Pregled gradiva

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?**

   Najpre se odlučujemo koju ćemo metodu mašinskog učenja koristiti. To radimo na osnovu toga da li je u pitanju problem regresije ili klasifikacije, kao i na osnovu veličine dataset-a. Naravno, možemo izabrati više različitih metoda, napraviti više modela, pa ih uporediti na validacionom ili test skupu. Prvi korak u izgradnji modela je izbor atributa koje ćemo koristiti. Ako imamo sirove podatke, možda će biti neophodno da konstruišemo atribute od sirovih podataka (neuronske mreže u velikoj meri same rade taj korak). Takođe, često je potrebno kategoričke atribute transformisati u numeričke. Zatim se po potrebi radi standardizacija podataka i nakon toga prelazi se na treniranje modela.

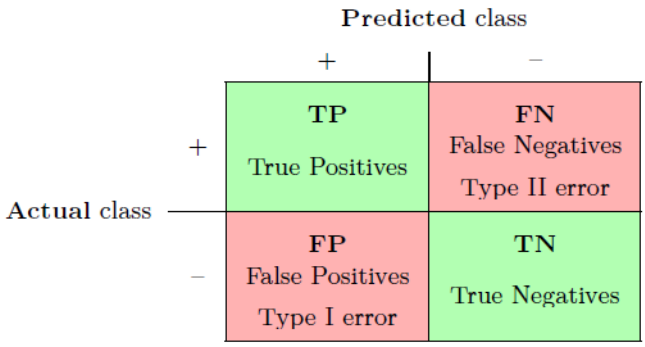
3. **U kom odnosu se dele podaci na skupove za trening, validaciju i testiranje? I čemu služi svaki od skupova?**

   Ne postoji jedno univerzalno pravilo, ali se mali i srednji skupovi podataka često dele u odnosu 70:15:15 ili 80:10:10. Trening skup služi za pravljenje modela, odnosno za dobijanje optimalnih parametara modela za fiksirane hiperparametre. Validacioni skup služi za izbor hiperparametara i za poređenje više modela. Test skup služi za konačnu evaluaciju modela.

4. **Šta je unakrsna validacija?**

   Unakrsna validacija je postupak u kome se trening skup deli na više delova. Model se više puta trenira, pri čemu se svaki put jedan deo koristi za validaciju, a preostali delovi za trening. Na kraju se posmatra prosečna preciznost kroz sve te podele. Ovaj postupak je naročito koristan kada nemamo mnogo podataka, jer omogućava pouzdaniju procenu kvaliteta modela nego jedna jedina podela na trening i validacioni skup.

5. **Kako izgleda matrica konfuzije?**



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

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

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

   sensitivity (true positive rate, recall) $= \frac{TP}{TP+FN}$

   specificity (true negative rate) $= \frac{TN}{TN+FP}$

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

   AUC (površina ispod ROC krive).

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

   Kod nebalansiranih podataka accuracy često nije dobra mera, pa se češće koriste precision, sensitivity, F_1 score i AUC.

8. **Koja je razlika između 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 pas, a regresija se koristi kada je zavisna promenljiva neprekidnog tipa, na primer kada hoćemo da predvidimo temperaturu vazduha ili broj prodatih proizvoda.

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

   Standardizacija podataka je proces koji se sastoji u sledećim koracima. Posmatramo neki fiksirani prediktor, odnosno neku odabranu kolonu u podacima. Prvo računamo uzoračku sredinu i standardnu devijaciju te kolone na trening skupu. Zatim od vrednosti u toj koloni na trening, validacionom i test skupu oduzimamo izračunatu uzoračku sredinu i delimo sa izračunatom uzoračkom standardnom devijacijom. Na taj način dobijamo da su vrednosti tog prediktora centrirane oko nule i da imaju standardnu devijaciju 1.
   Ovaj proces je naročito važan kada nam je bitno da svi prediktori budu na istoj skali, na primer kada koristimo euklidsku metriku, *lasso* ili *ridge* regularizaciju, kao i kod neuronskih mreža. Kod modela zasnovanih na stablima standardizacija često nije neophodna.

10. **Šta je data leakage?**

    Data leakage je pojava kada tokom pravljenja modela koristimo informacije koje u realnoj primeni ne bismo imali na raspolaganju. Na primer, do curenja informacija dolazi ako standardizaciju radimo koristeći ceo skup podataka umesto samo trening skupa, ili ako na bilo koji način informacije iz test skupa utiču na treniranje modela. Data leakage dovodi do toga da model na papiru izgleda bolje nego što zaista jeste.

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

    Overfitting je pojava kada se model previše prilagodi podacima iz trening skupa. 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.
    - Korišćenje više podataka, ako su dostupni.

12. **Koja je razlika između parametara i hiperparametara?**

    Parametri modela su veličine koje model uči iz trening podataka. Na primer, kod linearne ili logističke regresije to su koeficijenti uz prediktore. Hiperparametri su veličine koje biramo pre treniranja modela i koje se ne uče direktno iz podataka. Na primer, kod kNN-a hiperparametar je broj

suseda k, a kod stabla odlučivanja maksimalna dubina stabla. U praksi se parametri dobijaju treniranjem modela, a hiperparametri se biraju pomoću validacionog skupa.

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

    U slučaju regresije jedna od najčešće korišćenih funkcija greške je Mean Squared Error (MSE), definisana sa

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

    U slučaju klasifikacije često se koristi unakrsna entropija (Cross-Entropy Loss), koja kažnjava model kada daje malu verovatnoću tačnoj klasi. U slučaju višeklasne klasifikacije ona se može zapisati kao

    $$L = -\sum_{j=1}^{n}\sum_{i=1}^{k} y_{ji}\log(\hat{p}_{ji}),$$

    gde je k broj mogućih kategorija, y_{ji} indikator da li j-ta instanca pripada i-toj klasi, a $\hat{p}_{ji}$ procenjena verovatnoća da j-ta instanca pripada i-toj klasi.

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

    Neuronske mreže se često koriste kada imamo dovoljno veliki skup podataka. Ako imamo mali skup podataka, druge metode često postižu bolje rezultate. Takođe, ako su podaci sirovi, tj. neobrađeni, neuronske mreže često postižu bolje rezultate nego ostale metode, jer u velikoj meri same konstruišu reprezentacije podataka na osnovu kojih vrše klasifikaciju ili regresiju.

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

    Potpuno povezane neuronske mreže koriste se pre svega za rad sa tabelarnim podacima, ali i kao sastavni deo složenijih arhitektura.
    Konvolutivne neuronske mreže koriste se za obradu signala, poput slike i zvuka, a mogu se koristiti i za neke druge tipove podataka.
    Rekurentne neuronske mreže koriste se za obradu sekvencijalnih podataka, kao što su obrada prirodnog jezika i vremenske serije. Rekurentne mreže se danas koriste ređe nego ranije, jer ih u mnogim zadacima zamenjuju transformeri koji često postižu bolje rezultate.

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

Prva prednost je specijalizovanost za topologiju signala. Naime, i potpuno povezane mreže bi mogle da se primenjuju na probleme vezane za zvuk, slike i slično, ali kod njih bi raspored piksela (ako radimo obradu slike) bio potpuno proizvoljan, pošto ne uzimaju u obzir susednost piksela na slici, dok su konvolutivne mreže konstruisane imajući u vidu da to što su pikseli jedni u okolini drugih ima poseban značaj. Druga prednost su proređene interakcije, pod čime se podrazumeva da je svaka jedinica povezana samo sa malim brojem jedinica iz prethodnog sloja, umesto sa svim, kao u slučaju potpuno povezanih mreža. Na taj način se smanjuje broj parametara modela. Treća prednost je deljenje parametara, tj. sve jedinice jednog kanala imaju iste parametre definisane filterom tog kanala. Kada parametri ne bi bili deljeni, kao kod potpuno povezane mreže, učeći različite parametre za različite delove slike, mreža bi učila da nekim delovima ulaza pridaje posebnu semantiku. To može zvučati dobro, ali npr. u slučaju detekcije lica na slici, ne bi bilo dobro da se nauči da nos treba da bude baš na sredini slike, jer bi mogao biti i na nekom drugom mestu. Deljenje parametara omogućava da se nauči filter koji traži nos bilo gde na slici.

17. **Koje su neke savremene arhitekture i pravci razvoja neuronskih mreža?**

Pored potpuno povezanih, konvolutivnih i rekurentnih mreža, danas su veoma značajni i modeli zasnovani na mehanizmu pažnje (*attention*). Ovaj mehanizam omogućava modelu da utvrdi koji delovi ulaza su najvažniji za dati zadatak i da im posveti veću pažnju. Na toj ideji zasnovani su transformeri, koji danas imaju veoma važnu ulogu u obradi prirodnog jezika, a često se koriste i u drugim oblastima.
Grafovske neuronske mreže koriste se kada podaci imaju strukturu grafa, na primer kod društvenih mreža, saobraćajnih mreža ili molekula. Njihova osnovna ideja je agregacija informacija iz susedstva svakog čvora. Neki modeli iz ove klase, kao što su *Graph Attention Networks* (GAT), koriste i mehanizam pažnje za određivanje značaja susednih čvorova, ali attention nije sastavni deo svih grafovskih neuronskih mreža.
Generativne suparničke mreže (GAN) koriste se za generisanje novih podataka, najčešće slika. Sastoje se od dve mreže: generatora, koji pokušava da proizvede što realističnije podatke, i diskriminatora, koji pokušava da razlikuje generisane podatke od stvarnih. Tokom treninga generator postaje sve bolji u stvaranju uverljivih uzoraka, a diskriminator u njihovom prepoznavanju. GAN-ovi se i dalje koriste u nekim specifičnim zadacima, kao što su generisanje slika i pravljenje sintetičkih podataka. Ipak, u mnogim savremenim generativnim zadacima glavnu ulogu imaju difuzioni modeli, dok modeli zasnovani na transformerima dominiraju u generisanju i obradi teksta i imaju veoma važnu ulogu i u multimodalnim sistemima.
Veliki jezički modeli (LLM) predstavljaju veoma velike modele zasnovane na transformerima, trenirane na ogromnim količinama tekstualnih poda-

taka. Oni se koriste za obradu i generisanje prirodnog jezika i mogu da generišu tekst, odgovaraju na pitanja, rezimiraju sadržaj i obavljaju mnoge druge zadatke.

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

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

19. **Koji algoritmi se koriste za smanjivanje dimenzije 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. Više o UMAP algoritmu možete pogledati na sledećem linku.

20. **Ukratko objasniti logističku regresiju?**

    Logistička regresija se koristi za binarnu klasifikaciju, tj. kada zavisna promenljiva ima samo dve moguće klase. Ovaj metod se sastoji u tome da se za ulazni vektor x izračuna vrednost logističke funkcije

    $$p(x) = \frac{1}{1 + e^{-(\beta_0 + \beta^T x)}},$$

    gde su β_0 i β parametri modela dobijeni minimizovanjem odgovarajuće funkcije gubitka. Vrednost $p(x)$ se može tumačiti kao procenjena verovatnoća da instanca pripada klasi 1. Ako je ta verovatnoća veća od izabranog threshold-a, dodeljuje se klasa 1, a u suprotnom klasa 0. Threshold je najčešće postavljen na 0.5, ali može biti i neki drugi broj.

21. **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 izabranu 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 među posmatranim najbližim k susedima.