MSL1: Resampling

THA1268, 04.02.2019

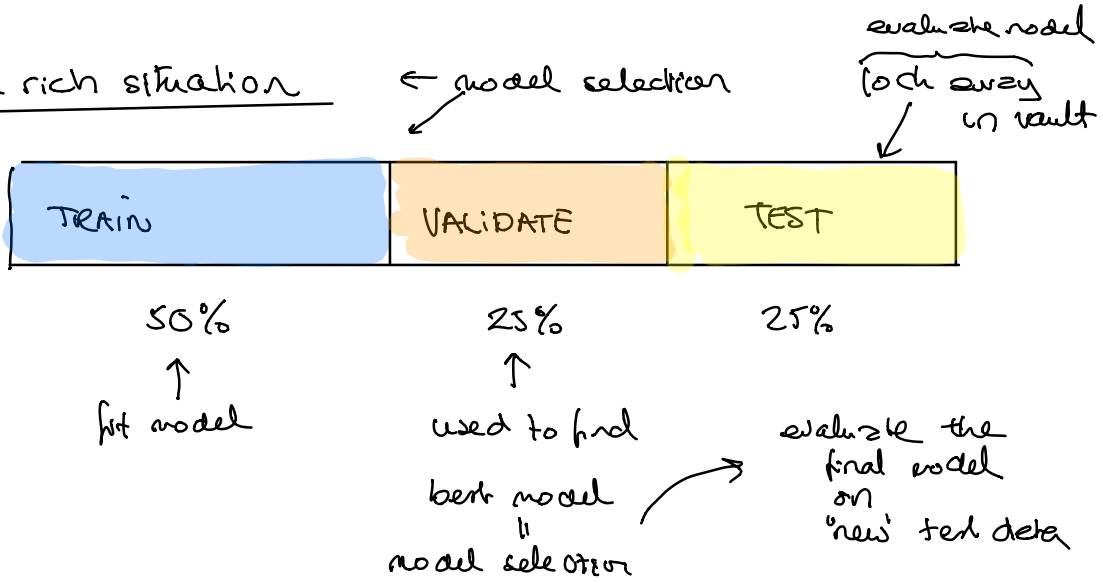
Model selection: choose between different models or different complexity of one model (e.g. kNN).

Model assessment: evaluate performance of chosen model on new data.

Criteria: regression $E[(Y - \hat{f}(x))^2]$ ← MSE (quadratic loss)

classification 0/1 loss ← misclassification rate

Data rich situation



Can we do this in a more data-efficient way?

CROSSVALIDATION

→ keep the test set in the vault → focus on model selection with the rest of the data.

1) Validation set approach

50-75%. 50-25%

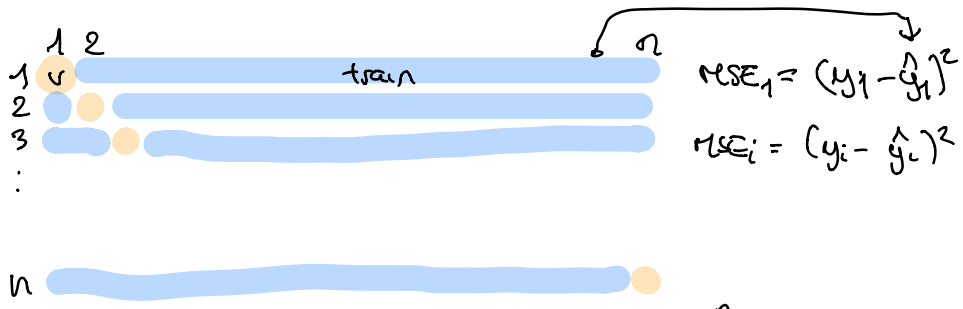
Ex: Y = miles pr gallon
X = horsepower ← poly 1, ..., 10 } fit train data for 10 models, use validation set → calc MSE

smallest validation MSE for poly 7

→ the result is dependent on the split of the data into a train/validation set.

→ if we have few data ⇒ exists better solutions!

2) LOOCV leave-one-out-cross-validation



$$CV_n = \frac{1}{n}\sum_{i=1}^{n} MSE_i$$

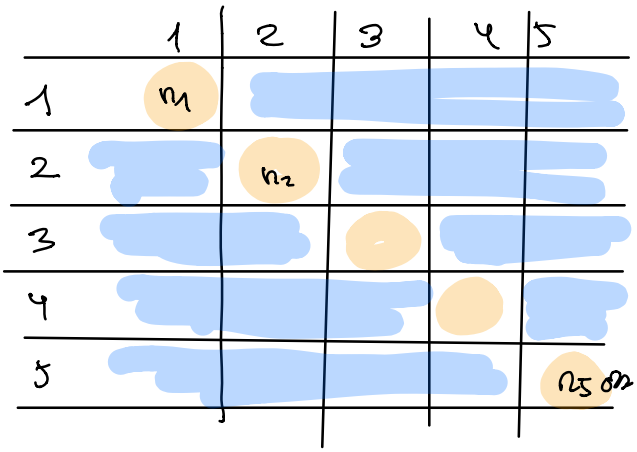
+: no randomness in splits, sample size $(n-1)$

÷ expensive + high variance

3) k-fold CV

Is $\frac{n}{k}$ an integer? If yes $n_1 = n_2 = \dots = n_k$

Shuffle data, create k folds. Draw k=5.



from train data

$$MSE_1 = \frac{1}{n_1}\sum_{i \in C_1} (y_i - \hat{y}_i)^2$$

MSE_2

.

MSE_k

C_1 = indecies for obs. in fold 1

different fold

want $\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_{i0})^2$

$$CV_5 = \frac{1}{n}\sum_{j=1}^{5} n_j MSE_j$$

often ok with $\frac{1}{5}\sum_{j=1}^{5} MSE_j$

often $n_1 \approx n_2 \approx \dots \approx n_k$ the folds of nearly eq. size

3

$k = 5$ and $k = 10$ are very popular, and recommended solutions.

How to choose the best model?

θ = model parameter

↖ poly: $\theta = 7$ is best for 5 fold

θ = some parameter to give the model

* Smallest CV error:

$$\hat{\theta} = \text{argmin } CV(\theta)$$

* One standard error rule

$MSE_1, \dots, MSE_5$: $n_1 = n_2 = \dots = n_5$

$$CV_5 = \frac{1}{n}\sum_{j=1}^{5} n_j MSE_j \approx \frac{1}{5}\sum_{j=1}^{5} MSE_j$$

$\widehat{var}(MSE)$

$$SE(\hat{\theta}) \leq \left(SE(CV_5)\right)^2 = \left(\frac{1}{5}\right)^2 \sum_{j=1}^{5} \frac{1}{4}\sum_{j=1}^{5}(MSE_j - CV_5)^2$$

$\frac{1}{5}$ — estimator for $var(MSE_j$'s)

Choose the simplest model where

$$CV(\theta) \leq CV(\hat{\theta}) + SE(\hat{\theta})$$

↑ min rule

give poly 2 in my example

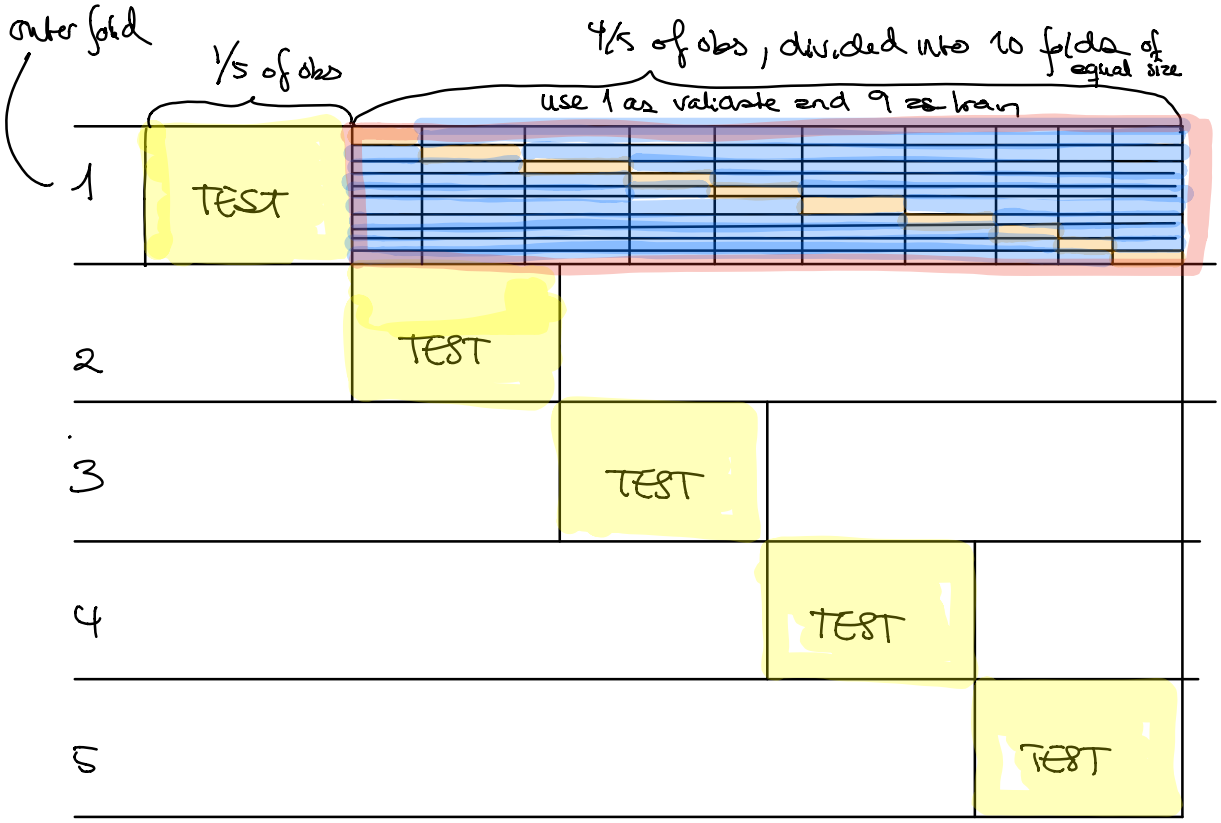
calculable as $\frac{\widehat{var}(mse)}{5}$ in R.

CV for both model selection and model assessment?

→ two nested layers of CV

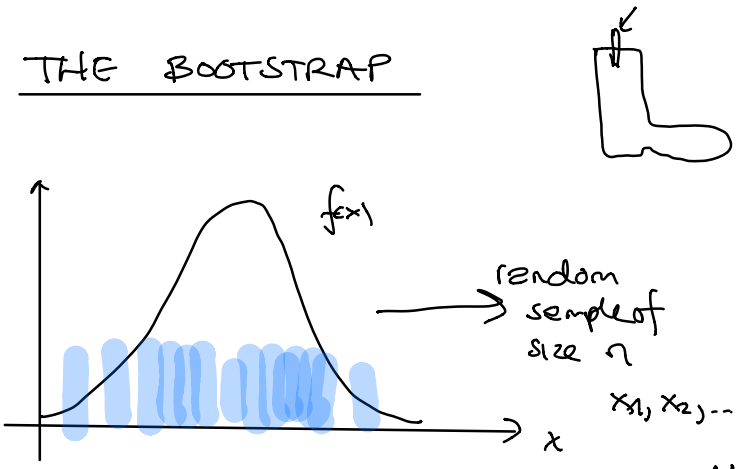
Outer loop: model assessment (draw 5-fold)

Inner loop: model selection (draw 10-fold) ← you choose!



Inner loop: decide on best model → fit to all ___ and then test on test. ↻ repeat 5 times for the outer loop

THE BOOTSTRAP



How can I find $SD(\tilde{X})$

Suggestion: We can draw many random samples, $b = 1, \dots, B$ and calculate $\tilde{X}_b$ for each sample end

$$SD(\tilde{X}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\tilde{X}_b - \frac{1}{B}\sum_{b=1}^{B}\tilde{X}_b\right)^2} \leftarrow \begin{cases} \tilde{X}_1 \\ \tilde{X}_2 \\ \tilde{X}_3 \\ \vdots \\ \tilde{X}_B \end{cases}$$

[Oh, but in general I have one sample end
I do not know f]

New = bootstrapping $\underbrace{\hat{f}(x) = \frac{1}{n} \quad \text{for each } x_1, x_2, \dots, x_n}$

same strategy as before, but now the B samples are drawn from $\hat{f}$.