Računarska statistika

Snježana Lubura Strunjak

Zagreb, 3. lipnja 2021.

Metode ponovnog uzorkovanja (re-sampling)

Podjele podataka (Data partitioning)

Pretpostavimo da su naši podaci potekli iz slijedećeg (općeg) statističkog modela (za učenje)

$$Y = f(X) + \epsilon,$$

gdje slučajna pogreška ϵ ima sredinu $\mathbb{E}(\epsilon) = 0$, i nezavisna je od X.
X–prediktori, nezavisne varijable (INPUTS u strojnom učenju).
Y–kriteriji, zavisne varijable (OUTPUTS ili TARGETS u strojnom učenju)

- Cilj statističkog učenja je naći korisnu aproksimaciju $\hat{f}(x)$ funkcije $f(x)$
- Podjela podataka se koristi za određivanje sposobnosti generalizacije (tj. prediktivne moći na nezavisnim test podacima) modela ili metode (za učenje).

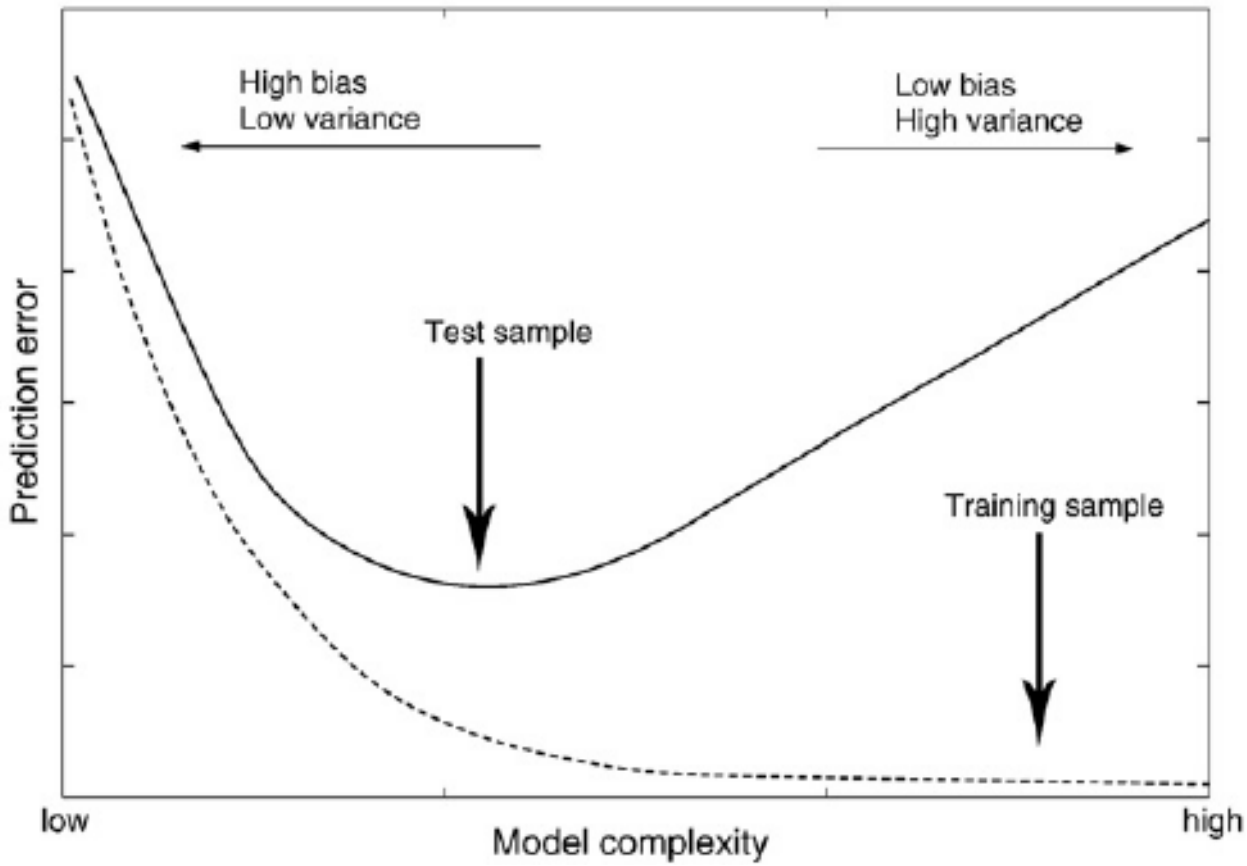
Podjele podataka (Data partitioning)

Tipična podjela uzorka na tri manja dijela: $50\%, 25\%, 25\%$ (nije pravilo!). Tipična metoda podjele: jednostavan slučajan uzorak (ili stratificiran- ako je npr. zastupljenost neke vrijednosti u početnom uzorku 10%, tražimo da zastupljenost i u malim uzorcima bude ista, itd.)
Uzorak se dijeli na tri dijela zbog tri osnovna koraka:

- Training - Razvoj modela (tj.aproksimiranje funkcije) - tome nam služi prvi dio početnog uzorka,
- Validation - Izbor modela (usporedba ponašanja različitih aproksimacija sa ciljem odabira (aproksimativno) najboljeg modela) - tome nam služi drugi dio početnog uzorka,
- Test - Ocjenjivanje modela (nakon odabira završnog modela, procjenjuje se pogreška predikcije (pogreška generalizacije) na novim, nezavisnim podacima) - tome nam služi treći dio početnog uzorka.

Podjele podataka (Data partitioning)

Slika: C. Ballabio, Spatial prediction of soil properties in temperate, Geoderma 151 (2009) 338-350



Krosvalidacija (Cross-validation)

Metoda za procjenjivanje pogreške predikcije (za izbor modela) u situacijama kada nema dovoljno podataka za podjelu na 3 dijela.
Algoritam:

- Podijeli podatke u K podjednakih dijelova.
- Za $k = 1, 2, \ldots, K$:
  - Izbaci k-ti dio
  - Fitaj model na ostalih $K - 1$ dijelova
  - Izračunaj pogrešku predikcije (pogreška predikcije je najčešće kvadrat od razlike između stvarne vrijednosti i procijenjene) fitanog modela primjenjenog na predikciju k-tog dijela.
- Združi K procjena pogreške predikcije na način da se izračuna procjena pogreške krosvalidacijom, u oznaci CV, kao srednja vrijednost svih dobivenih procjena pogrešaka.

Krosvalidacija (Cross-validation)

Kako odabrati vrijednost za K?

- $K = N$ (tj. "leave-one-out")
  - CV ima nisku pristranost, ali visoku varijancu
  - Računarski intenzivno
  - U nekim situacijama (npr.u linearnoj regresiji) računanje je jednostavno ("H" matrica)
- Za manji K (npr. 5) CV ima nižu varijancu, ali može imati visoku pristranost.

Diskriminativna analiza: procjena pogreške klasifikacije po grupama krosvalidacijom

Važno je prepoznati grupe u uzorku koje se sastoje od sličnih opservacija jer na taj način bolje razumijemo same podatke i fenomene koji su opisani takvim podacima.
Ako su podaci podijeljeni u grupe, onda diskriminativnom analizom dolazimo do odgovora po čemu se te grupe razlikuju i razvijamo metodu klasificiranja pojedine opservacije u neku od grupa.

Diskriminativna analiza: procjena pogreške klasifikacije po grupama krosvalidacijom

Ciljevi diskriminativne analize:

- Interpretacija: "Kako se grupe razlikuju?" Naći i interpretirati linearne kombinacije varijabli koje optimalno predviđaju grupne razlike.
- Klasifikacija: "Koliko se točno mogu observacije klasificirati u grupe?" Primjenom funkcija varijabli predviđa se pripadnost pojedinoj grupi i procjenjuje pogreška.

Diskriminativna analiza: procjena pogreške klasifikacije po grupama krosvalidacijom

Podjela metoda diskriminativne analize (DA):

- Klasična Fisherova diskriminativna analiza (FDA)
  - Linearni model (za jednake matrice kovarijanci po grupama)
  - Kvadratni model (za nejednake matrice kovarijanci po grupama)
- Kanonička diskriminativna analiza (KDA)
- Neparametarske metode

Linearna diskriminativna analiza (Fisherova DA)

Izvedite primjer Example 37.4: Linear Discriminant Analysis of Remote-Sensing Data on Crops (SAS Help)
Kolika je pogreška predikcije u slučaju kada se radi diskriminativna analiza bez krosvalidacije, a kolika s krosvalidacijom?
(Za bolje razumijevanje ispisa programa pročitati slide-ove 64-85 u materijalima Kvantitativne metode VLS PMF Matematika.pdf u folderu PDF materijali na Merlinu)

Re-sampling metode: zaključak

- Jackknife i Bootstrap metode su (primarno) metode za mjerenje statističke točnosti.
  - Jackknife se provodi sistematičnim uzorkovanjem iz podataka
  - Bootstrap se provodi
    - Slučajnim uzorkovanjem iz podataka ili
    - Slučajnim uzorkovanjem po procijenjenom (fitanom) modelu
- Podjela podataka i krosvalidacija su (primarno) metode za mjerenje pogreške predikcije.
  - Podjela podataka se primjenjuje u situacijama sa dovoljno podataka za slučajno dijeljenje na 3 dijela: za učenje, validaciju i test.
  - Krosvalidacija se primjenjuje u situacijama sa nedovoljno podataka za dijeljenje na K dijelova iste veličine: $(K-1)$ za učenje, jedan za test.