

Evaluarea modelelor

Recapitulare: învățarea supervizată

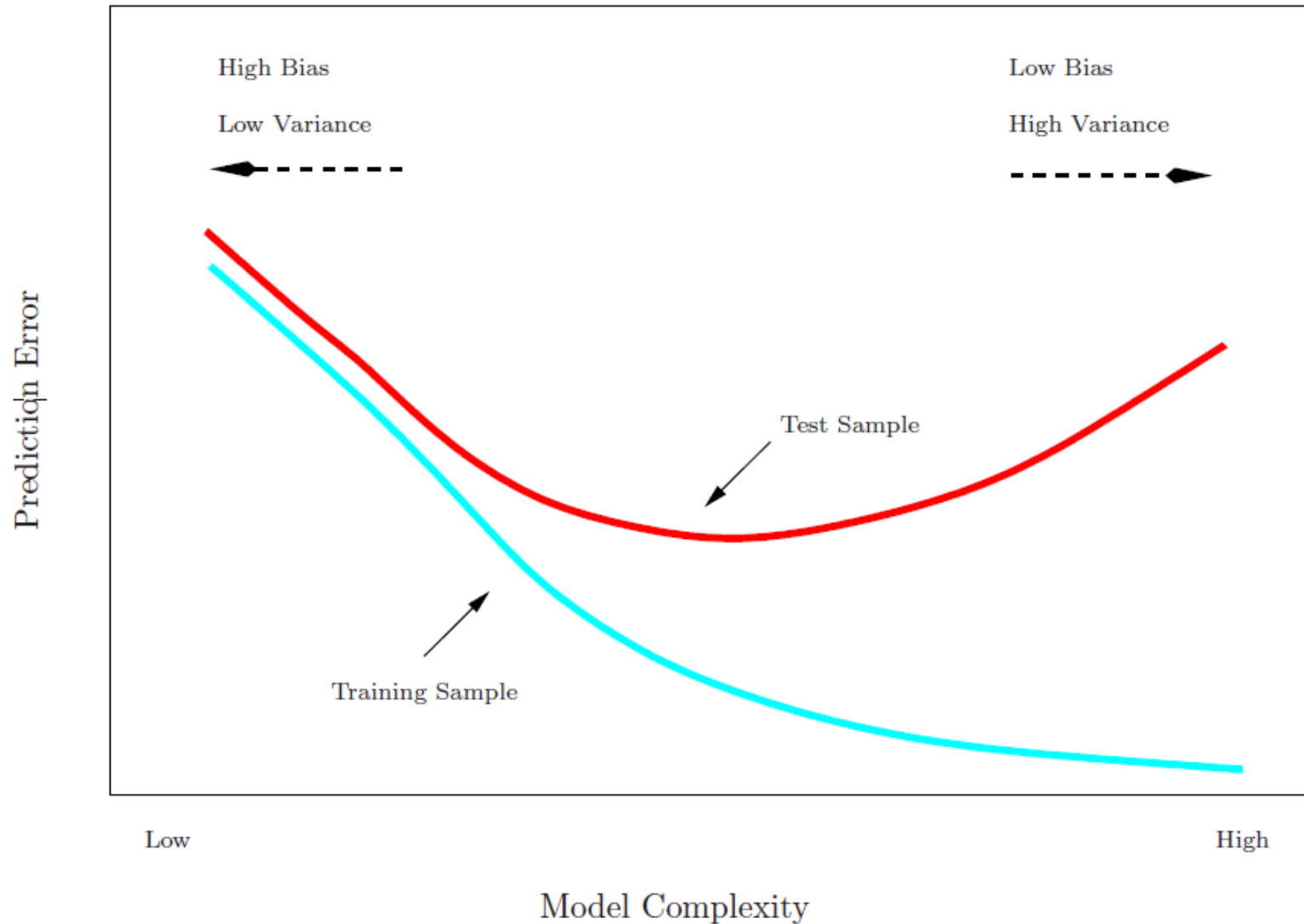
- Scop
 - Furnizarea unei ieșiri corecte pentru o nouă intrare
- Definiere
 - Se dă un set de date (exemple, instanțe, cazuri)
 - Date de antrenament – sub forma unor perechi (attribute_data_i, iesire_i) unde
 - $i=1, n$ (n =nr. datelor de antrenament)
 - attribute_data_i =(atr_i1, atr_i2, ..., atr_im), 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 regresie
 - Date de test – sub forma (attribute_data_i), $i=1, t$ (t = nr. datelor de test)
 - Să se determine
 - O funcție (necunoscută) care realizează corespondența atribut-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

- Definiere
 - 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 $k = \text{nr de date}$ (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**
 - Eficiența gestionării seturilor mari de date
- **Interpretabilitatea**
 - Modelului de clasificare

Calitatea învățării - Măsuri de performanță - Măsuri statistice

- Matrice de confuzie: rezultate reale vs. rezultate calculat
- Acuratețea
 - 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	<i>True positiv (TP)</i>	<i>False positiv (FP)</i>
	Clasa(ele) negativă(e)	<i>False negative (FN)</i>	<i>True negative (TN)</i>

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 și rapelul, facilitând compararea a 2 algoritmi
 - Media armonică a preciziei și rapelului
 - $2PR / (P + R)$

Metode de validare

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

Metoda de validare simplă

Seturi disjuncte de antrenare și testare

- Setul de date se împarte arbitrar în:
 - Setul de date de antrenare
 - Setul de date de validare
- Modelul este învățat pe datele de antrenare, 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

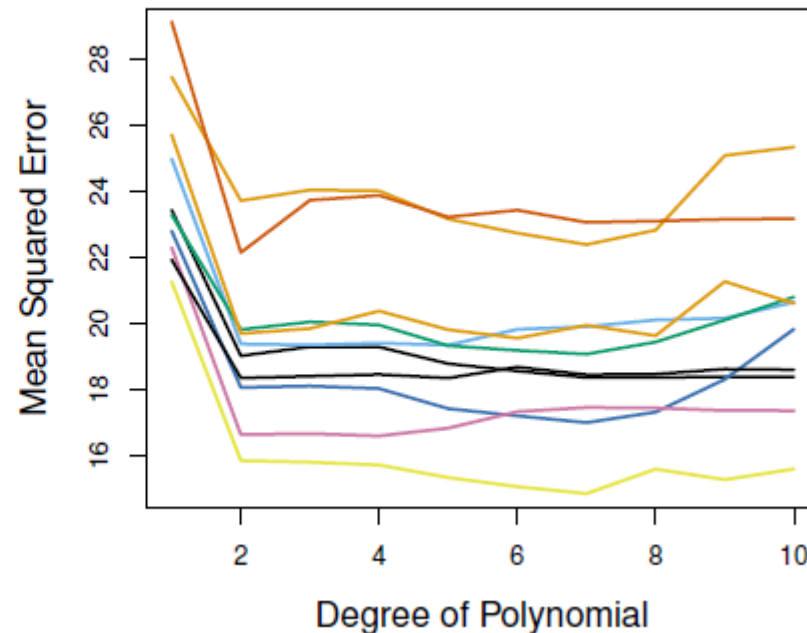
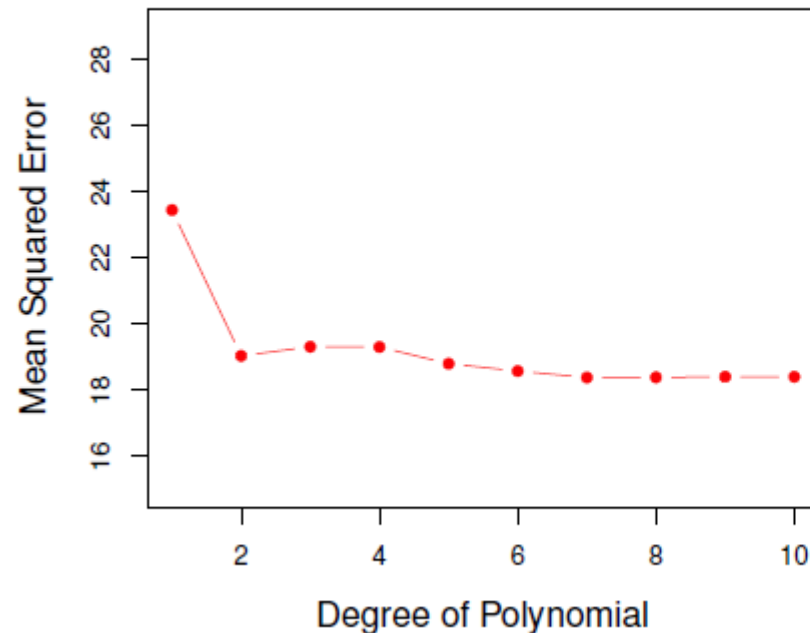
- Datele se împart aleator în două jumătăți: pe o jumătate se face antrenarea și pe cealaltă jumătate se face validarea:



Exemplu:

Setul de date Auto

- 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

Dreapta:
mai multe
împărțiri
aleatorii a
datelor ¹⁰

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 validare sub-estimează eroarea de test obținută prin antrenarea pe întregul set de date

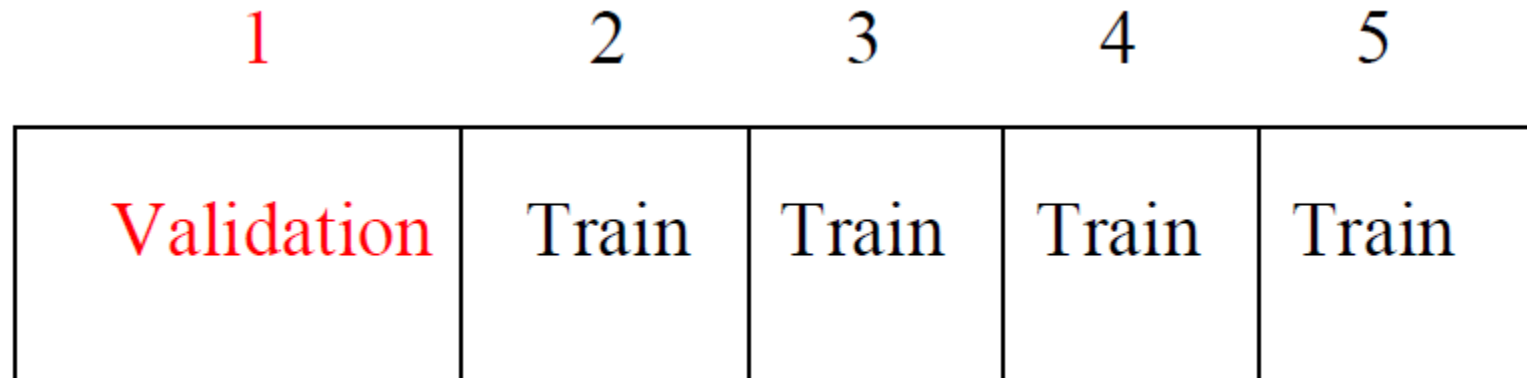
Metode de validare

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

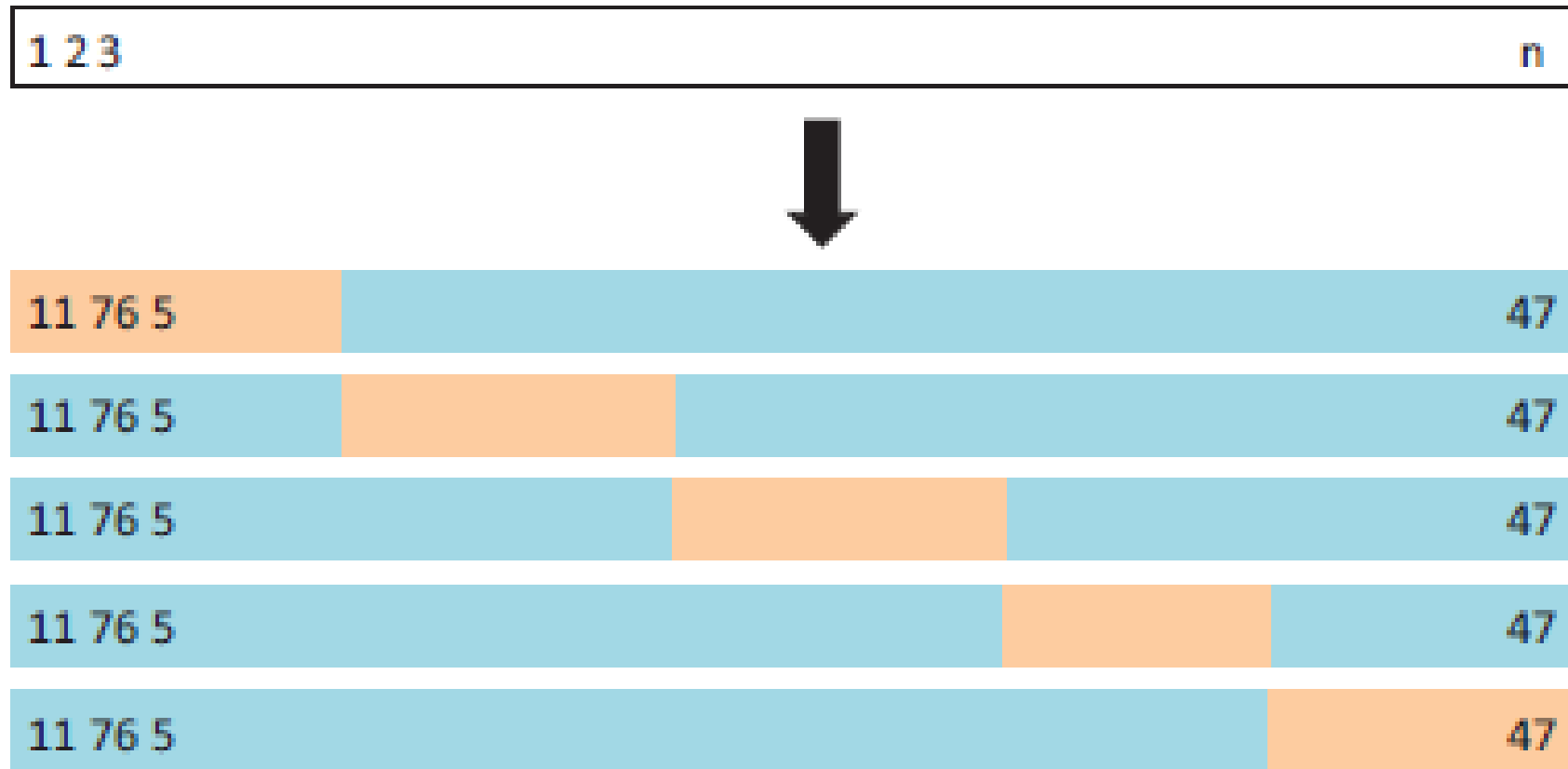
K-fold Cross-validation

Validare încrucișată pe mai multe sub-seturi

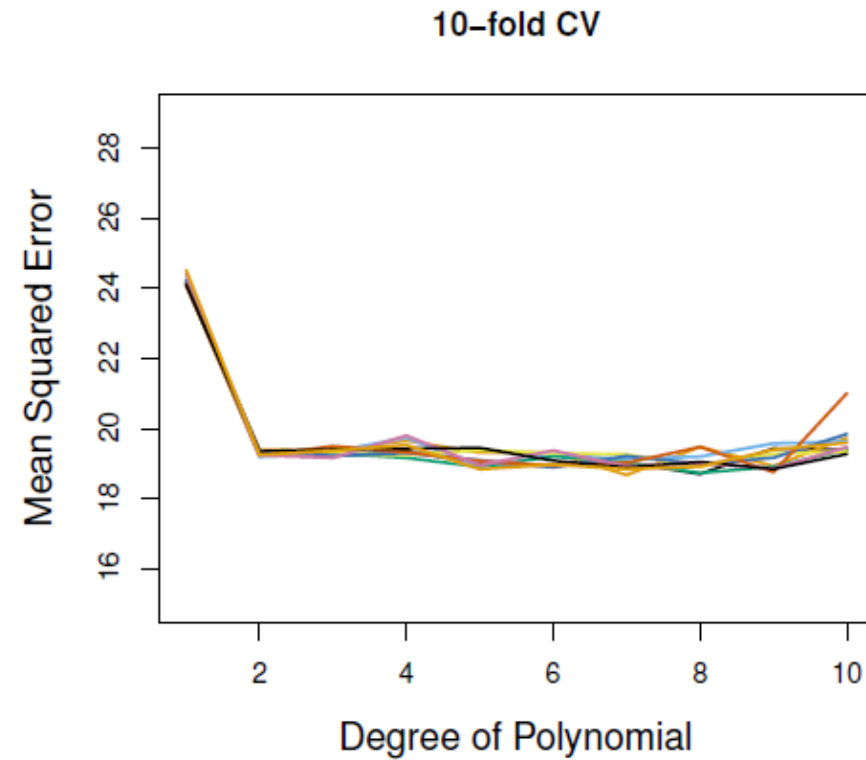
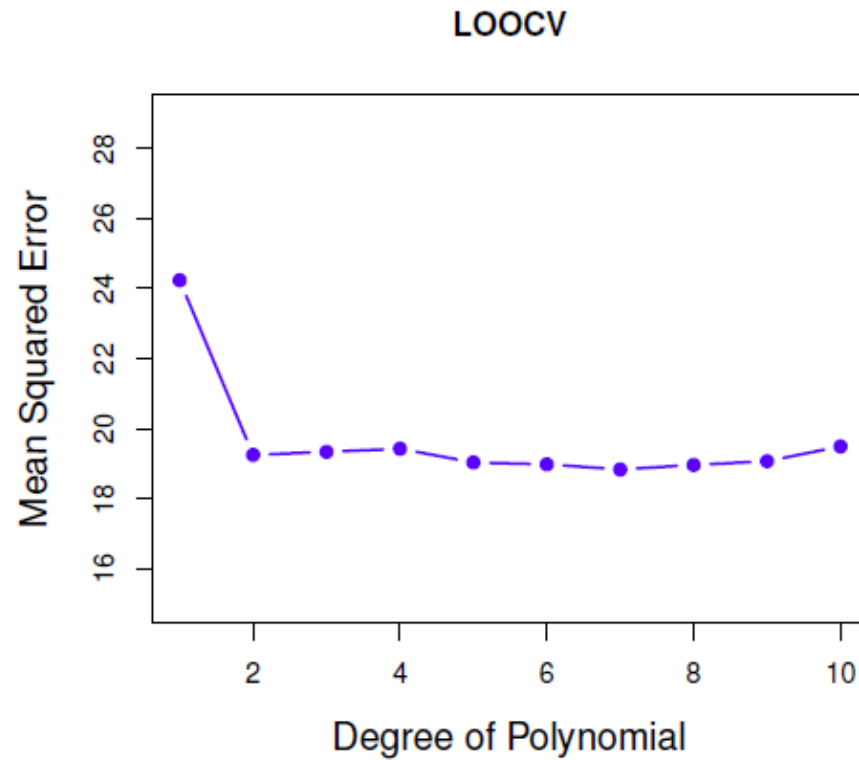
- Este o metodă foarte folosită pentru estimarea erorii la testare
- Se poate folosi pentru a alege cel mai bun model, și pt a estima eroarea la testare a celui mai bun model
- Ideea:
 - se împart aleator datele în mod egal în k sub-seturi. Sub-setul k se lasă deoparte, celelalte $k-1$ sub-seturi se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



Exemplu: 5 folds CV



Exemplu: rezultatele obținute folosind validarea încrucișată Setul de date Auto



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 $K=5$, $K=10$

Metode de validare

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

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/>)