Istraživanje podataka 1

Vežbe 6

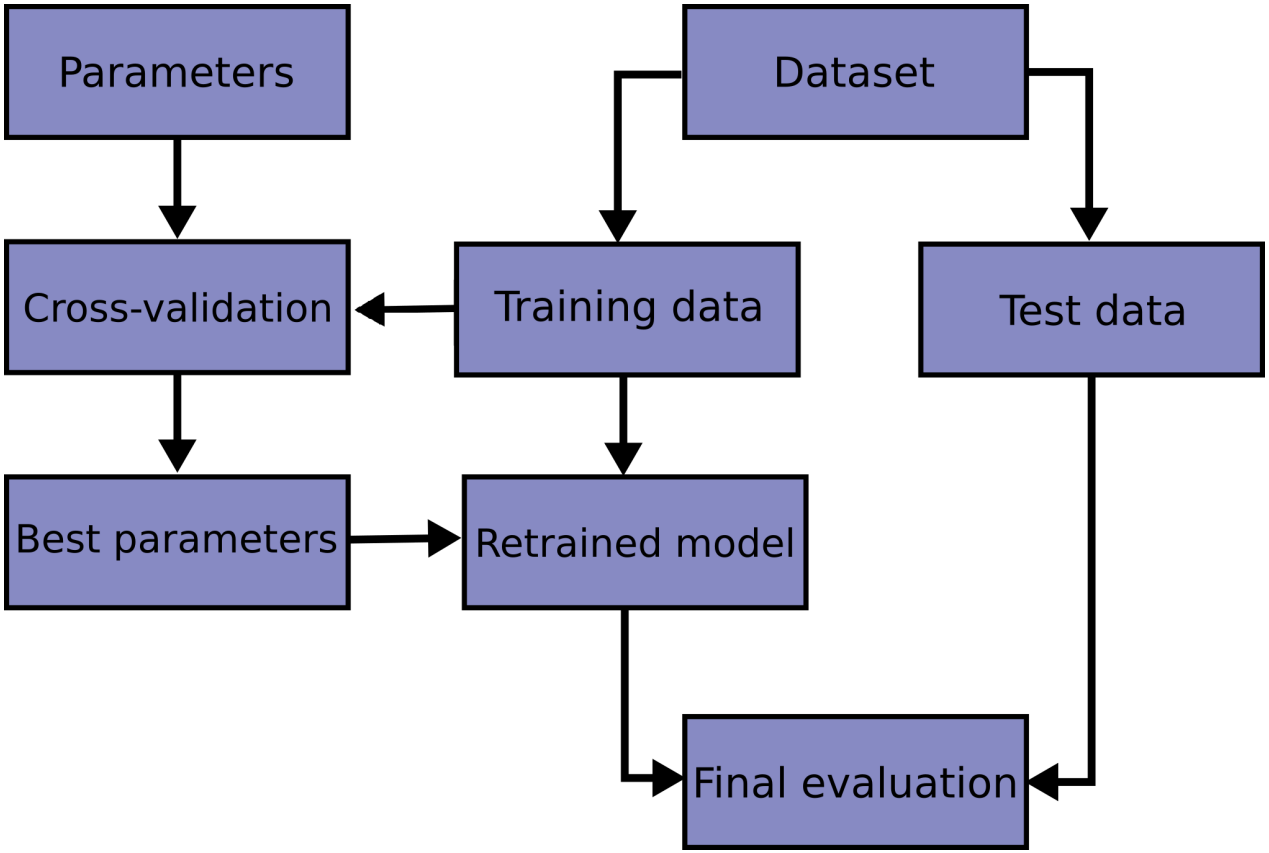
Outline

Outline

Podela skupa

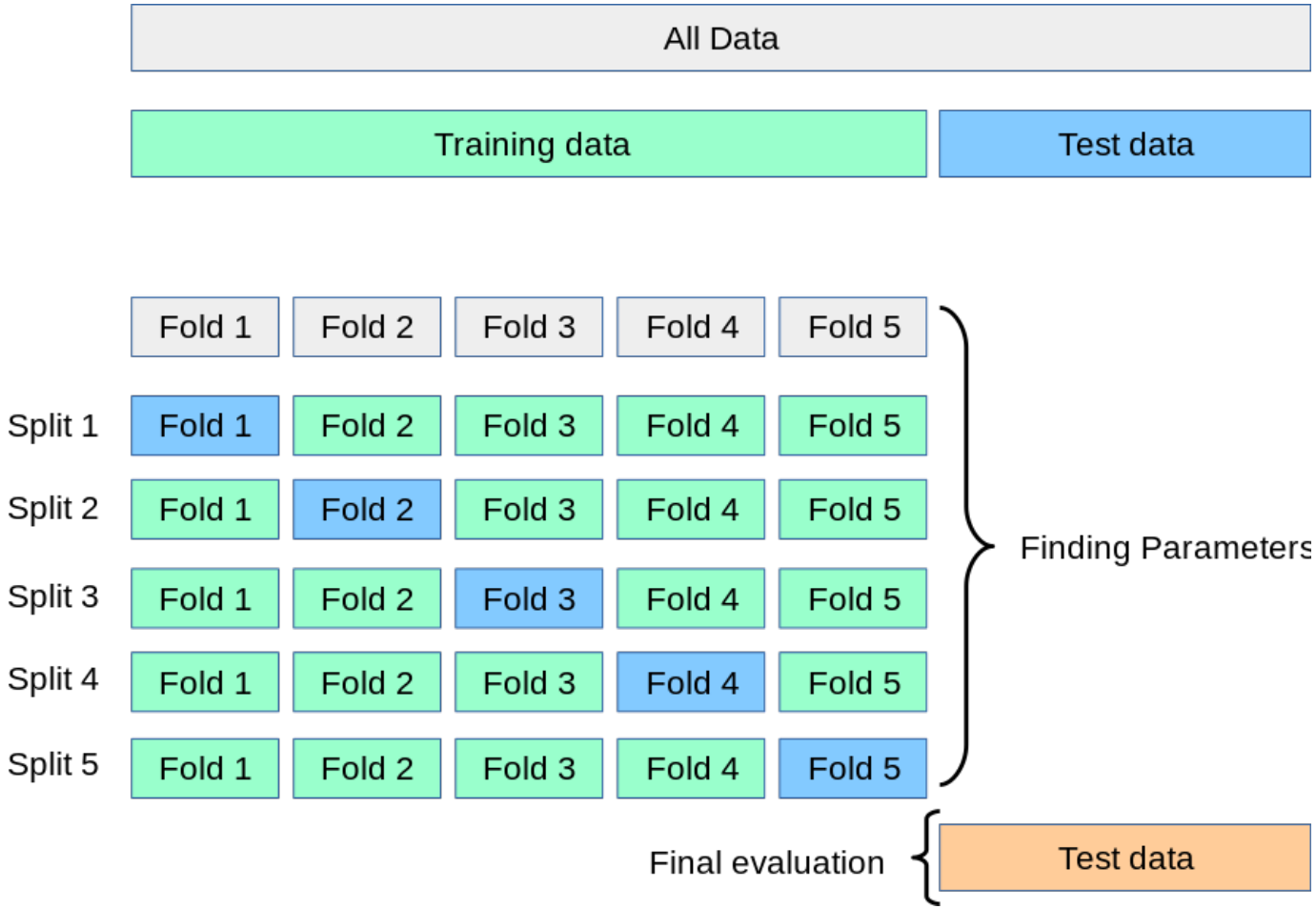
- Trening i test deo
- Trening deo, test deo i deo za validaciju

Unakrsna validacija



- Podešavanje hiper-parametara procenjivača

Unakrsna validacija



Unakrsna validacija

- sklearn.model_selection.GridSearchCV
- iscrpna pretraga korišćenjem zadatih vrednosti parametara za procenjivača
- parametri
  - *estimator* - procenjivač (npr. KNeighborsClassifier)
  - *param_grid* - rečnik ili lista rečnika sa definisanim mogućim vrednostima za parametre procenjivača
  - *scoring* - mera za proveru modela
  - *cv* - broj podskupova za unakrsnu validaciju, default= 3-fold
  - *refit* - da li ponovo napraviti model sa najboljim parametrima nad celim skupom podataka. Da bi mogla da se radi predikcija nad drugim skupom potrebno je staviti True. (default=True)

Unakrsna validacija

- atributi
  - *cv_results_* - rečnik sa podacima o zadatim parametrima i rezultatima
  - *best_estimator_* - najbolji klasifikator
  - *best_score_* - najbolji skor
  - *best_params_* - parametri koji daju najbolji rezultat
  - *scorer_* - funkcija za skor

Unakrsna validacija

- metode
  - *fit(x,y)* - pravljnje modela sa optimalnim parametrima na osnovu skupa (x,y)
  - *predict(x)* - predviđanje klase za test instance

Outline

Algoritam K najbližih suseda (KNN)

klasifikacija instance je zasnovana na sličnosti sa drugim instancama.
Neka je:

- k zadati broj suseda
- D skup instanci za treniranje; svaka instanca je predstavljena sa (x, y) gde su x vrednosti atributa za predviđanje, a y klasa kojoj instanca pripada
- svaka test instanca z je predstavljena sa (x', y') gde su x' vrednosti atributa za predviđanje, a y' dodeljena klasa

Algoritam K najbližih suseda (KNN)

1: **for all** $z = (x', y')$ **do**
2: izračunatu $dist(x', x)$ (rastojanje između instance z i svake instance $(x, y) \in D$)
3: izdvojiti $D_z \subseteq D$, skup k najbližih trening instanci test instanci z
4: $y' = \arg\max_v \sum_{(x_i, y_i) \in D_z} I(v = y_i)$
5: **end for**

Preprocesiranje

- sklearn.preprocessing
- svaki atribut se zasebno obrađuje
- standardizacija
  - *scale(X)* - vraća standardizovane vrednosti
  - *StandardScaler* - klasa za standardizaciju
    - *fit(x)*
    - *transform(x)*
    - *inverse_transform(x)*
    - *fit_transform(x)*

Pretprocesiranje

- skaliranje na određeni interval
  - *MinMaxScaler(feature_range=(0, 1))* - klasa za skaliranje
    - *fit(x)*
    - *transform(x)*
    - *fit_transform(x)*
    - *inverse_transform(x)*

Klasifikacija - KNN

- sklearn.neighbors.KNeighborsClassifier
- parametri
  - *n_neighbors* - broj suseda (default = 5)
  - *p* - stepen za rastojanje Minkovskog
  - *metric* - izbor metrike za rastojanje, (default 'minkowski'), moguće vrednosti su definisane u sklearn.neighbors.DistanceMetric
  - *weights* - težina suseda
    - *uniform* - svi susedi imaju istu težinu (default)
    - *distance* - težina suseda je obrnuta rastojanju - najbliži sused ima najveću težinu, najudaljeniji sused ima najmanju težinu

Klasifikacija - KNN

- metode
  - *fit(x)* - pravi model
  - *predict(x)* - određuje klase instancama
  - *kneighbors(x)* - vraća rastojanje i poziciju suseda

Klasifikacija - KNN

- Preprocesiranje
  - Normalizacija numeričkih atributa

$$x' = \frac{2 * (x - x_{min})}{x_{max} - xmin} - 1$$

  - Kodiranje kategoričkih atributa primenom kodiranja *jedan-od-c*, gde je c broj kategorija. Vrednost se kodira kao binarni vektor dimenzije c. Prva kategorija ima vrednost (1,0,...,0), druga (0,1,0,...,0), a poslednja (0,0,...,1).

Klasifikacija - KNN

- Metrike za računanje rastojanja između instance koja se klasifikuje i suseda
  - (Težinsko) Euklidsko rastojanje
  - (Težinsko) Menhetn rastojanje
  - težina atributa se računa prema značajnosti atributa

$$w_i = \frac{FI_i}{\sum_{p=1}^{d} FI_p}$$

Klasifikacija - KNN

- Unakrsna validacija za određivanje najboljeg k iz opsega $[k_{min}, k_{max}]$
  - bira se k sa najmanjom prosečnom stopom greške

Klasifikacija - KNN

- Izbor atributa (izbor unapred)
  - moguće zadavanje obaveznih atributa
  - u svakom koraku se bira po jedan atribut koji nije među izabranim atributima i koji najviše doprinosi smanjenju greške stope. Postupak se ponavlja dok se ne zadovolji jedan od kriterijuma za zaustavljanje.

Klasifikacija - KNN

- Instance sa nedostajućim vrednostima se ne uzimaju u obzir
- Instanca se klasifikuje prema glasovima k najbližih suseda - dodeljuje se klasa sa najviše glasova.