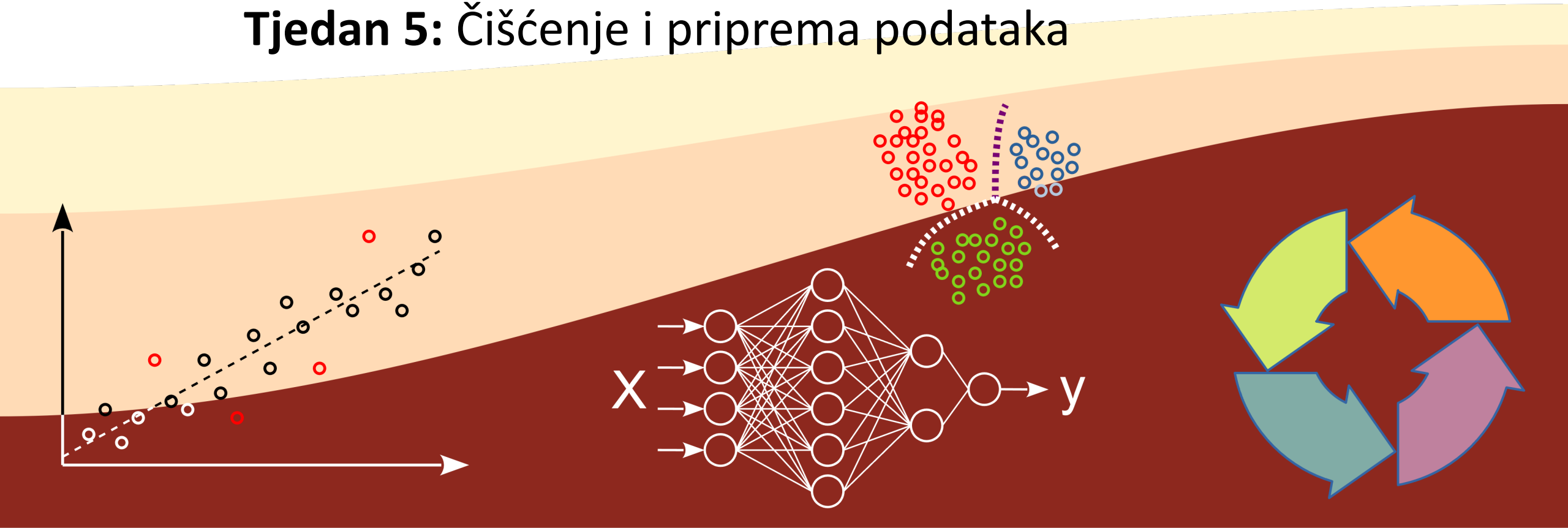
Arhitektura i Razvoj Inteligentnih Sustava

Tjedan 5: Čišćenje i priprema podataka



Creative Commons

- slobodno smijete:
  - dijeliti — umnožavati, distribuirati i javnosti priopćavati djelo
  - prerađivati djelo
- pod sljedećim uvjetima:
  - imenovanje: morate priznati i označiti autorstvo djela na način kako je specificirao autor ili davatelj licence (ali ne način koji bi sugerirao da Vi ili Vaše korištenje njegova djela imate njegovu izravnu podršku).
  - nekomercijalno: ovo djelo ne smijete koristiti u komercijalne svrhe.
  - dijeli pod istim uvjetima: ako ovo djelo izmijenite, preoblikujete ili stvarate koristeći ga, preradu možete distribuirati samo pod licencom koja je ista ili slična ovoj.

Osnovni postupci u pripremi podataka

- Glavni problemi s kojima se susrećemo
  - Podaci nedostaju – NULL vrijednosti
  - Kategoričke vrijednosti
  - Višestruke instance
  - ”Loša statistika”
    - Značajke s malo jedinstvenih vrijednosti
    - Niska varijanca – značajke koji ne doprinose klasifikaciji
  - Značajke koje koreliraju
  - Skaliranje značajki
  - Disbalans klasa

NULL vrijednosti

- U vremenskim serijama
  - *Backward* i *forward filling* – uzimanje prethodne vrijednosti u vremenskoj seriji ili slijedeće ako prethodne nema
- U klasičnim skupovima podataka
  - Medijan značajke
  - Aritmetička sredina značajke
- Uklanjamo iz skupa podataka
  - Ako su sve značajke instance NULL
  - Ako je značajka NULL u svim instancama

NULL vrijednosti

- Koliko nam je vrijedna određena značajka?
  - Popunjenost po instancama
  - <15% NULL vrijednosti ima smisla nadopunjavati s *backward fill* i *forward fill* strategijama
  - Sve više od 15% – možda je bolje koristiti strategiju s medijanom ili aritmetičkom sredinom

Višestruke instance

- Potrebno je razumjeti statistiku skupa podataka
  - Ako je dupli zapis individualan, vjerojatno se radi o grešci – uklonimo zapis
  - Ako postoje višestruki zapisi koji pripadaju određenoj klasi
    - Treba ocijeniti da li više instanci klase doprinosi utvrđivanju klasifikacije
    - Recimo kod algoritama za grupiranje (*clustering*) ima smisla zadržati sve instance
    - Kod ANN nema
- Treba poznavati i poslovnu semantiku podataka

Kategoričke vrijednosti

- Korištenjem *label encodera* pretvaramo ih u diskretne vrijednosti
  - Koje se mogu koristiti od raznih algoritama
    - Uglavnom skup $\mathbb{N}$
  - Ne treba se bojati diskretizacije prostora značajki
  - Ne smijemo upasti u problem niske varijance
    - Recimo „spol”: muški / ženski
- scikit-learn nudi takve enkodere – recimo *LabelEncoder*

Značajke s „lošom statistikom”

- Niska varijanca
  - Premalo razlike između vrijednosti značajke svih instanci
  - Skaliranjem bi se ta razlika mogla pojačati?
  - Ne i za mali broj kategoričkih vrijednosti
    - Muško / žensko, dan / noć, da / ne i slično...
  - Možemo odbaciti značajke koje imaju recimo varijancu ispod 0.15
- Možemo provjeriti koliko jedinstvenih vrijednosti imamo u određenoj značajki
  - Odbacimo značajke koje imaju mali broj jedinstvenih vrijednosti

Korelacija

- Provjerimo da li postoji značajna korelacija između značajki
  - Ako značajke koreliraju – postoji neka uzročno-posljedična veza između njih
  - Pitanje donje granice korelacije – što smatramo kao dovoljno zavisne značajke
    - 1 = vrijednost značajki je jednaka za svaku instancu – u relacijskoj algebri se to zove funkcijska zavisnost
  - Klasifikacija bi trebala imati dobru metriku i bez jedne od tih značajki
    - Niža varijanca – postoji veća šansa da neke instance u dimenziji s višom varijancom bude drukčije klasificirana
  - Ako značajke uopće ne koreliraju?
    - Efekt se vidi u težinskim matricama ANN
    - Univarijatna vs. multivarijatna statistika

Skaliranje

- Zbog gradijentnih metoda optimizacije
  - Idealno, sve značajke slično skalirane
  - Standardno skaliranje

$$(v_i - \mu)/\sigma$$

    - Od vrijednosti značajke u instanci i se oduzme aritmetička sredina značajke i ta se razlika podijeli sa standardnom devijacijom značajke
  - Min-max skaliranje
    - Minimalna vrijednost značajke je 0
    - Maksimalna vrijednost značajke je 1
    - Sve ostale vrijednost skalirano popunjavaju taj prostor
- scikit-learn nudi klase za ova skaliranja
  - StandardScaler
  - MinMaxScaler

Disbalans klasa

- Broj instance jedne klase je značajnije veći od druge klase
  - Recimo 5% prema 95%
  - Podešavanje granica klase postaje problem
    - Često rezultira sa velikim FP (*false-positive*) ili FN (*false-negative*) vrijednostima
  - Posebni problem kod podržanog učenja (*reinforcement*)
- Rješenje je sintetički *upsampling* ili *downsampling*
  - Recimo SMOTE (*synthetic minority oversampling*)
  - python modul *imblearn*
  - Neki algoritmi imaju mogućnost definiranja težinske vrijednosti klase (*class weight*)
    - Ovo se može definirati u optimizatorima pytorcha, u kerasu i slično

Velike količine podataka i skalabilnost

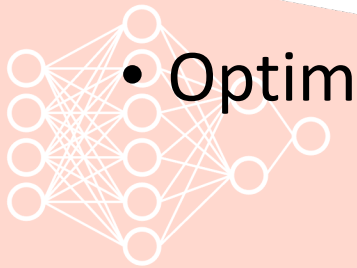
- Osnova su *clusteri* za obradu podataka
  - Apache Spark i Flink
    - Mikro-*batch* naprama prava obrada toka podataka
  - Map / reduce koncept – jedan čvor je *master*, ostali čvorovi *workeri*
  - Koncepti podijeli pa vladaj – particioniranje
    - Horizontalno particioniranje
      - skup podataka se podijeli na više particija redova
    - Vertikalno particioniranje
      - skup podataka se podijeli na više atributa (značajki) koje imaju neke funkcijske zavisnosti

Odabir algoritama

- Vremenske serije
  - Nestrukturirano – audio recimo, tekst, itd...
  - Ne želimo *sparse* podatke – idealno bi trebalo biti manje od 15% praznih vrijednosti
  - Koristimo *backward fill* i *forward fill* za popunjavanje
  - Klasične metode
    - LSTM – Long Short-Term Memory – vrsta rekurzivne mreže
    - GRU – Gated recurrent unit – novija vrsta LSTM-a
    - klasični RNN
  - MLP
  - Konvolucijske neuronske mreže (CNN)

Podaci s labelama

- Tip labele
  - binarna klasifikacija – ili instanca pripada klasi ili ne
  - klasifikacija s više klasa
- Klasifikatori s oštrim izlazima (*crisp*)
  - funkcija tranzicije igra bitnu ulogu
  - više klasa
    - svaka klasa ima svoj izlaz koji je binarnog tipa
- Klasifikatori sa softmax izlazima (vjerojatnost da ulaz pripada jednoj klasi)
  - uzima se klasa s najvišom vjerojatnošću
- Optimizatori: SGD, Adam



Podaci s kontinuiranom vrijednošću cilja

- Neuronske mreže kao regresori
  - Standardni podaci s kontinuiranim izlazima
  - Ne želimo iskriviti linearnu funkciju neurona uvođenjem tranzicijske funkcije
  - Optimizatori: MSELoss
- Standardni regresori
- *Random forest,* stabla odluke
- Algoritmi za grupiranje (*clustering*)
- SVM (*support vector machine*)

Nestrukturirani podaci

- Audio, video, slike
  - Imamo vremenske serije – audio i video
  - Konvolucijske mreže (CNN)
  - Može i standardni višeslojni perceptron

Hibridni pristup

- Učenje strukturiranog ulaza s nestrukturiranim
  - Spojimo mreže i dodamo slojeve za izlaz
- Vremenska serija i standardni pristup
- Više različitih prostora značajki koje trebaju završiti jednom klasifikacijom
- Spoj binarnih labela i kontinuiranog cilja
- Posebne arhitekture koje se prilagođavaju podacima koje imamo
- Moduli
  - pytorch, keras, tensorflow, ...