestalošću te kategorije u trening skupu) i mean encoding (kategorija se zamenjuje srednjom vrednošću zavisne promenljive koja pripada toj kategoriji), jer ne proizvode dodatne atribute.

Ostaje nam još da pomenemo da je kao i kod klasifikacije važno da raspodela kategorija bude ravnomerna, jer model teško prepoznaje značaj kategorija koje su slabo zastupljene u bazi. U tim situacijama se često kategorije čija je zastupljenost manja od 5% posmatraju kao jedna.

Definisanje novih kategorija na osnovu postojećih se može koristiti i kada je broj kategorija jako veliki. Prisustvo velikog broja kategorija od kojih se svaka pojavljuje svega 2-3 puta u bazi nije dobro za interpretaciju modela, jer ne postoji dovoljno informacija o svakoj od njih da bi se izveo konkretan zaključak. Na primer, ako u bazi imamo kolonu u kojoj se nalazi 100 različitih marki automobila, korisnije je da od njih napravimo novu kolonu koja predstavlja recimo zemlju ili kontinent porekla autombila.