Cross-validation

Part 1

Jeremy Brown

24 January 2022

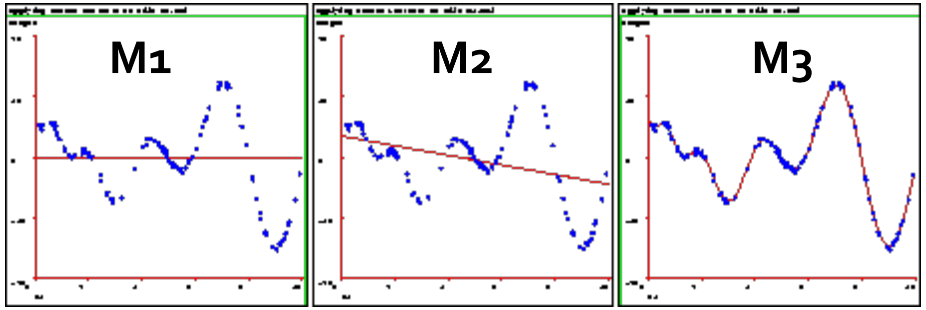
LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Breakout room group discussions (5 mins)

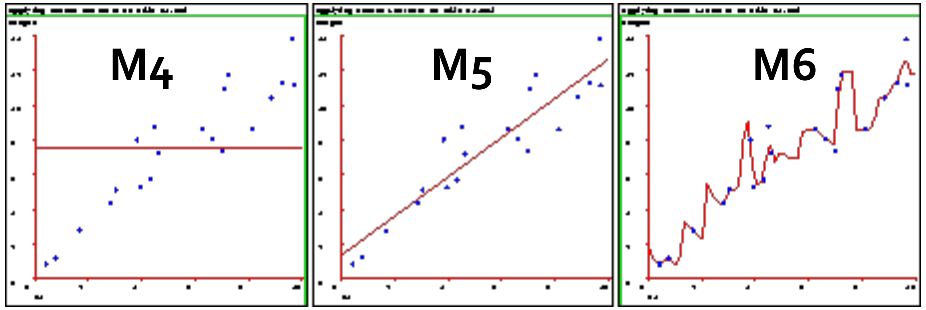
In breakout groups, spend 5 minutes working through one of these discussions:

A. You are fitting a regression model:

   1) Which model fits better? Why?



   2) Which model fits better? Why?

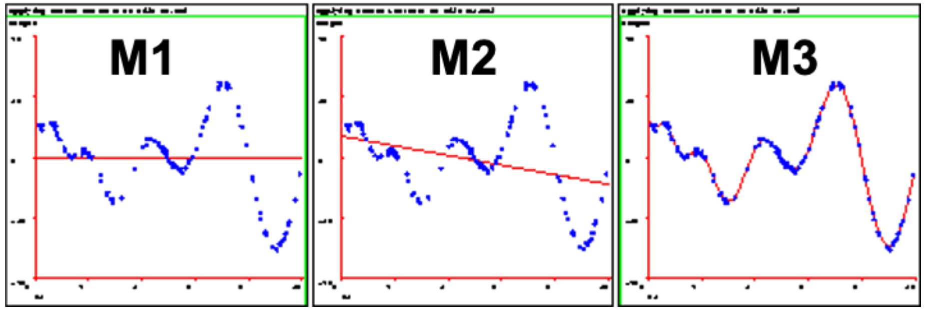


B. Why do you think it is important to do cross-validation?
   - Share any experience you might have with cross-validation

Figure source: Schneider J

1) Which model fits better?

One method of assessing the quality of a given model is by *a loss function*.



Very large MSE | Large MSE | ~0 MSE

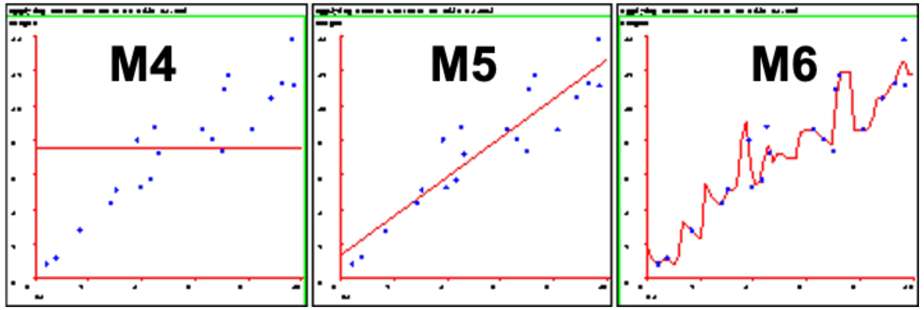
$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

Models with lower error are deemed to be better.

Figure source: Wikimedia commons, Schneider J

2) Which model fits better?

Only looking at *error on training data* can lead us astray



Very large MSE	Large MSE	~0 MSE

M4 underfits the data

M6 overfits the data

Recap week 1: over-fitting

Test/train error for classification models

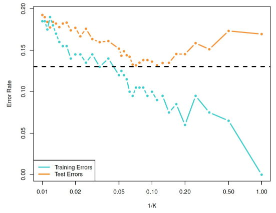
Training and Test error

Training error keeps on decreasing as K decreases.

Test error decreases at first: this shows model is getting better at fitting the test data.

When K gets too small, test error increases: the training data is not representative of the test data - this is **over-fitting**.

- Blue: training error, Orange: test error
- K decreases to the right



Alex Lewin (LSHTM) K-NN Spring 2021 14 / 16

Figure source: Wikimedia commons, Schneider J

What is cross-validation (CV)?

Previously…

Goal: build a generalisable model that can **make good predictions**: low test error rate

Training error rate

- Can be calculated on the data used to train the model
- Often can underestimate the test error rate

Test error rates

- Requires a separate set of data or cross-validation

Cross-validation refers to a set of methods for measuring the performance of a given predictive model on new data.

What is cross-validation (CV)?

A ***resampling method*** because it involves fitting the same algorithm multiple times using different subsets of the data.

Basic recipe of cross-validation techniques:

1. Divide the data into ***two* sets**:
    a. the **training data set**, used to train or build the model; and
    b. the **testing set**
2. Train the model using the **training set**
3. Use the testing set to test the model by estimating the prediction error. This will help you in gauging the effectiveness of your model's performance.

There are different cross-validation methods.

Minimising error or expected loss

Data: $D = \big((x_1, y_1), \dots, (x_n, y_n)\big)$

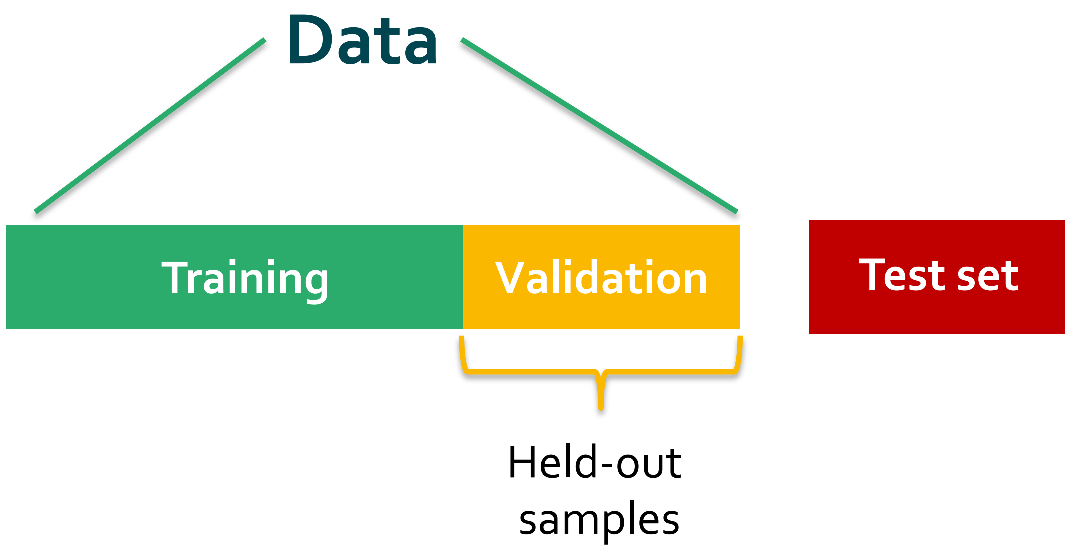
Model: $M \in (1, \dots, C)$

Expected loss: $\text{Err} = \mathbb{E}\big[\mathrm{L}\big(Y, f_m(X)\big)\big]$

True distribution of the data: $(X, Y) \sim p$

$\boldsymbol{D} = \big((\boldsymbol{x}_1, \boldsymbol{y}_1), \dots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\big)$ and $\boldsymbol{T} = \big((\boldsymbol{x'}_1, \boldsymbol{y'}_1), \dots, (\boldsymbol{x'}_t, \boldsymbol{y'}_t)\big)$ i.i.d. of p

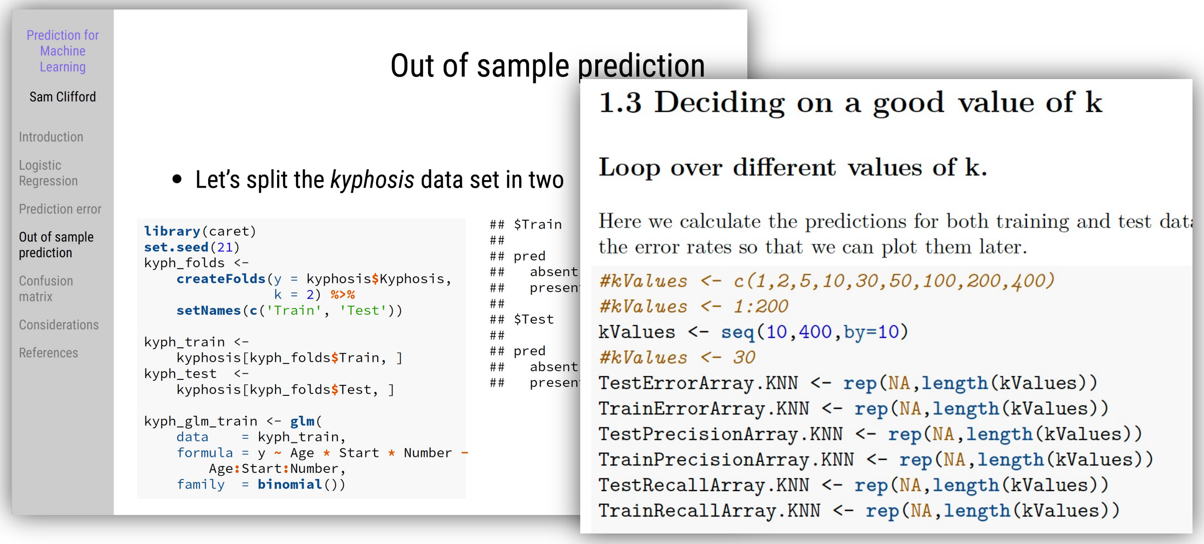
Validation set approach



Validation set approach is the most straightforward kind of cross-validation

1. Spilt ***Data*** into two sets:
   - training set
   - validation set
2. Use the **training data** to build the model
   - Train for each of the M models
3. Use **validation data** to evaluate performance
4. Choose the model with the smallest empirical error on the validation set

The validation set approach is also known as the **hold-out method.**



Prediction for Machine Learning

Sam Clifford

Introduction

Logistic Regression

Prediction error

Out of sample prediction

Confusion matrix

Considerations

References

Out of sample prediction

- Let's split the *kyphosis* data set in two

```
library(caret)
set.seed(21)
kyph_folds <-
    createFolds(y = kyphosis$Kyphosis,
                k = 2) %>%
    setNames(c('Train', 'Test'))

kyph_train <-
    kyphosis[kyph_folds$Train, ]
kyph_test  <-
    kyphosis[kyph_folds$Test, ]

kyph_glm_train <- glm(
    data    = kyph_train,
    formula = y ~ Age * Start * Number -
        Age:Start:Number,
    family  = binomial())
```

```
## $Train
##
## pred
##   absent
##   presen
##
## $Test
##
## pred
##   absent
##   presen
```

1.3 Deciding on a good value of k

Loop over different values of k.

Here we calculate the predictions for both training and test dat
the error rates so that we can plot them later.

```
#kValues <- c(1,2,5,10,30,50,100,200,400)
#kValues <- 1:200
kValues <- seq(10,400,by=10)
#kValues <- 30
TestErrorArray.KNN <- rep(NA,length(kValues))
TrainErrorArray.KNN <- rep(NA,length(kValues))
TestPrecisionArray.KNN <- rep(NA,length(kValues))
TrainPrecisionArray.KNN <- rep(NA,length(kValues))
TestRecallArray.KNN <- rep(NA,length(kValues))
TrainRecallArray.KNN <- rep(NA,length(kValues))
```

k-fold cross-validation

1. Randomly permute (or shuffle) the data

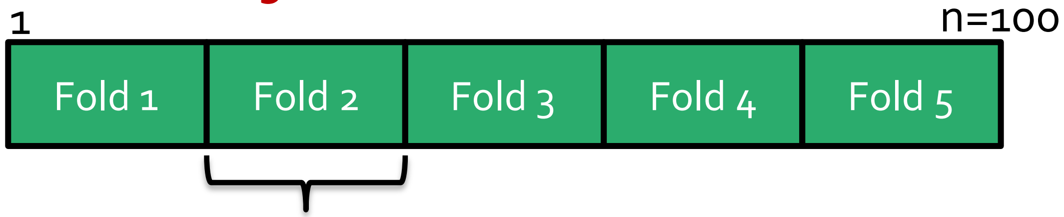
k-fold cross-validation

1. Randomly permute (or shuffle) the data
2. **Split the data into equally sized *k*-folds**
   – Choose a value for the parameter *k*

k-fold cross-validation

1. Randomly permute (or shuffle) the data
2. **Split the data into equally sized *k*-folds**
   - Choose a value for the parameter *k*

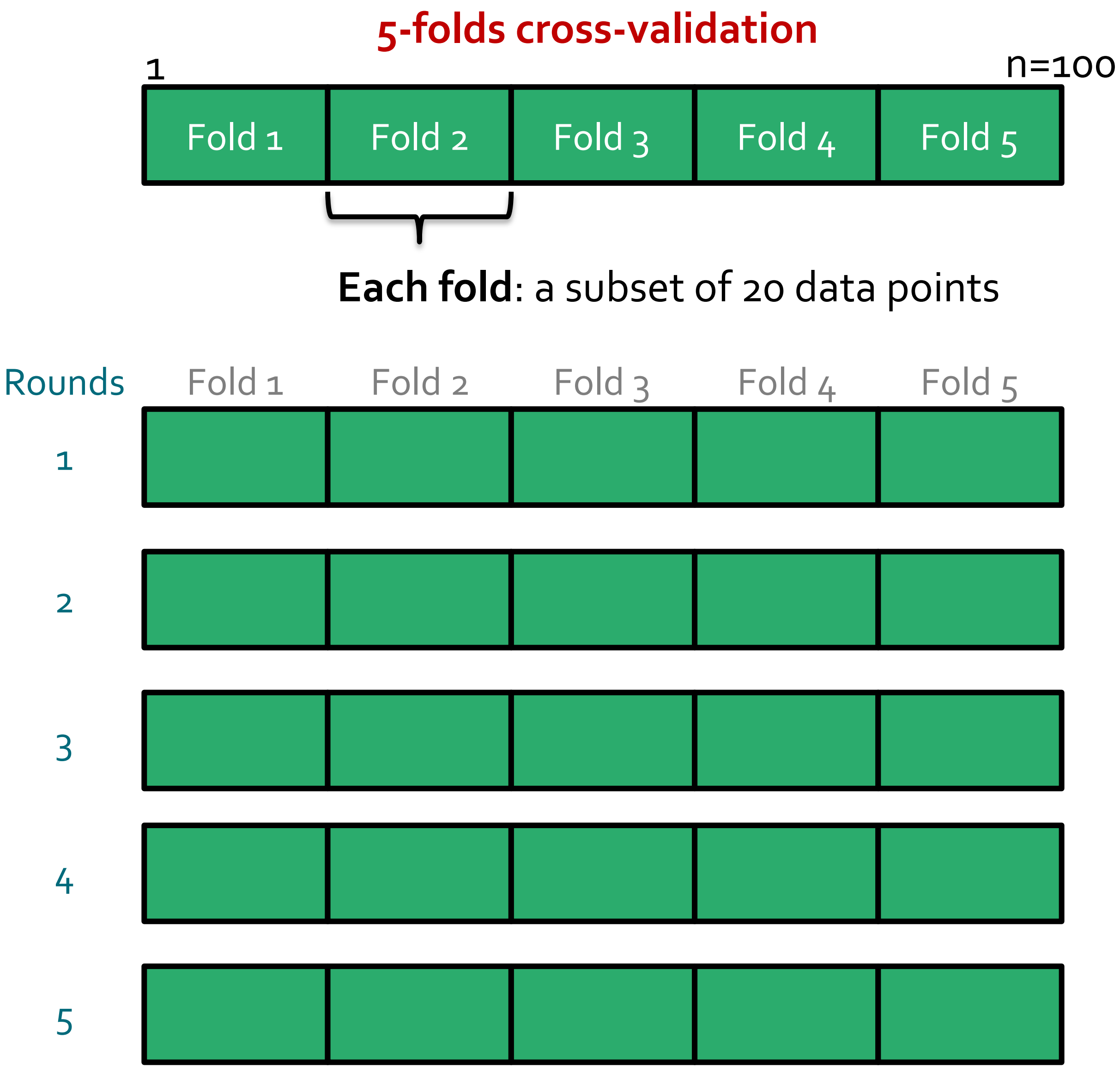
5-folds cross-validation



Each fold: a subset of 20 data points

k-fold cross-validation

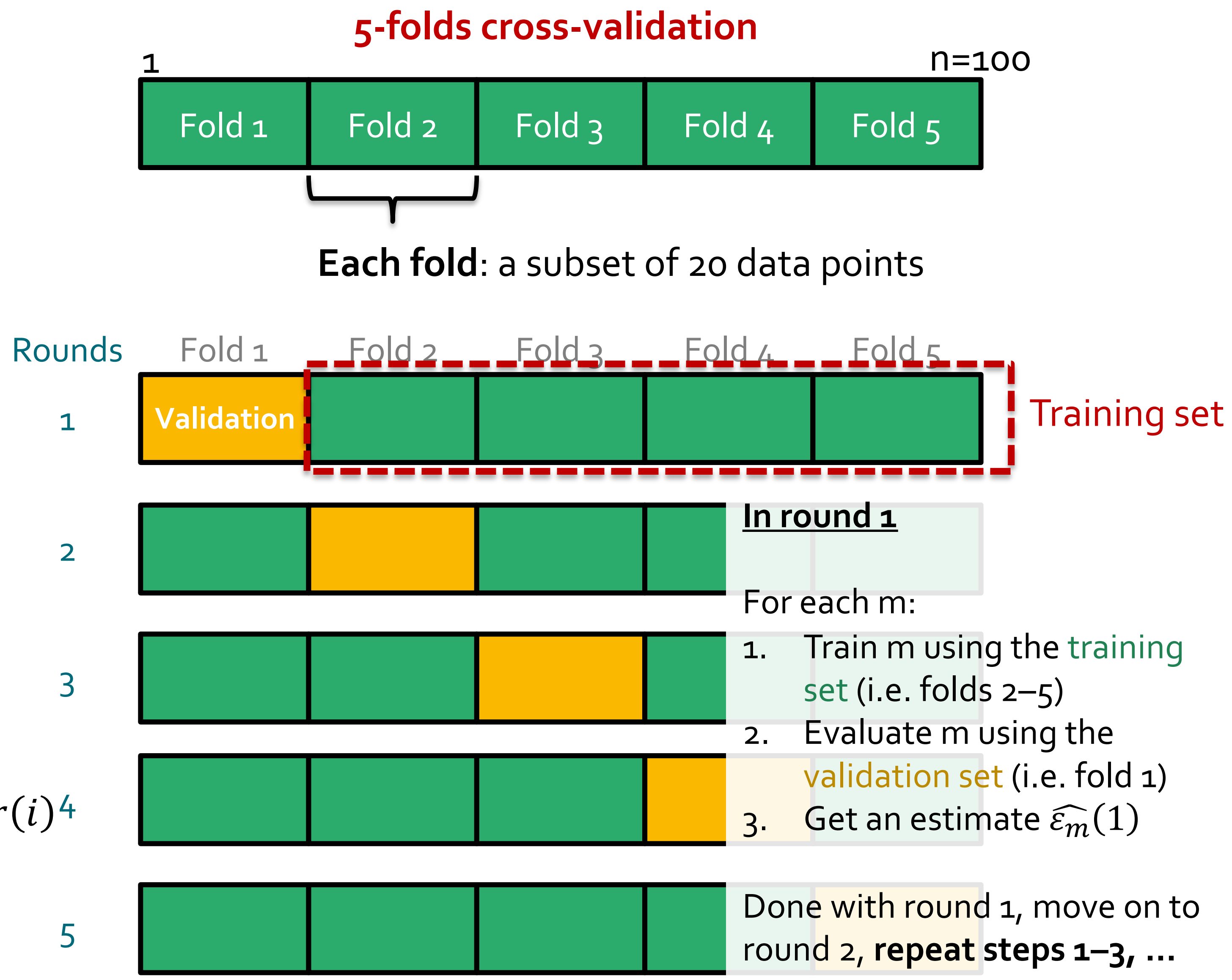
1. Randomly permute (or shuffle) the data
2. **Split the data into equally sized *k*-folds**
   – Choose a value for the parameter ***k***
3. **Validation!**
   – #rounds = #folds
   – For each round, there will be a different fold that is the validation set



k-fold cross-validation

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

1. Randomly permute (or shuffle) the data
2. **Split the data into equally sized *k*-folds**
   - Choose a value for the parameter ***k***
3. **Validation!**
   - #rounds = #folds
   - For each round, there will be a different fold that is the validation set
   - Calculate $\widehat{Err}$ across the rounds $\widehat{Err} = \frac{1}{k}\sum_{i=1}^{k=5}\widehat{Err}(i)$
4. **Choose m* to minimise** $\widehat{Err}_m$
5. **Retrain** using m* on all of D



Leave-one-out cross-validation (LOOCV)

LOOCV is a special case of the *k*-fold cross-validation

- *K = n :* the number of data points in the training set
- there will be n rounds (you don't really need to permute the data: Step 1 for k-fold cv)
  - **<u>In round 1</u>**

    For each m:

    1. Train m using the training set (i.e. folds 2–n)
    2. Evaluate m using the validation set (i.e. fold 1)
    3. Get an estimate $\widehat{Err_m}(1)$

Done with round 1, move on to round 2, **repeat steps 1–3, …**

Cross-validation

Part 2

Jeremy Brown

24 January 2022

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Breakout room group discussions (20 mins)

In breakout groups, spend 20 minutes to review k-fold cross-validation.

1. **What are the <u>advantages</u> and <u>disadvantages</u> of k-fold cross-validation relative to:**
   1. the validation set approach?
   2. the leave-one-out cross-validation (LOOCV)?

What is the optimal number of folds in k-folds cross-validation?

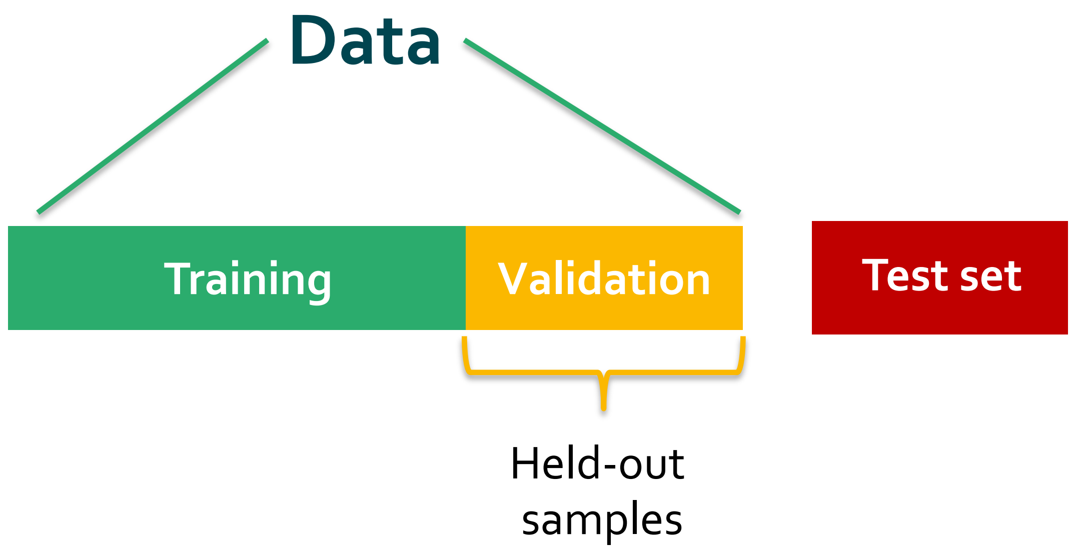
A recommended value for *k* is 10

How do we know that this configuration is appropriate for our dataset and our algorithms?

Validation set approach vs k-folds cross-validation

Model evaluation may depend heavily on which data points end up in the training set and which end up in the test set

- may be significantly different depending on how you divide the data



Optimal number of folds in k-folds cross-validation

It depends.

During cross-validation, you were averaging over independent estimates

- **LOOCV** → lower bias

What happens when training sets are highly correlated?

Correlation may increase with k (LOOCV is when k = n)

Model performance also depends on the training size.

1. If there were 100 observations in the training set
2. If there were 50 observations in the training data set

Breakout room group discussions (10 mins)

In breakout groups, spend 10 minutes

1. What is data leakage when training a model?
2. Why it is a problem?
3. How can you minimize data leakage?

Nested cross-validation

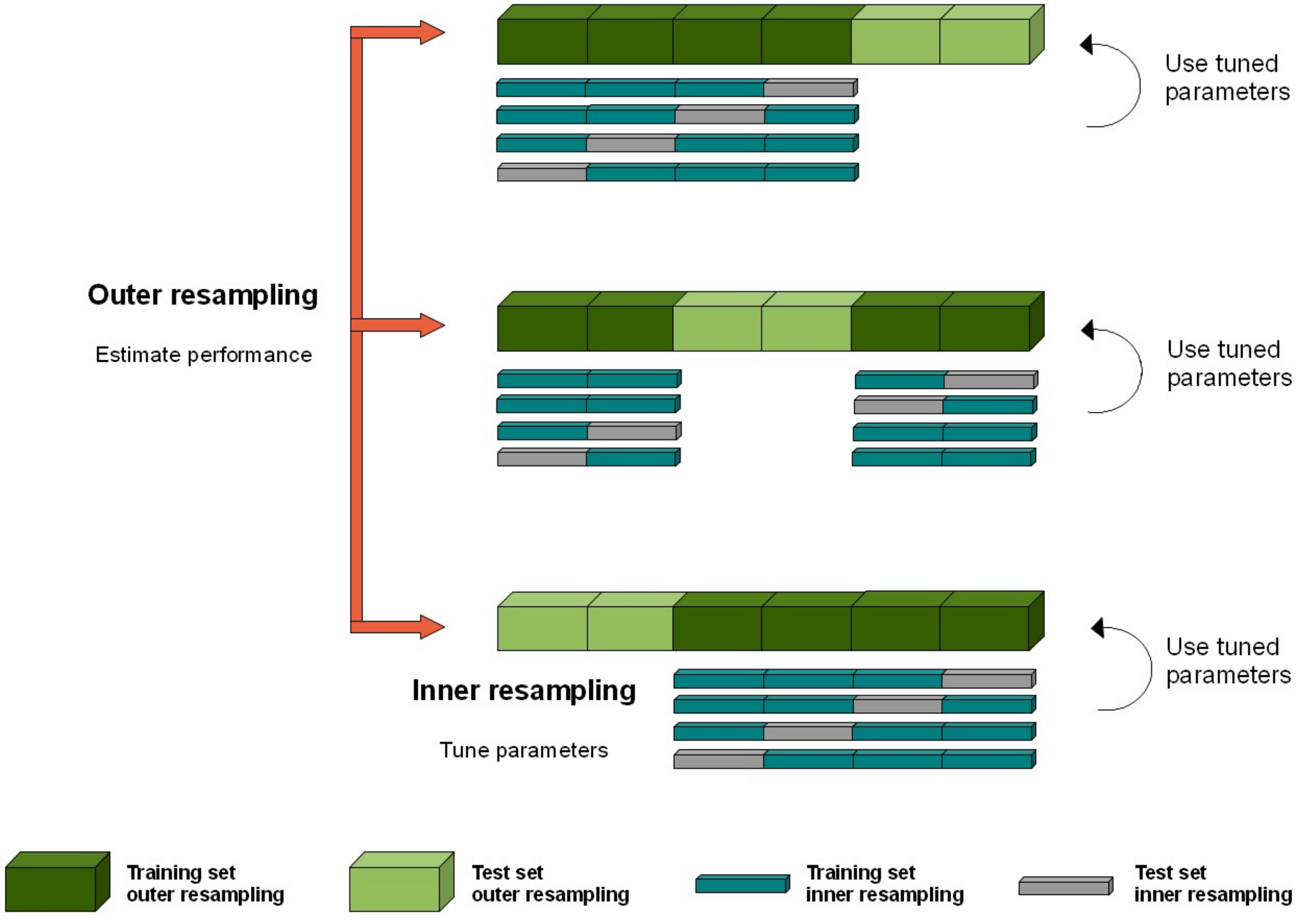
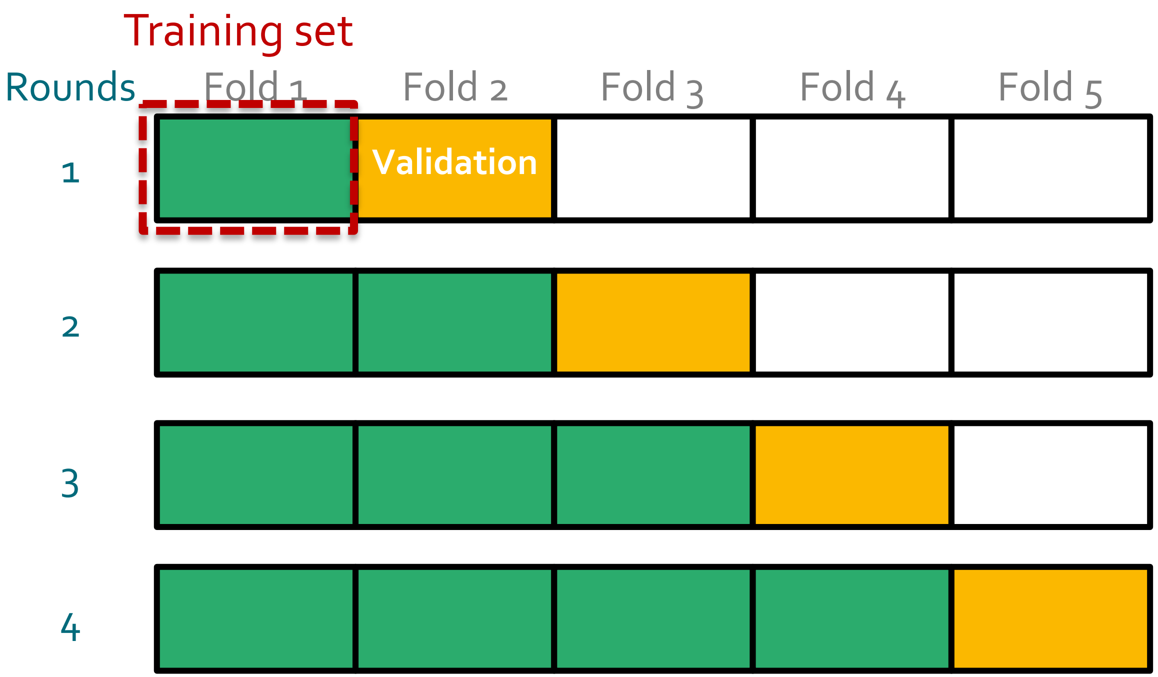


Image source: https://mlr.mlr-org.com/articles/tutorial/nested_resampling.html

Time series cross-validation

How to implement time series (or other intrinsically ordered data) cross-validation in R ?



Cross-validation on a rolling basis

https://github.com/robjhyndman/forecast

https://robjhyndman.com/hyndsight/tscv/