UNIVERSITÀ DEGLI STUDI DI PADOVA

METODI STATISTICI PER LA BIOINGEGNERIA (B)

PARTE 11: REGOLARIZZAZIONE

A.A. 2024-2025

Prof. Martina Vettoretti

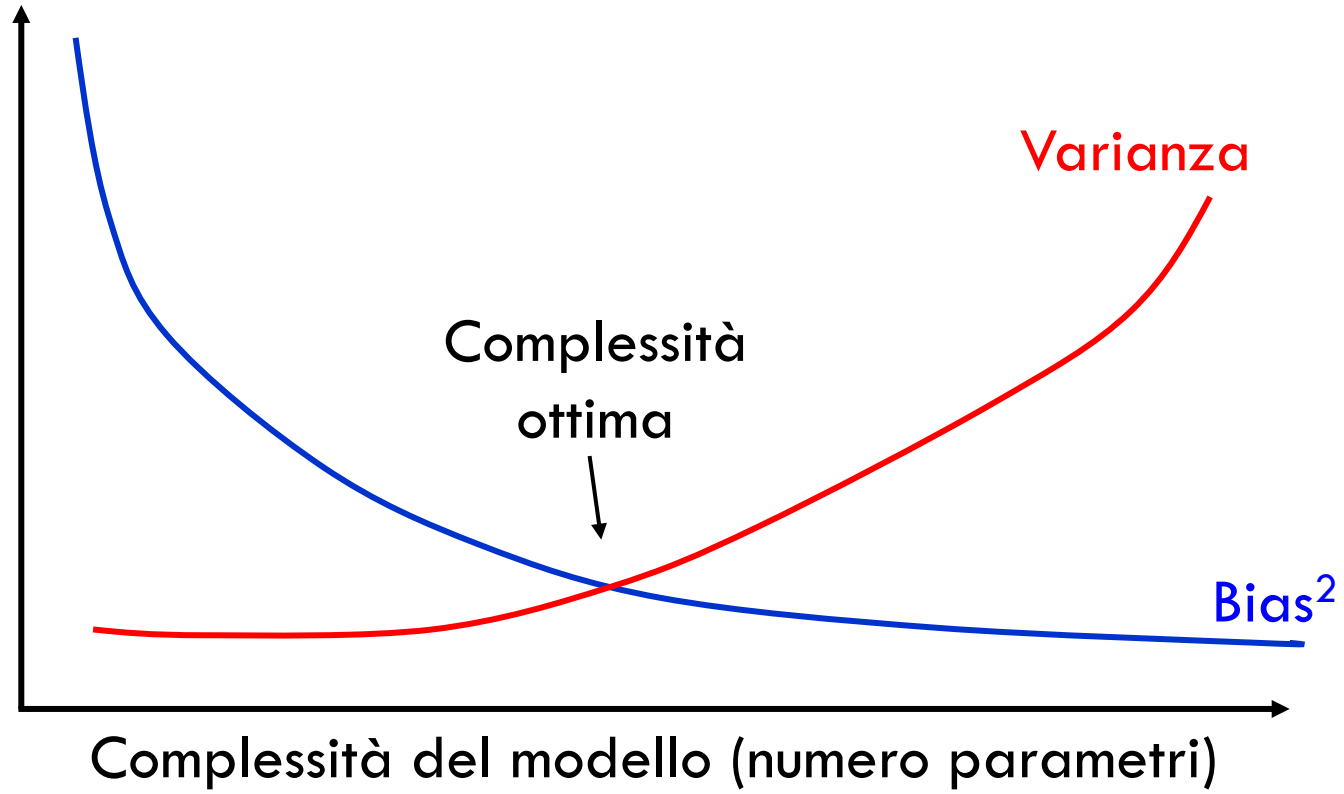
SCOMPOSIZIONE DELL'ERRORE

- Relazione vera tra X_1, X_2, …, X_m e Y: $Y = f(X_1, \dots, X_m) + \delta, \;\; E[\delta] = 0, \; Var(\delta) = \sigma_\delta^2$
- Approssimazione tramite regressione lineare multipla: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_m X_m + \varepsilon$
- Stime dei parametri: $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m$
- Stimatore di Y: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_m X_m$
- Errore quadratico del modello di regressione lineare multipla: $(Y - \hat{Y})^2$
- Scomposizione dell'errore quadratico medio:

$$E[(Y - \hat{Y})^2] = \underbrace{Var(\delta)}_{\text{Errore irriducibile, non dipende dal modello}} + \underbrace{(E[\hat{Y}] - Y)^2}_{\text{Bias}^2} + \underbrace{Var(\hat{Y})}_{\text{Varianza}}$$

IL COMPROMESSO TRA BIAS E VARIANZA

- Idealmente vorremmo minimizzare sia il bias che la varianza dello stimatore $\hat{Y}$. Essi però variano in direzioni opposte al variare della complessità del modello → occorre trovare un compromesso (trade-off)



- Per minimizzare l'errore di predizione del modello conviene limitare la complessità del modello, aumentando un po' il bias per mantenere bassa la varianza.

METODI DI SHRINKAGE

- I metodi di *shrinkage* (restringimento) introducono **bias nelle stime** dei coefficienti $\hat{\boldsymbol{\beta}}$ (e quindi di $\hat{\boldsymbol{Y}}$) al fine di mantenere bassa la varianza di $\hat{\boldsymbol{Y}}$ e quindi rendere meno incerta la predizione.

- Il bias viene introdotto introducendo una **regolarizzazione** della stima dei coefficienti $\boldsymbol{\beta}$ che penalizza valori grandi in valore assoluto delle stime dei coefficienti $\hat{\boldsymbol{\beta}}$ → Le stime sono «ristrette» in valore assoluto verso lo zero.

- I metodi di shrinkage sono particolarmente utili quando abbiamo tante variabili correlate tra loro → La stima non regolarizzata potrebbe portare alla stima di coefficienti grandi in valore assoluto e di segno opposto per variabili tra loro correlate.

STIMA STANDARD VS STIMA REGOLARIZZATA

- Problema di ottimizzazione standard (senza regolarizzazione):

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}}(F(\boldsymbol{\beta}))$$

Quando usiamo i minimi quadrati lineari:

$$F(\boldsymbol{\beta}) := (\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta}) = SSE$$

- Problema di ottimizzazione con regolarizzazione:

$$\widehat{\boldsymbol{\beta}}_{reg,\lambda} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}}(F(\boldsymbol{\beta}) + P(\lambda; \beta_1, \beta_2, \dots \beta_m))$$

Termine di penalità: penalizza valori grandi (in modulo) dei coefficienti

Parametro di regolarizzazione: regola il grado di regolarizzazione.

Nota: nella formulazione classica non si regolarizza l'intercetta, β_0.

REGOLARIZZAZIONE L2 O REGRESSIONE RIDGE (1/2)

- **Regolarizzazione L2:** il termine di penalità è il quadrato della norma 2 del vettore dei coefficienti di regressione (intercetta esclusa) → il modello di regressione prende il nome di **regressione Ridge.**

$$\widehat{\boldsymbol{\beta}}_{Ridge,\lambda} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}}(F(\boldsymbol{\beta}) + \lambda \cdot \sum_{j=1}^{m} \beta_j^2)$$

$$F(\boldsymbol{\beta}) := (\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta}) = SSE$$

- Si penalizzano valori elevati (in modulo) dei coefficienti di regressione.
- Effetto risultante: le stime dei parametri vengono «schiacciate» verso lo 0. Tuttavia nessun coefficiente viene posto a 0. La complessità del modello risulta dunque invariata.

REGOLARIZZAZIONE L2 O REGRESSIONE RIDGE (2/2)

- λ è uno scalare positivo. Più λ è grande, maggiore è la penalità imposta su valori grandi dei coefficienti.
  - Quando $\lambda \to \infty, \hat{\beta}_j \to 0$
- Il valore ottimale di λ va trovato.
- Fissato λ si può dimostrare che la soluzione al problema di regressione Ridge è:

$$\widehat{\boldsymbol{\beta}}_{Ridge,\lambda} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \cdot \boldsymbol{Q})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

dove $\boldsymbol{Q}$ è la matrice identità m+1 x m+1 con uno zero sulla diagonale in corrispondenza dell'intercetta (parametro non penalizzato).

Esempio:

$\beta = [\beta_0, \beta_1, \beta_2]$ con β_0 intercetta

$$Q = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- Valori come λ, che non sono parametri del modello lineare (cioè non sono coefficienti β), ma hanno un impatto sulla forma finale del modello si chiamano **iperparametri.**

ESEMPIO: EFFETTO DELLA REGOLARIZZAZIONE L2

- Modello per la predizione del diametro della componente acetabolare della protesi all'anca. Variabili indipendenti: altezza, girovita, lunghezza piede, età, sesso, patologia (2 variabili dummy: frattura e necrosi).

Coefficienti	$\boldsymbol{\lambda = 0}$	$\boldsymbol{\lambda = 1}$	$\boldsymbol{\lambda = 10}$	$\boldsymbol{\lambda = 100}$
Intercetta	46.04	46.53	48.86	51.97
Altezza	6.91	6.19	3.58	1.02
Girovita	5.61	5.22	3.26	0.85
Lunghezza piede	4.35	4.27	3.10	1.02
Età	-0.56	-0.61	-0.71	-0.36
Sesso	-0.49	-0.23	0.95	1.44
Frattura	-0.26	-0.27	-0.28	-0.18
Necrosi	0.11	0.10	0.03	-0.01

REGOLARIZZAZIONE L1 O REGRESSIONE LASSO

- **Regolarizzazione L1**: il termine di penalità è la norma 1 del vettore dei coefficienti di regressione (intercetta esclusa) → il modello di regressione prende il nome di **regressione LASSO** (Least Absolute Shrinkage and Selection Operator).

$$\widehat{\boldsymbol{\beta}}_{LASSO,\lambda} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(F(\boldsymbol{\beta}) + \lambda \cdot \sum_{j=1}^{m} |\beta_j|\right)$$

$$F(\boldsymbol{\beta}) := (\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta})^T (\boldsymbol{Y} - \boldsymbol{X} \cdot \boldsymbol{\beta}) = SSE$$

- Effetto risultante: le stime dei parametri vengono «schiacciate» verso lo 0 e, per valori sufficientemente grandi di λ, i coefficienti più piccoli vengono posti a 0 → si riduce la complessità del modello.
- Il valore ottimo per l'iperparametro λ va trovato.

ESEMPIO: EFFETTO DELLA REGOLARIZZAZIONE L1

- Modello per la predizione del diametro della componente acetabolare della protesi all'anca. Variabili indipendenti: altezza, girovita, lunghezza piede, età, sesso, patologia (2 variabili dummy: frattura e necrosi).

Coefficienti	$\lambda = 0$	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 1$
Intercetta	46.04	46.52	49.23	52.41	53.86
Altezza	6.91	6.45	3.98	0.62	0
Girovita	5.61	5.16	2.29	0	0
Lunghezza piede	4.35	3.93	2.23	0.26	0
Età	-0.56	-0.48	0	0	0
Sesso	-0.49	-0.15	0.94	2.26	0
Frattura	-0.26	-0.24	0	0	0
Necrosi	0.11	0.05	0	0	0

REGOLARIZZAZIONE ELASTIC NET

- **Regolarizzazione Elastic Net:** termine di penalità dato da una combinazione lineare dei termini di penalità delle regolarizzazioni L1 e L2.

$$\widehat{\boldsymbol{\beta}}_{ENet,\lambda,\alpha} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left(F(\boldsymbol{\beta}) + \lambda \cdot \sum_{j=1}^{m}\left[(1-\alpha)\cdot\beta_j^2 + \alpha \cdot \left|\beta_j\right|\right]\right)$$

$$F(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\cdot\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\cdot\boldsymbol{\beta}) = SSE$$

- Iperparametri: λ, α
  - λ regola il grado di regolarizzazione
  - $\alpha \in [0,1]$ indica quanto prevale il termine di penalità L1 su quello L2.
    - Se $\alpha = 0$ → Elastic net è equivalente a Ridge
    - Se $\alpha = 1$ → Elastic net è equivalente a LASSO
- I valori ottimi di entrambi gli iperparametri vanno trovati.

QUALE REGOLARIZZAZIONE ?

- Ridge vs LASSO:
  - Ridge non riduce la complessità del modello, mentre LASSO lo fa.
  - Se nel nostro dataset ipotizziamo esserci pochi predittori forti e molte variabili «rumore» (cioè non associate all'outcome) → preferiamo LASSO
  - Se ipotizziamo non esserci variabili indipendenti dall'outcome → preferiamo Ridge
- Elastic net consente di avere sia i vantaggi della Ridge che della LASSO
  - Tuttavia con elastic net abbiamo 2 iperparametri da stimare → la ricerca degli iperparametri ottimi può diventare computazionalmente onerosa

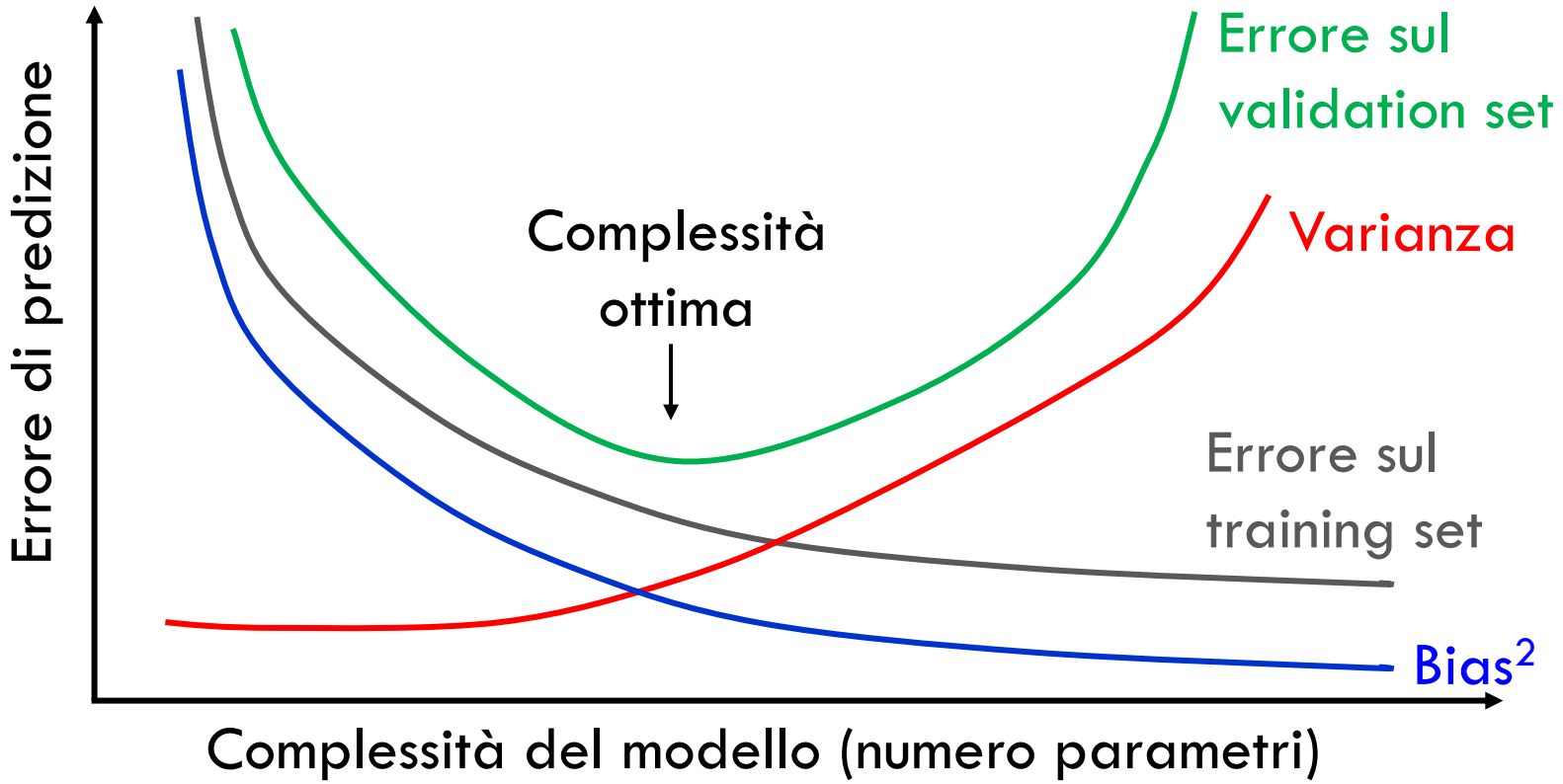
TUNING DEGLI IPERPARAMETRI

- Non è possibile stabilire a priori il valore degli iperparametri.
- Non esistono neanche stimatori che permettono di calcolare il valore degli iperparametri.
- Il loro valore viene stimato empiricamente. Si testano diversi valori per λ (ed eventualmente α) e si seleziona il parametro (o la combinazione di parametri) che minimizza l'errore di predizione del modello, tipicamente RMSE o MSE.

$$MSE = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

ERRORE DI PREDIZIONE SU DATI DI VALIDAZIONE

- **Training set:** set di dati utilizzati per stimare i parametri del modello.
- **Validation set:** nuovo set di dati, non utilizzato per stimare i parametri del modello.



- Dobbiamo stimare i valori degli iperparametri che minimizzano l'errore di predizione su un set di dati di validazione, non utilizzato per stimare i parametri $\boldsymbol{\beta}$ del modello.

OTTIMIZZAZIONE DEGLI IPERPARAMETRI MEDIANTE VALIDATION SET (1/2)

- Dividiamo i dati a disposizione in due set:
  - Training set: set di dati impiegato per stimare i parametri del modello di regressione.
  - Validation set: set di dati impiegato per valutare la miglior combinazione di iperparametri.
- Scegliamo una griglia di possibili valori per ciascun iperparametro:

$\lambda_1, \lambda_2, \ldots, \lambda_L$

$\alpha_1, \alpha_2, \ldots, \alpha_A$

→ A x L possibili coppie di valori per gli iperparametri

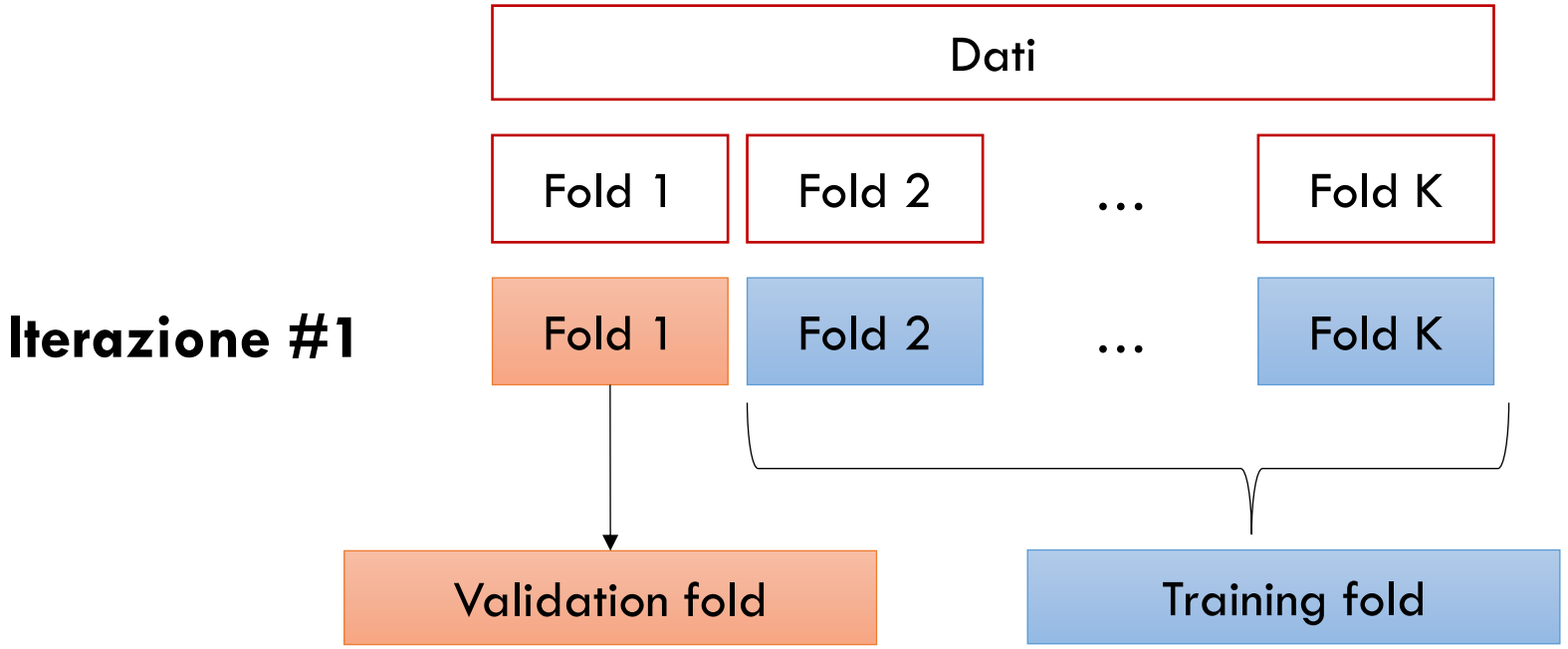
OTTIMIZZAZIONE DEGLI IPERPARAMETRI MEDIANTE VALIDATION SET (2/2)

- Per ogni coppia di valori (λ_i, α_j):
  - alleniamo il modello sul training set (cioè ne stimiamo i parametri $\boldsymbol{\beta}$ usando la stima regolarizzata)
  - Usiamo il modello con $\boldsymbol{\beta}$ stimato sul training set per predire l'outcome sui dati del validation set → calcoliamo l'MSE delle predizioni sul validation set.
  - Selezioniamo la coppia di iperparametri $(\lambda_{opt}, \alpha_{opt})$ che minimizza l'MSE sul validation set.
  - Usiamo $(\lambda_{opt}, \alpha_{opt})$ per stimare i coefficienti del modello finale sul training set.

- Limitazione: siamo sicuri che il particolare split dei dati che facciamo per ricavare il training set e il validation set non influenzi il risultato?

K-FOLD CROSS-VALIDATION (1/2)

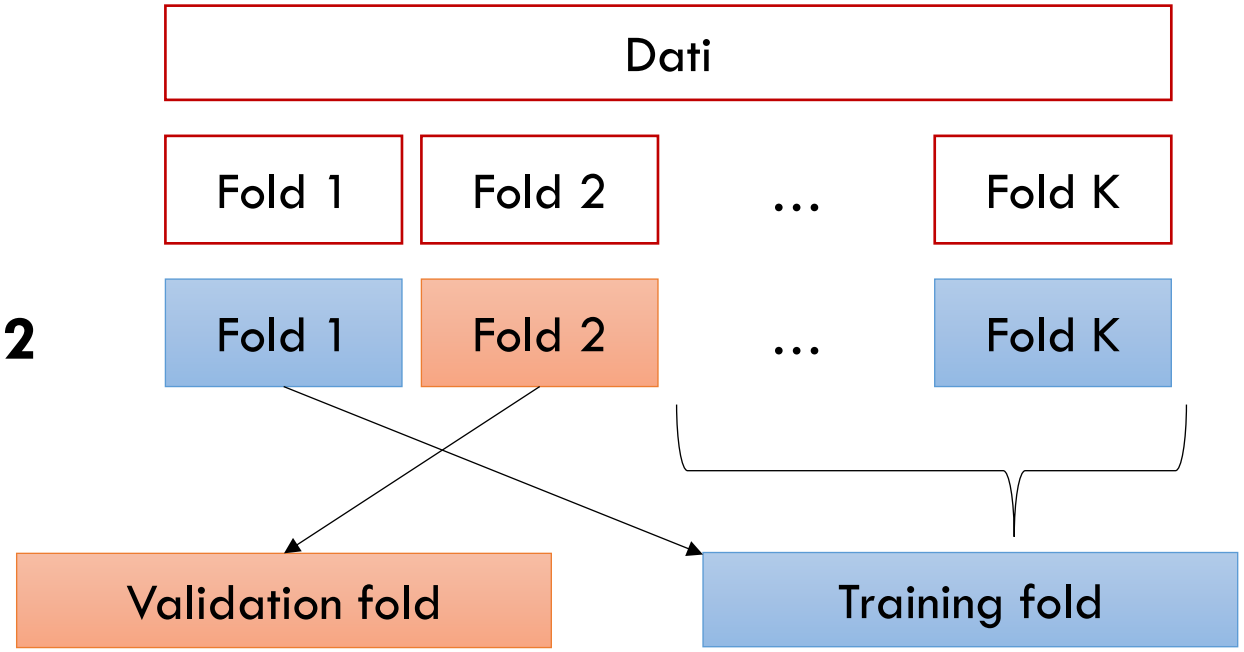
- Dividiamo i dati in K sottoinsiemi, detti fold. Realizziamo K iterazioni.
- Iterazione 1:
  - Fold 2-K → Dati da usare per il training del modello (stima dei parametri $\boldsymbol{\beta}$)
  - Fold 1 → Set di validazione per calcolare le performance del modello (MSE)



K-FOLD CROSS-VALIDATION (2/2)

- Dividiamo i dati in K sottoinsiemi, detti fold. Realizziamo K iterazioni.
- Iterazione 2:
  - Fold 1, 3-K → Dati da usare per il training del modello (stima dei parametri $\boldsymbol{\beta}$)
  - Fold 2 → Set di validazione per calcolare le performance del modello (MSE)

Iterazione #2



- In generale alla i-esima iterazione usiamo la fold i-esima come set di dati per la validazione, l'unione delle altre fold come set di dati per il training del modello.

K-FOLD CROSS-VALIDATION PER L'OTTIMIZZAZIONE DEGLI IPERPARAMETRI

N = A x L diverse combinazioni di iperparametri candidate:

- ad ogni iterazione, alleniamo sulle fold di training gli N modelli corrispondenti alle N combinazioni di iperparametri. Testiamo ciascun modello sulla fold di validazione, calcolando l'MSE.
- Al termine delle iterazioni, calcoliamo la media dell'MSE sulle fold di validazione per ciascuna combinazione di iperparametri.
- Combinazione ottima di iperparametri: quella che minimizza l'MSE medio sulle fold di validazione.

MSE sulla fold i-esima				
	λ_1	λ_2	…	λ_L
α_1	$MSE^i_{1,1}$	$MSE^i_{1,2}$	…	$MSE^i_{1,L}$
α_2	$MSE^i_{2,1}$	$MSE^i_{2,2}$	…	$MSE^i_{2,L}$
…				
α_A	$MSE^i_{A,1}$	$MSE^i_{A,2}$	…	$MSE^i_{A,L}$

$$\overline{MSE}_{a,l} = \frac{1}{K}\sum_{i=1}^{K} MSE^i_{a,l}$$

MSE medio sulle K fold				
	λ_1	λ_2	…	λ_L
α_1	$\overline{MSE}_{1,1}$	$\overline{MSE}_{1,2}$	…	$\overline{MSE}_{1,L}$
α_2	$\overline{MSE}_{2,1}$	$\overline{MSE}_{2,2}$	…	$\overline{MSE}_{2,L}$
…				
α_A	$\overline{MSE}_{A,1}$	$\overline{MSE}_{A,2}$	…	$\overline{MSE}_{A,L}$

$$\min(\overline{MSE}_{a,l}) \Rightarrow \alpha_{opt}, \lambda_{opt}$$

NOTE

- Una volta scelti i valori degli iperparametri ottimi, $\alpha_{opt}, \lambda_{opt}$, si utilizzano questi valori per stimare i coefficienti $\boldsymbol{\beta}$ del modello finale sull'intero set di dati a disposizione (tutte le fold insieme).
- Scelta del numero di fold K:
  - Valori tipici di K sono 5 o 10.
  - Se K è pari numero di osservazioni nel dataset n, la K-fold cross-validation prende il nome di **leave-one-out cross-validation**.
    - Approccio computazionalmente oneroso per dataset di grande dimensione.
- Occorre fare attenzione alla scelta della griglia di valori degli iperparametri da testare.

SCELTA DELLA GRIGLIA DEGLI IPERPARAMETRI

- Sappiamo che α_{opt} è compreso tra 0 e 1

  → possibile griglia per α: {0, 0.1, 0.2, 0.3,…,0.9, 1}

- Sappiamo che $\lambda_{opt} > 0$, ma non ne conosciamo l'ordine di grandezza

  → possibile **griglia iniziale** per λ: {10^{-p}, 10^{-p+1}, …,1, 10, …, 10^{p}}

  - Se al primo giro di ottimizzazione finisco in uno degli **estremi della griglia** dei λ → non ho davvero trovato l'ottimo! Espando la griglia dei λ.

λ_{min}	…	λ_{max}

$\lambda_{opt} = \lambda_{min}$ → **Il valore ottimo è davvero λ_{min}** o una quantità $< \lambda_{min}$?

→ **espando la griglia** aggiungendo valori $< \lambda_{min}$

  - Quando λ_{opt} cade all'interno della griglia dei λ possiamo fermarci qui o decidere di testare una griglia più fitta di valori di λ attorno al valore ottimo trovato.

$\lambda_{min,new}$	…	$\lambda_{min,old}$	…	λ_{max}

$\lambda_{min,new} < \lambda_{opt} < \lambda_{max}$ → **OK**

ESEMPIO

- Modello per la predizione del diametro della componente acetabolare della protesi all'anca. Variabili indipendenti: altezza, girovita, lunghezza piede, età, sesso, patologia (2 variabili dummy: frattura e necrosi).
- Regolarizzazione elastic net con ottimizzazione degli iperparametri mediante **5-fold cross-validation**.
- Griglia di valori per α: {0.1, 0.4, 0.7, 1}
- Griglia di valori per λ: $\{10^{-10}, 10^{-7}, 10^{-4}, 10^{-1}, 10^{2}, 10^{5}, 10^{8}\}$

ESEMPIO: RISULTATI

Minimo dell'MSE medio sulle fold di validazione

```
mean_MSE_CV =

    2.1617    2.1617    2.1615    3.6503    7.4811    7.4811    7.4811
    2.1617    2.1617    2.1615    3.5567    7.4811    7.4811    7.4811
    2.1617    2.1617    2.1616    3.3916    7.4811    7.4811    7.4811
    2.1617    2.1617    2.1617    2.9100    7.4811    7.4811    7.4811
```

α

λ

$$\alpha_{opt} = 0.1$$

$$\lambda_{opt} = 10^{-4}$$

ESEMPIO: RISULTATI

Coefficienti	No regolarizzazione	$\alpha = 0.1, \lambda = 10^{-4}$
Intercetta	46.04	46.52
Altezza	6.91	6.88
Girovita	5.61	5.59
Lunghezza piede	4.35	4.35
Età	-0.56	-0.56
Sesso	-0.49	-0.48
Frattura	-0.27	-0.27
Necrosi	0.11	0.11

In questo caso il modello regolarizzato si discosta poco dal modello non regolarizzato. In generale però la regolarizzazione può impattare in maniera importante i risultati, soprattutto per modelli con tanti predittori.

INFLUENZA DELLA SCALA DELLE VARIABILI

- I parametro di regolarizzazione λ rappresenta il **grado di regolarizzazione**, che idealmente vorremmo essere **uniforme** per tutte le variabili indipendenti.
- Attenzione: **se le variabili hanno scale diverse**, il grado di regolarizzazione non è lo stesso per tutte le variabili!
  - Le variabili che assumono valori più piccoli tenderanno ad avere coefficienti più grandi → maggiore penalizzazione.
- Per rendere uniforme il grado di regolarizzazione per tutte le variabili, prima di applicare la regressione regolarizzata, è buona norma **normalizzare le variabili indipendenti** per riportarle ad avere la stessa scala.
- Approcci di normalizzazione più diffusi:
  - Standardizzazione
  - Min-max scaling

STANDARDIZZAZIONE

$$Z = \frac{X - \bar{X}}{S_X}$$

- X: variabile originale
- Z: variabile standardizzata
- $\bar{X}$: media campionaria di X
- S_X: deviazione standard campionaria di X

➢ Dopo la standardizzazione, tutte le variabili avranno media 0 e varianza 1.

MIN-MAX SCALING

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- X: variabile originale
- X': variabile trasformata
- X_{min}: valore minimo della variabile originale
- X_{max}: valore massimo della variabile originale

- Dopo min-max scaling, tutte le variabili avranno range tra 0 e 1.
- Questo approccio è preferibile quando abbiamo molte variabili qualitative che entreranno nel modello come variabili binarie di valori 0 o 1.