Evaluarea modelelor

Recapitulare: învățarea supervizată

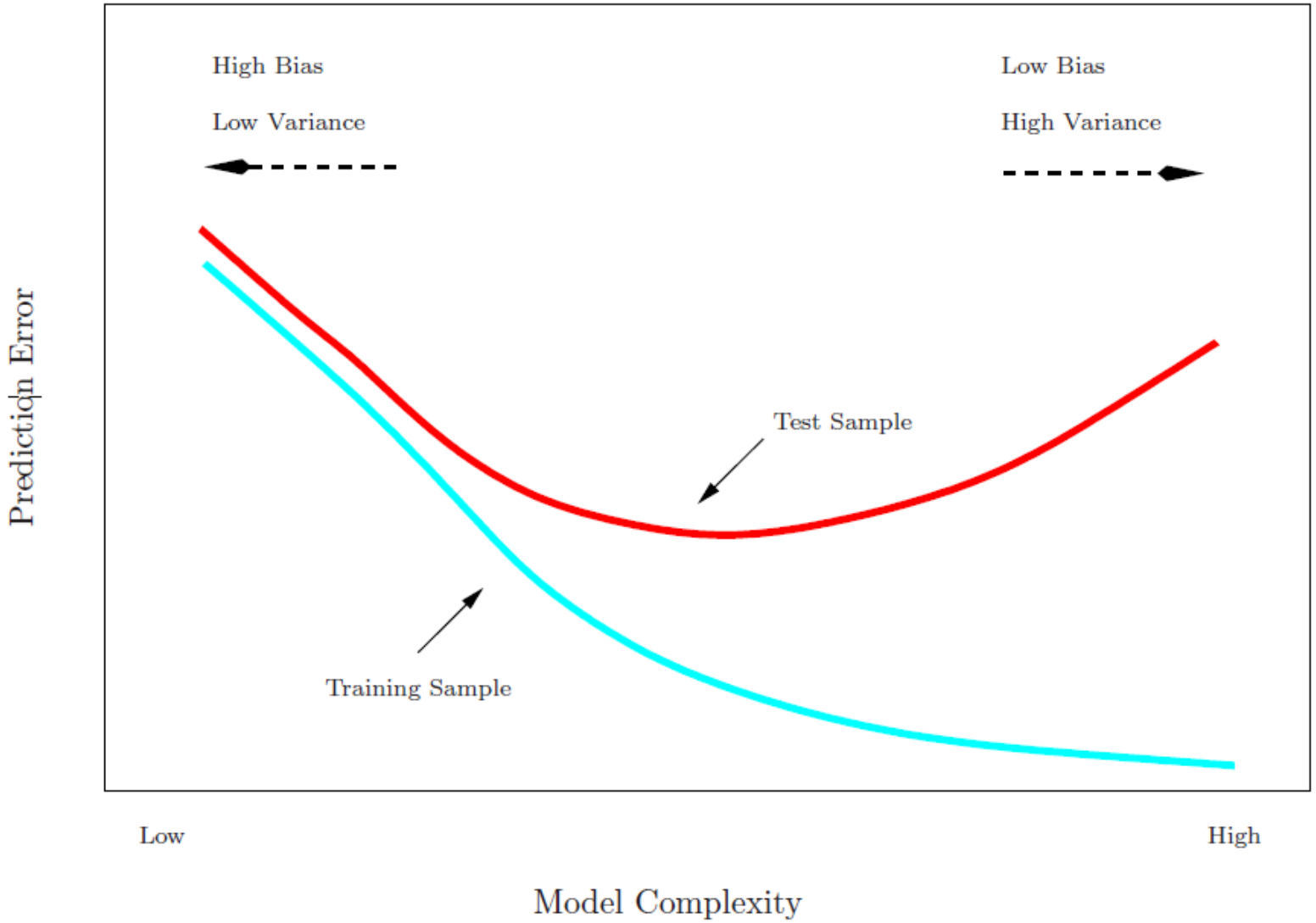
- Scop
  - Furnizarea unei ieșiri corecte pentru o nouă intrare
- Definire
  - Se dă un set de date (exemple, instanțe, cazuri)
    - Date de antrenament – sub forma unor perechi (atribute_data_i, iersire_i) unde
      - i=1,n (n=nr. datelor de antrenament)
      - atribute_date_i =(atr_i1, atr_i2,...., atr_im), m – nr atributelor (caracteristicilor, proprietăților) unei date
      - Iesire_i
        - o categorie dintr-o mulțime dată (predefinită) cu k elemente (k-nr de clase) – problemă de clasificare
        - un număr real – problemă de reresie
    - Date de test – sub forma (atribute_data_i), i =1,t (t = nr. datelor de test)
  - Să se determine
    - O funcție (necunoscută) care realizează corespondența atribute-ieșire pe datele de antrenament
    - Ieșirea (clasa/valoarea) asociată unei date de test folosind funcția învățată pe datele de antrenament
- Proces în 2 etape
  - Antrenarea
    - Învățarea, cu ajutorul unui algoritm, a modelului de clasificare
  - Testarea
    - Testarea modelului folosind date de test noi (unseen data)

Calitatea învățării

- Definire
  - O măsură de performanță a algoritmului
    - Ex. Acuratețea (nr de exemple corect clasificate / nr total de exemple)
  - Calculată în
    - Faza de antrenare
    - Faza de testare
- Metode de evaluare
  - Seturi disjuncte de antrenare și testare
    - Setul de antrenare poate fi împărțit în date de învățare și date de validare
    - Setul de antrenare este folosit pentru estimarea parametrilor modelului (cei mai buni parametri obținuți pe validare vor fi folosiți pentru construcția modelului final)
    - Pentru seturi mari de date
  - Cross-validation: Validarea încrucișată pe mai multe (k) sub-seturi egale ale datelor (de antrenament)
    - Separarea datelor de k ori, k-1 sub-seturi pt învățare și 1 sub-set pt validare
    - Dimensiunea unui sub-set = dimensiunea setului / k
    - Performanța este dată de media celor k rulări (ex. K=5 sau k=10)
    - Pentru date puține
    - Leave-one-out cross-validation
      - Similar validării încrucișate dar k = nr de date (un sub-set conține un singur exemplu)
      - Pentru date foarte puține
  - Bootstrap
- Dificultăți:
  - **Over-fitting** (învățare „pe de rost„) – performanță bună pe datele de antrenament, dar foarte slabă pe datele de test

Overfitting

Diferența dintre eroarea obținută pe datele de antrenament și cea obținută pe datele de test



Calitatea învățării

Măsuri de performanță

- **Măsuri statistice**
  - Acuratețea
  - Precizia
  - Rapelul
  - Scorul F1
- Eficiență
  - În construirea modelului
  - În testarea modelului
- Robustețea
  - Tratarea zgomotelor și a valorilor lipsă
- Scalabilitatea
  - Eficienţa gestionării seturilor mari de date
- Interpretabilitatea
  - Modelului de clasificare

Calitatea învățării - Măsuri de performanță -Măsuri statistice

- Matrice de confuzie: rezultate reale vs. rezultate calculat
- Acurateţea
  - Nr de exemple corect clasificate / nr total de exemple
  - Opusul erorii
  - Calculată pe
    - Setul de validare
    - Setul de test

		Rezultate reale	
		Clasa pozitivă	Clasa(ele) negativă(e)
Rezultate calculate	Clasa pozitivă	*True positiv (TP)*	*False positiv (FP)*
	Clasa(ele) negativă(e)	*False negative (FN)*	*True negative (TN)*

În cazul în care este importantă doar o singură clasă (clasă pozitivă), restul claselor sunt negative

- Precizia (P)
  - nr. de exemple pozitive corect clasificate / nr. total de exemple clasificate ca pozitive
  - probabilitatea ca un exemplu clasificat pozitiv să fie relevant
  - TP / (TP + FP)
- Rapelul (R)
  - nr. de exemple pozitive corect clasificate / nr. total de exemple pozitive
  - Probabilitatea ca un exemplu pozitiv să fie identificat corect de către clasificator
  - TP/ (TP +FN)
- Scorul F1
  - Combină precizia şi rapelul, facilitând compararea a 2 algoritmi
  - Media armonică a preciziei şi rapelului
  - 2PR/(P+R)

Metode de validare

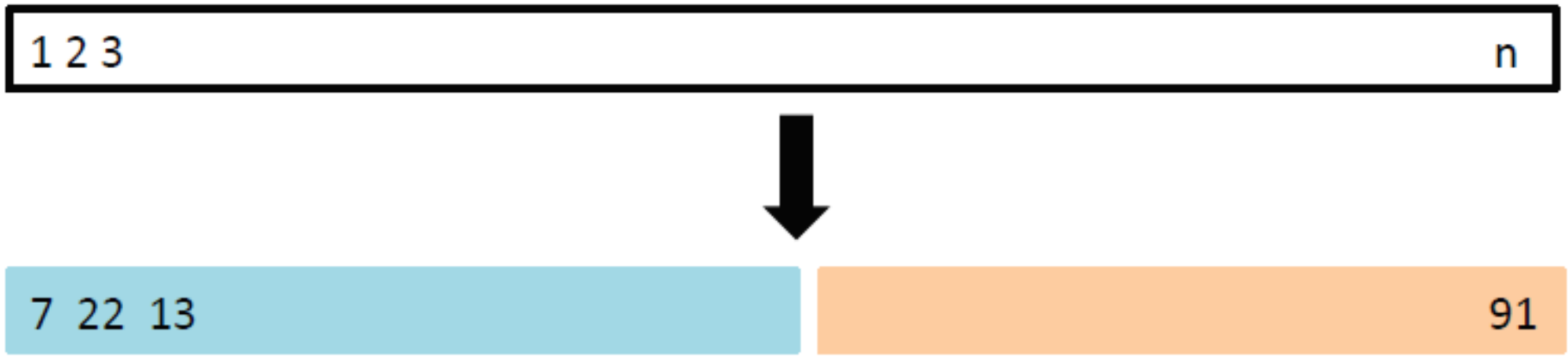
- Validarea simplă (seturi disjuncte de antrenament și testare)
- Validarea încrucișată (k-fold cross-validation)
- Bootstrap

Metoda de validare simplă
Seturi disjuncte de antrenament și testare

- Setul de date se împarte arbitrar în:
  - Setul de date de antrenament
  - Setul de date de validare
- Modelul este învățat pe datele de antrenament, și este folosit pentru a face predicții pe datele din setul de validare
- Eroarea pe setul de validare furnizează o estimare a erorii de testare

Procesul de validare

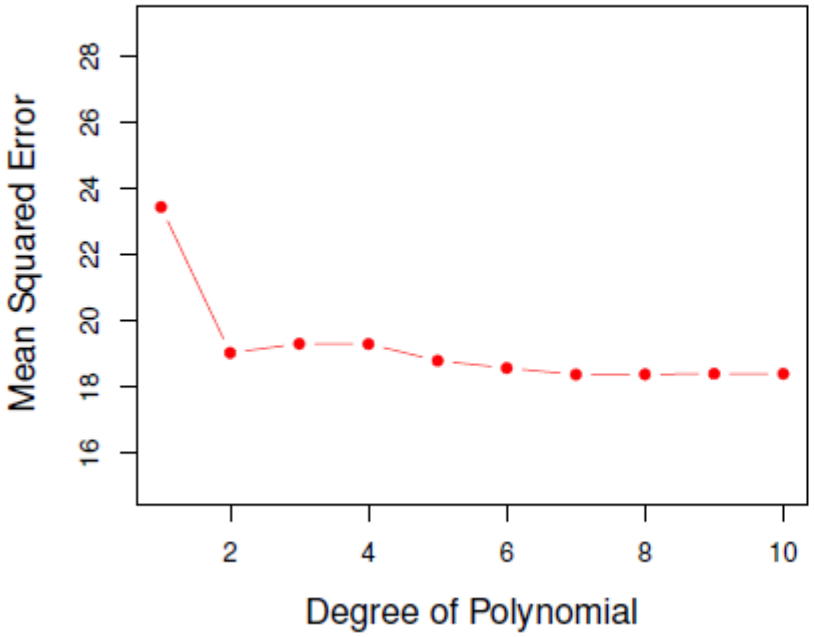
- Datele se împart aleator în două jumătăți: pe o jumătate se face antrenamentul și pe cealaltă jumătate se face validarea:

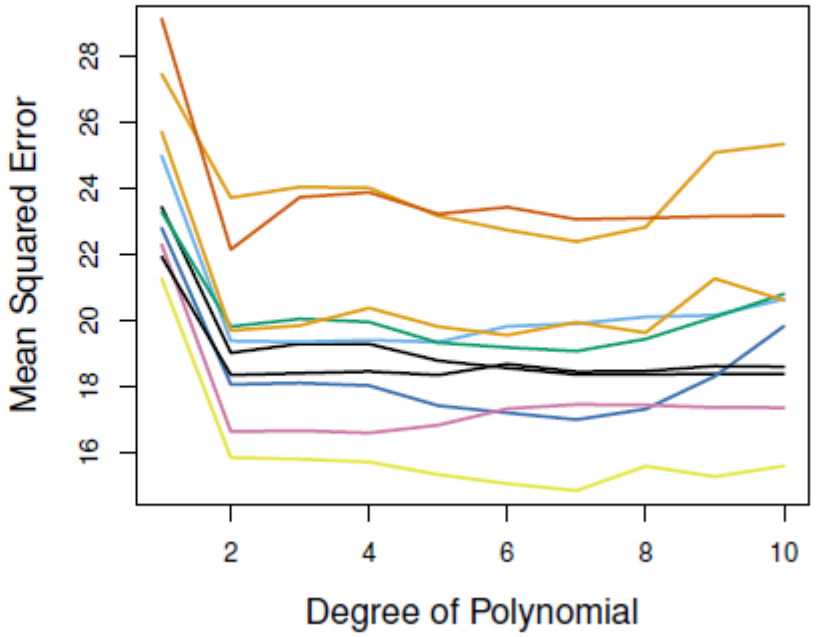


Exemplu:

Setul de date Auto (biblioteca ISLR din limbajul R)

- Obiectiv: compararea regresiei liniare cu un model polinomial
- Împărțim aleator cele 392 de observații în 2 seturi:
  - Datele de antrenament conținând 196 de observații
  - Datele de validare conținând celelalte 196 de observații





Stânga: eroarea obținută pe o singură împărțire a datelor

Dreaptă: mai multe împărțiri aleatorii a datelor

Dezavantajele metodei de validare

- Estimarea erorii de testare prin eroarea obținută pe datele de validare poate fi foarte variabilă, depinzând foarte mult de observațiile ce sunt incluse în setul de antrenament și în setul de validare
- Doar un sub-set al datelor (sub-setul datelor de antrenament) este folosit pentru a învăța modelul (nu și datele din setul de validare)
- Din acest motiv eroarea obținută pe setul de validarea sub-estimează eroarea de test obținută prin antrenarea pe întregul set de date

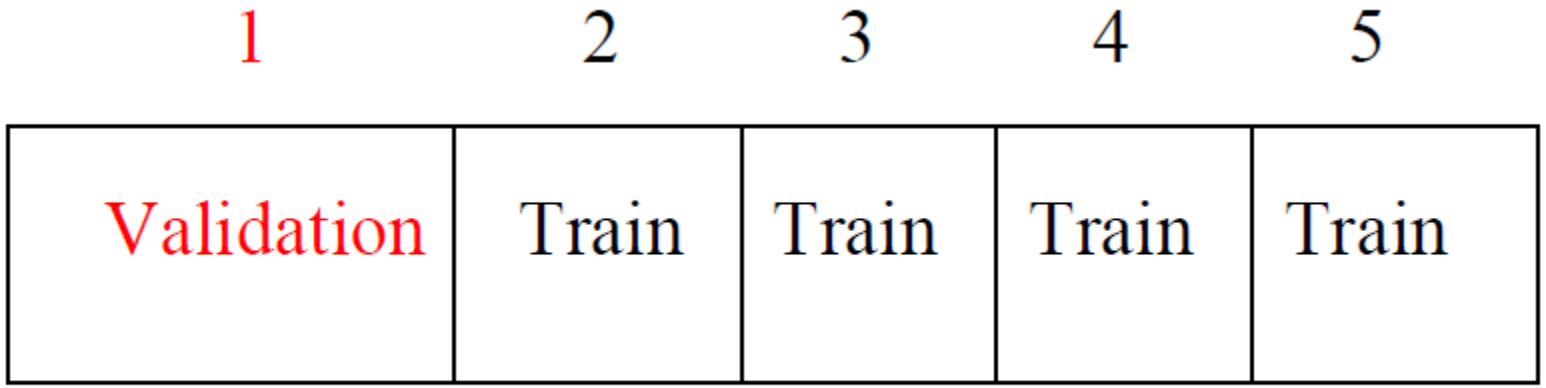
Metode de validare

- Validarea simplă (seturi disjuncte de antrenament și testare)
- Validarea încrucișată (k-fold cross-validation)
- Bootstrap

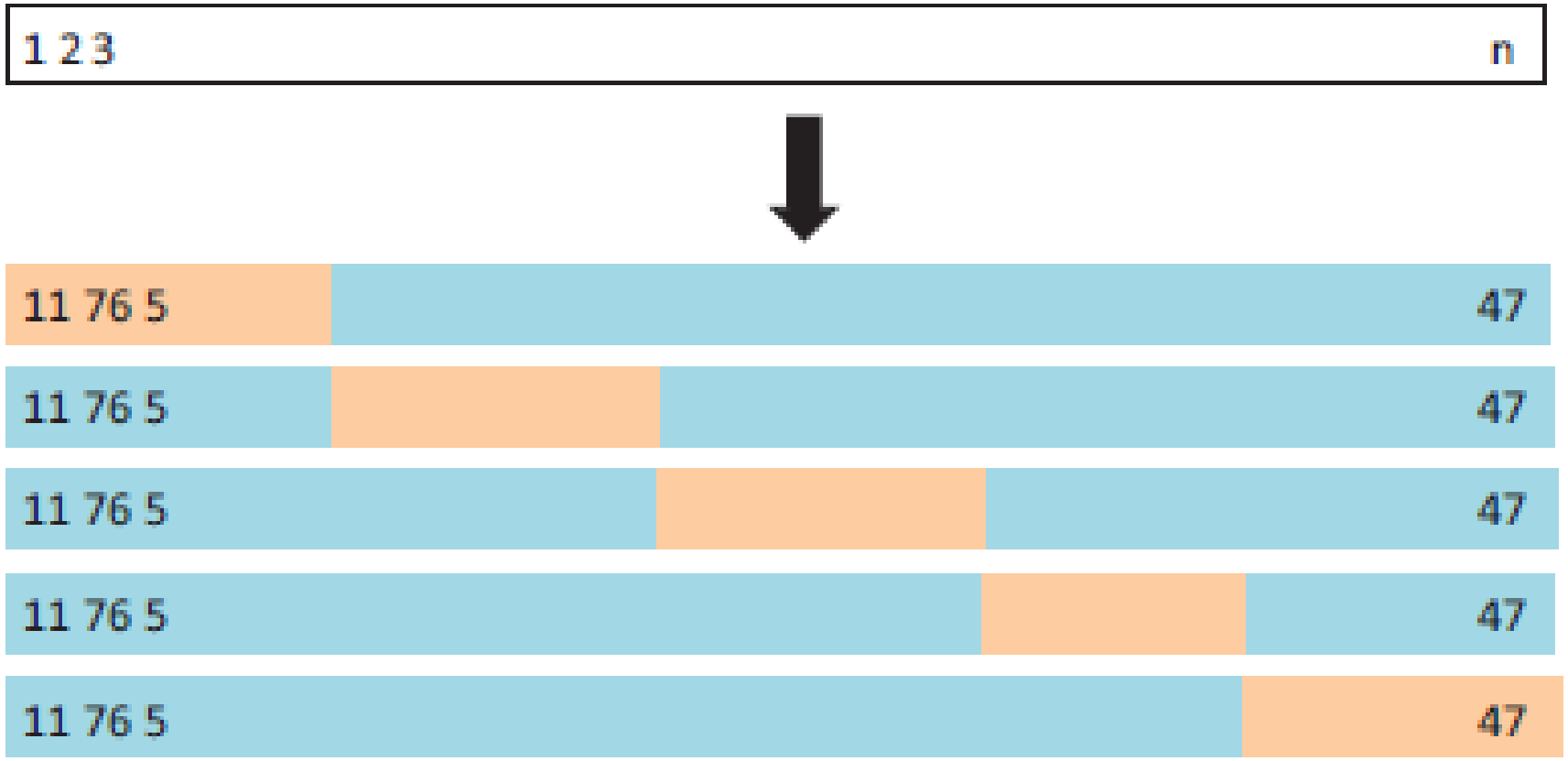
K-fold Cross-validation

Validare încrucișată pe mai multe sub-seturi

- Este o metodă foarte folosită pentru estimarea erorii la testare
- Se poate folosi pentru a alege cel mai bun model, și pt a estima eroarea la testare a celui mai bun model
- Ideea:
  - se împart aleator datele în mod egal în k sub-seturi. Sub-setul k se lasă deoparte, celelalte k-1 sub-seturi se folosesc pt a construi modelul cu care se fac predicții pe datele din sub-setul k
  - acest proces se repetă pt fiecare sub-set k =1,2,...,K și la sfârșit se face media erorii obținute



Exemplu: 5 folds CV



Validarea încrucișată pt clasificare

- We divide the data into K roughly equal-sized parts $C_1, C_2, \ldots C_K$. C_k denotes the indices of the observations in part k. There are n_k observations in part k: if n is a multiple of K, then $n_k = n/K$.
- Compute

$$\mathrm{CV}_K = \sum_{k=1}^{K} \frac{n_k}{n} \mathrm{Err}_k$$

  where $\mathrm{Err}_k = \sum_{i \in C_k} I(y_i \neq \hat{y}_i)/n_k$.
- The estimated standard deviation of CV_K is

$$\widehat{\mathrm{SE}}(\mathrm{CV}_K) = \sqrt{\sum_{k=1}^{K} (\mathrm{Err}_k - \overline{\mathrm{Err}_k})^2/(K-1)}$$

- This is a useful estimate, but strictly speaking, not quite valid.

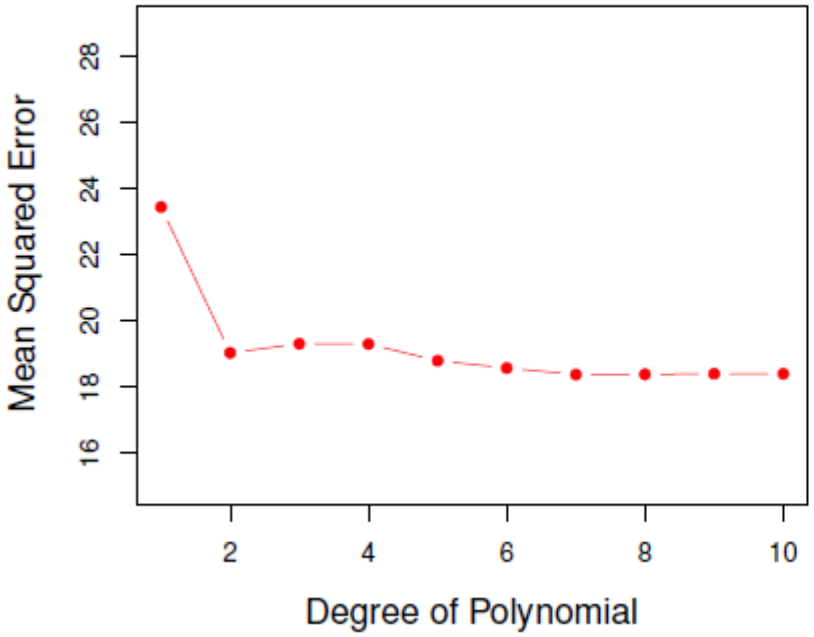
Observații legate de validarea încrucișată

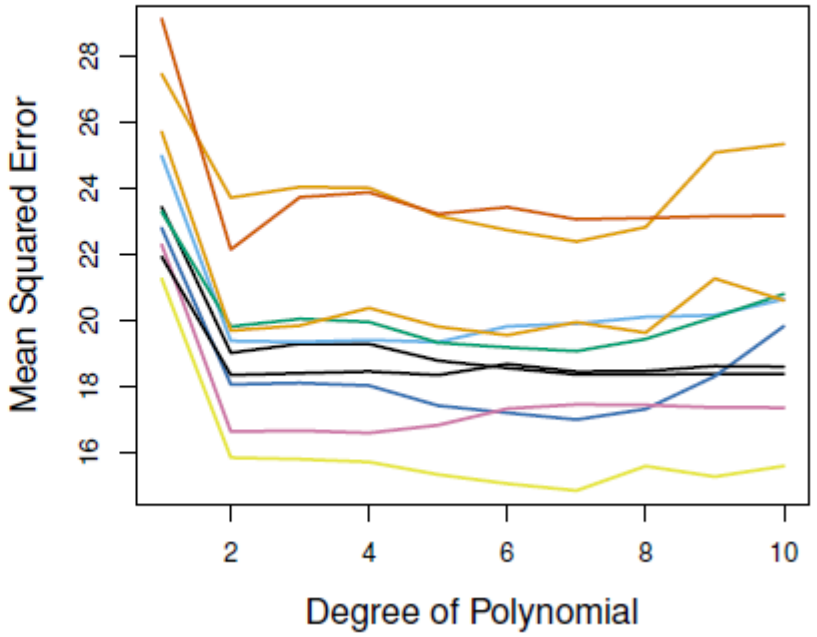
- Deoarece datele de antrenament reprezintă (K-1)/K din setul total de date, eroarea obținută va fi mai mare decât eroarea de testare obținută folosind întreg setul de date pt antrenament
- Acest dezavantaj este minimizat atunci când folosim **leave-one-out cross-validation (LOOCV)** în care K=N, și fiecare sub-set conține o singură observație
- De regulă se folosește K=5, K=10

Revenim la exemplul de mai devreme de la validarea simplă

Setul de date Auto (biblioteca ISLR din R)

- Obiectiv: compararea regresiei liniare cu un model polinomial

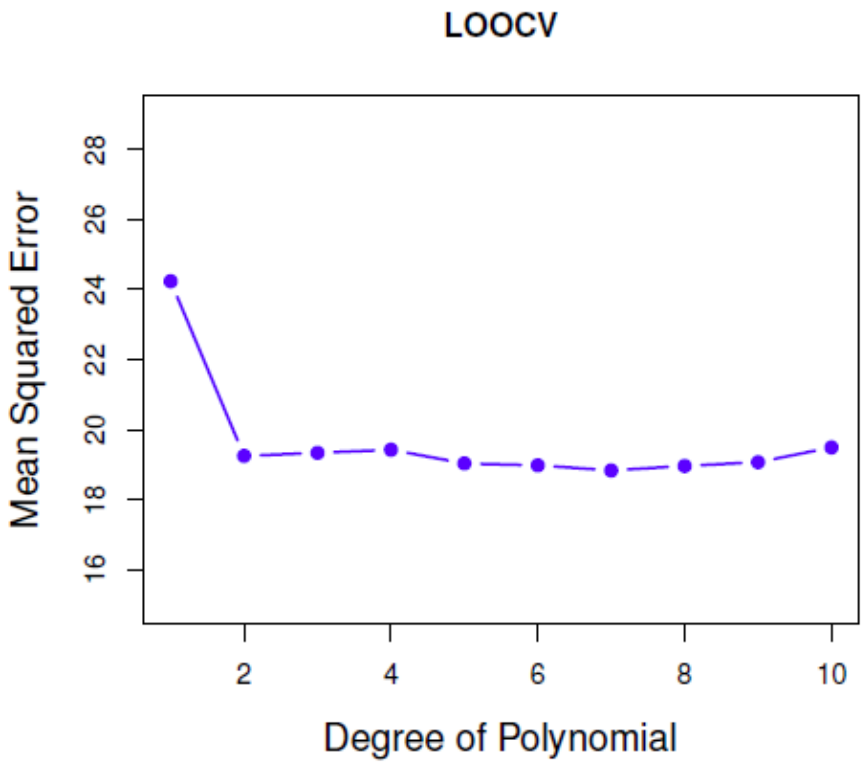


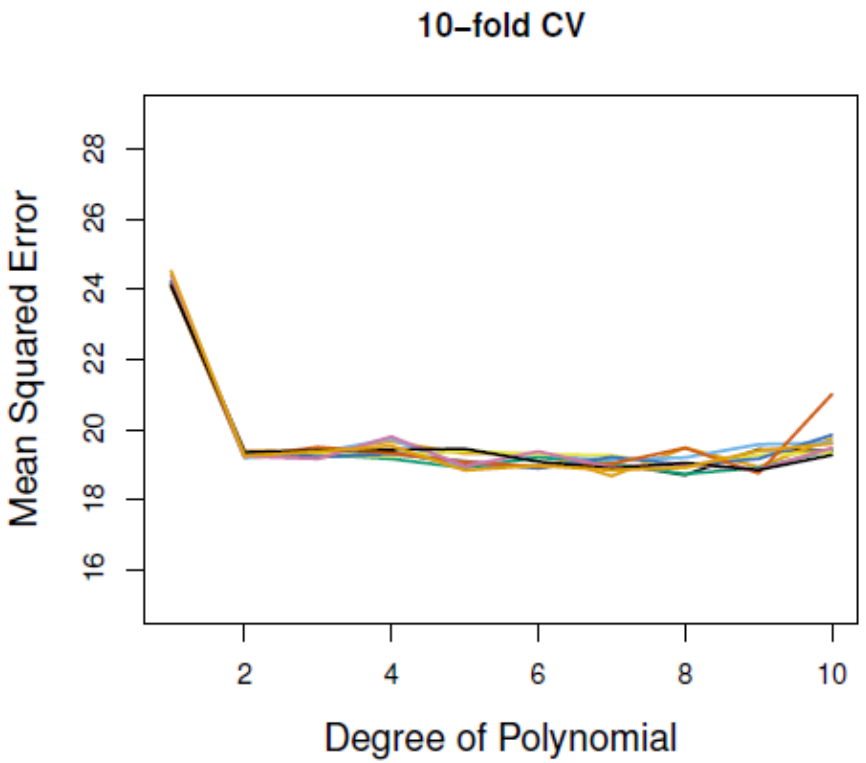


Stânga: eroarea obținută pe o singură împărțire a datelor

Dreaptă: mai multe împărțiri aleatorii a datelor

Exemplu: rezultatele obținute folosind validarea încrucișată
Setul de date Auto (biblioteca ISLR din R)





Metode de validare

- Validarea simplă (seturi disjuncte de antrenament și testare)
- Validarea încrucișată (k-fold cross-validation)
- Bootstrap

Bootstrap

- Este un instrument statistic folosit pt a cuantifica incertitudinea asociată cu un anumit estimator sau cu o metodă statistică
- De ex, poate furniza o valoarea estimativă a variației standard a unui coeficient, sau un interval de încredere pt acel coeficient
- Termenul „bootstrap„ (șireturi) folosit in ML își are originea într-o povestire din cartea „Suprinzătoarele aventuri ale baronului Munchausen„ de Rudolph Erich Raspe:
  - *The Baron had fallen to the bottom of a deep lake. Just when it looked like all was lost, he thought to pick himself up by his own bootstraps.*
- Nu are același sens cu a boot-a un calculator

Motivație pentru metoda bootstrap

- Problemă: vrem să investim o anumită sumă de bani la bursă în 2 active care generează cîștigurile X și Y
- Vom investi o fracție (alfa) din suma totală în X și (1-alfa) din suma totală în Y
- Vrem să alegem alfa a.i. să minimizăm riscul investiției noastre
- Vrem să minimizăm $\mathrm{Var}(\alpha X + (1-\alpha)Y)$
- Valoarea lui alfa care minimizează riscul este dată de formula

$$\alpha = \frac{\sigma_Y^2 - \sigma_{XY}}{\sigma_X^2 + \sigma_Y^2 - 2\sigma_{XY}},$$

$$\text{where } \sigma_X^2 = \mathrm{Var}(X), \sigma_Y^2 = \mathrm{Var}(Y), \text{ and } \sigma_{XY} = \mathrm{Cov}(X, Y).$$

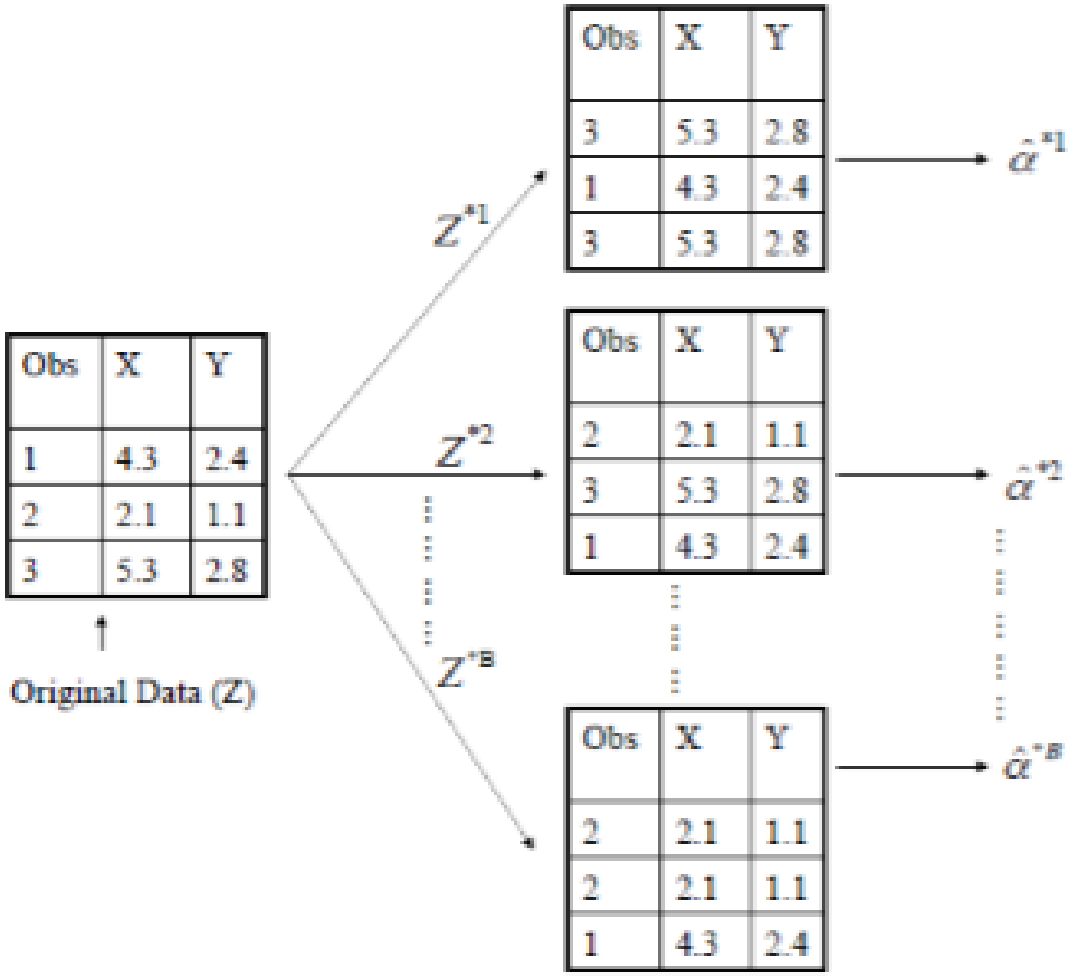
Procedura bootstrap

- Metoda *bootstrap* folosește calculatorul pentru a genera date, a.î. să putem estima variabilitatea estimatorului nostru fără a avea alte date noi
- Pentru că nu putem extrage seturi de date independente din populație, metoda bootstrap extrage seturi de date independente prin eșantionare cu înlocuire din setul de date original
- Fiecare dintre aceste seturi de date aare același număr de observații ca și setul de date original
- Din cauza eșantionării cu înlocuire, unele observații pot să apară de mai multe ori, iar altele deloc

Exemplu de aplicare a metodei bootstrap

O ilustrare grafică a metodei bootstrap pe un set de date format din 3 observații.

Fiecare sub-set de date obținut prin bootstrap conține n=3 observații care sunt eșantionate cu înlocuire din setul de date original.



Bootstrap pt estimarea erorii de test?

- În validarea încrucișată fiecare dintre cele K sub-seturi de date folosite pt validare este diferită de celelalte K-1 sub-seturi folosite pentru antrenament (nu există nici o obs comună). Acest lucru este foare important pt estimarea corectă a erorii!
- Dacă folosim Bootstrap pt estimarea erorii, putem să folosim un set de date obținut prin bootstrap ca și set de antrenament, și setul original de date ca și validare
- Pb: multe observații sunt comune, apar atât în setul de antrenament cât și în setul de validare (aproximativ 2 treimi din setul de date original apare într-un set de date obținut prin bootstrap).Din acest motiv, eroarea obținută va fi o sub-estimare a erorii reale

Materiale de citit

- Capitolul 5 din cartea: **An Introduction to Statistical Learning with Applications in R**. Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Springer-Verlag, 2013. (disponibilă gratuit online aici: http://www-bcf.usc.edu/~gareth/ISL/)