

La cross-validazione

(applicabile sia alla previsione numerica che alla classificazione;
applicabile sia alla *model selection* che al *model assessment* (con k differenti))

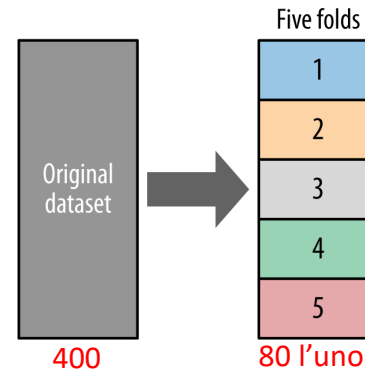
3 grandi vantaggi della CV:

- si sfrutta tutto il dataset disponibile, sia per il training che per il test; no spreco!
- eventuali combinazioni anomale sono «viste» sia dal training che dal test
- si ottiene una distribuzione della performance di test (anziché un solo valore, per quel seme)

Rimescolamento dati iniziale

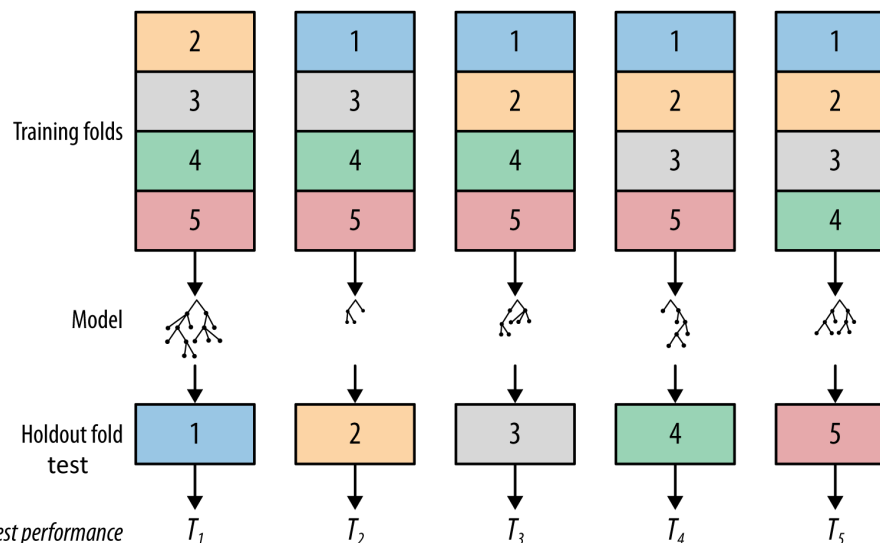
(metodo pandas *sample* oppure l'argomento *shuffle* della classe *kfold* di scikit-learn, oppure il widget *Data Sampler* di Orange)

Regola pratica di Hastie & Tibshirani:
5-10 fold (*n_splits* in scikit_learn)



K alberi per il test, 1 solo albero deployato (allenato su tutto il dataset)

classificazione: **error rate**
regressione: **RMSE**
(Rooted Mean Square Error)



I vari alberi (algoritmi) possono differire o per il differente training set (sempre) oppure, anche, per i parametri del modello (ad es. *max_depth* oppure *min_samples_leaf* degli alberi)

Mean and standard deviation of test sample performance

CV NESTED: per OGNI iterazione (una certa suddivisione tra training e test set) sono allenati / testati differenti modelli, ognuno con un differente valore del parametro di tuning. (CV nested non implementata in Orange!) → **vedi slide successiva**