Pregled gradiva

OSNOVNA PITANJA

1. **Koje vrste mašinskog učenja postoje?**

   - Nadgledano učenje
   - Nenadgledano učenje
   - Učenje potkrepljivanjem (primer je autonomna vožnja automobila, ili igranje šaha)

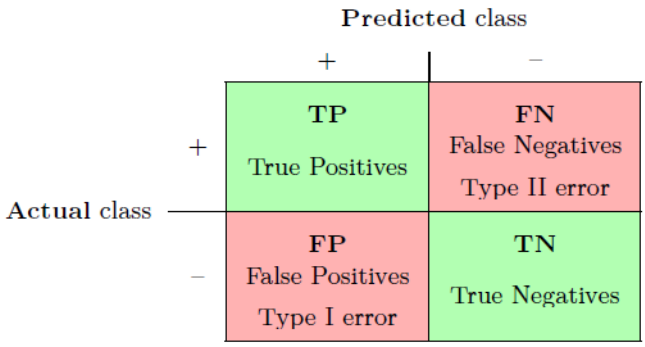
2. **Koji su koraci u izgradnji modela?**

   Prvi korak je izbor atributa koje ćemo koristiti. Ako imamo sirove podatke, možda će biti neophodno da konstruišemo atribute. Kada ih konstruišemo, možda ćemo neke da izbacimo ako se ispostavi da su linearno zavisni od ostalih, zatim je neophodno da ih standardizujemo. Nakon toga biramo koju ćemo metodu mašinskog učenja da koristimo. To radimo na osnovu toga da li je u pitanju problem regresije ili klasifikacije, kao i na osnovu toga koje veličine je dataset. Naravno, možemo da izaberemo više različitih metoda i da napravimo više modela, pa da ih uporedimo na test skupu i izaberemo najbolji.

3. **U kom odnosu se dele podaci na training, validation i test skup? I čemu služi svaki od skupova?**

   Ako je ukupan broj podataka mali, onda ih najčešće delimo u odnosu 70:15:15. A ako imamo veliki skup podatka (npr. preko 10000), onda je dovoljno uzeti po 1500-2000 podataka u validation i test skupu, jer nam ti skupovi služe samo za ispitivanje precizosti modela, a već sa 1500-2000 podataka ćemo dobiti verodostojan rezultat. Trening skup služi za pravljenje modela, odnosno za dobijanje optimalnih parametara modela, validation skup služi za dobijanje optimalnih hiperparametara modela, a test skup za konačnu evaluaciju modela.

4. **Kako izgleda matrica konfuzije?**



5. **Koje mere preciznosti koristimo u slučaju klasifikacije?**

   accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$

   precision $= \frac{TP}{TP+FP}$

   recall $= \frac{TP}{TP+FN}$

   specifity $= \frac{TN}{TN+FP}$

   F_1 score $= \frac{2 \cdot precision \cdot recall}{precision+recall}$

   AUC, tj. verovatnoća da slučajno odabrani element iz pozitivne kategorije ima veći skor (skor se izračunava modelom) od slučajno odabranog elementa iz negativne kategorije.

6. **Koje mere preciznosti koristimo ako imamo nebalansirane podatke?**

   F_1 score i AUC (površina ispod ROC krive).

7. **Koja je razlika izmedju regresije i klasifikacije?**

   Klasifikacija se koristi kada je zavisna promenljiva kategoričkog tipa, na primer kada hoćemo da odredimo da li je na slici mačka ili nije, a regresija se koristi kada je zavisna promenljiva neprekidnog tipa, npr. kada hoćemo da predvidimo visinu nekog čoveka ili broj prodatih proizvoda.

8. **Šta je standardizacija podataka i čemu služi?**

   Standardizacija podataka je proces koji se sastoji u sledećim koracima. Posmatramo neki fiksiran prediktor, odnosno neku odabranu kolonu u podacima. Prvo računamo uzoračku sredinu i standardnu devijaciju te kolone na trening skupu. Zatim od vrednosti te kolone na trening, validation i test skupu oduzimamo izračunatu uzoračku sredinu i delimo sa izračunatom standardnom devijacijom. Na taj način dobijamo da su vrednosti tog prediktora raspodeljeni oko nule, sa disperzijom 1.
   Ovaj proces je neophodan kada nam je bitno da svi prediktori budu na istoj skali. A to nam je bitno npr. ako koristimo euklidsku metriku ili ako koristimo *lasso* ili *ridge* regularizaciju. Standardizacija svakako ne može da odmogne, pa je poželjno uraditi je. Npr. ispostavilo se da optimizacija parametara neuronske mreže traje kraće ako su podaci standardizovani.

9. **Šta je overfitting, kako ga prepoznati i kako ga izbeći?**

   Overfitting je pojava kada se model previše prilagodi trening podacima. Prepoznajemo ga tako što izračunamo preciznost modela i na trening i na test skupu. Ako je preciznost na trening skupu značajno veća nego na test skupu, to je znak da je došlo do overfittinga. Primer je Lagranžov interpolacioni polinom koji prolazi kroz sve tačke trening skupa, tj. greška na trening skupu je nula, ali će greška na test skupu biti velika. Postoji više načina za borbu sa overfittingom:

   - Regularizacija, koja služi da ne damo pojedinim parametrima da budu previše veliki.
   - Pojednostavljivanje modela, odnosno smanjivanje broja parametara.

10. **Kada najčešće koristimo neuronske mreže?**

    Neuronske mreže se koriste kada imamo dovoljno veliki skup podataka. Ako imamo mali skup podataka, onda druge metode postižu bolje rezultate. Takodje, ako su podaci sirovi, tj. neobradjeni, neuronske mreže postižu mnogo bolje rezultate nego ostale metode, jer one u suštini same kreiraju atribute na osnovu kojih vrše klasifikaciju.

11. **Kako biste napravili model koji na osnovu teksta mejla predvidja da li je mejl spam?**

    Pretpostavimo da imamo označenu bazu podataka, tj. da imamo tekstove mejlova i da li su spam ili nisu. Cilj je odabrati relevantne atribute. Posmatraćemo reči koje se često pojavljuju u spam mejlovima, a skoro nikad u regularnim mejlovima, kao i reči koje se retko pojavljuju u spam mejlovima, a često u regularnim mejlovima, i na osnovu tih reči ćemo napraviti kategoričke prediktore. Verovatno će nam prva analiza biti korisnija. Tom analizom ćemo ustanoviti da se reči kao što su lutrija, novac,

poklon, nasledstvo, popust, akcija, često pojavljuju u spam mejlovima, a retko u regularnim mejlovima, pa ćemo za svaku od tih reči napraviti prediktor. Dakle, kada dobijemo neki novi mejl, analiziraćemo njegove reči i ako sadrži neku od ovih reči (a pogotovo ako sadrži više njih), možemo taj mejl da označimo kao spam.

12. **Kako funkcionišu sistemi za preporuku, npr. kod Netflixa?**

    Netflix čuva podatke o svim serijama i filmovima koje je neki korisnik pogledao. Te informacije su prediktori koje koristi za predvidjanje. Na osnovu toga nam predlaže sadržaje koji su slični pogledanim, na primer na osnovu žanra kom pripadaju ti sadržaji. Takodje uporedjuje nas sa drugim korisnicima, i nalazi korisnike koji su pogledali dosta istih sadržaja kao mi, i onda nam predlaže sadržaje koje su ti korisnici pogledali, a mi nismo.

1. **Kako izgleda funkcija greške u slučaju regresije, a kako u slučaju klasifikacije?**

   U slučaju regresije najčešća funkcija greške je Mean squared error (MSE) i definisana je sa

   $$MSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

   U slučaju regresije funkcija greške je unakrsna entropija (Cross entropy loss) i jednaka je negativnoj vrednosti logaritma funkcije verodostojnosti, odnosno:

   $$Crossentropyloss = -\sum_{j=1}^{n} \sum_{i=1}^{k} p_{ji} log p_{ji},$$

   gde je k broj mogućih kategorija, a p_{ji} verovatnoća da j-ta instanca pripada i-toj kategoriji.

2. **Koje su prednosti konvolutivnih mreža u odnosu na potpuno povezane?**

   Glavna prednost je u tome što konvolutivne mreže imaju mnogo manje parametara, pa je treniranje mnogo brže. Još jedna prednost je što su konvolutivne mreže invarijantne u odnosu na translaciju, pa npr. daju bolje rezultate pri obradi slika.

3. **Koji su osnovni tipovi neuronskih mreža postoje i za šta se koriste?**

   Potpuno povezane neuronske mreže - najčešće se ne koriste samostalno, već kao sastavni deo drugih neuronskih mreža.
   Konvolutivne neuronske mreže - koriste se za obradu signala, poput zvuka i slike, ali takodje i teksta.
   Rekurentne neuronske mreže - koriste se za obradu sekvencijalnih podataka, kao što su obrada prirodnog jezika (eng *natural language processing* - NLP) i obrada vremenskih serija.

4. **Navesti primer kada je bitnije minimizovati false negatives u odnosu na false positives?**

   Pretpostavimo da treba na osnovu slike da klasifikujemo da li je tumor benigni ili maligni. Neka benigni predstavlja negativnu kategoriju, a maligni pozitivnu. Mnogo veće posledice će imati ako maligni klasifikujemo kao benigni, nego da benigni klasifikujemo kao maligni. Dakle, mnogo nam je bitnije da minimizujemo false negatives.

5. **Koji algoritmi se koriste za smanjivanje dimenzija podataka?**

   Razlikujemo algoritme zasnovane na kovarijacionoj matrici i algoritme zasnovane na grafovskoj analizi podataka. Najpoznatiji predstavnik prve grupe je PCA, a predstavnici druge grupe su t-SNE i UMAP.
   UMAP je trenutno najpopularniji algoritam i više o njemu možete pogledati na sledećem linku.

6. **Ukratko objasniti logističku regresiju?**

   Logistička regresija se koristi za binarnu klasifikaciju. Ovaj metod se sastoji u tome da se izračuna vrednost logističke funkcije za ulaz X:

   $$\frac{1}{1 + e^{-(1+\beta X)}}$$

   gde je β vektor parametara modela dobijen minimizovanjem cross entropy funkcije gubitaka. Ako je vrednost veća od threshold-a, ulazu se dodelju kategorija 1, a u suprotnom kategorija 0. Threshold je najčešće postavljen na 0.5, ali može biti i neki drugi broj.

7. **Ukratko objasniti kNN?**

   kNN je metoda nadgledanog učenja, koja može da se koristi i za regresiju i za klasifikaciju. Sastoji se u tome što se za neko fiksirano k posmatra k najbližih suseda (u odnosu na neku metriku) instance za koju želimo da izračunamo vrednost zavisne promenljive. Ako je u pitanju regresija, dodelićemo joj aritmetičku sredinu vrednosti zavisne promenljive posmatranih najbližih k suseda, a ako je u pitanju klasifikacija, dodelićemo joj onu klasu koja je najzastupljenija medju posmatranim najbližim susedima.