Imperial College London

4 December 2020

Machine Learning Workshop

Tim C.D. Lucas

tlucas@ic.ac.uk

MRC Centre for Environment & Health

UKRI Medical Research Council

Imperial College London

Imperial College London

2 Machine Learning Workshop: Workflow overview.

Please open R script lucas_ml_workshop.R and load packages.
We'll talk about the following code in a minute.

```
tr1 <- trainControl(
         method = 'LGOCV',         # Hold out data for testing
         p = 0.75,
         number = 1,
         savePredictions = TRUE)

m1 <-  train(
         time ~ .,                 # Define response and covariates
         data = melanoma,          # Select the data
         method = 'rpart2',        # Choose a model
         tuneLength = 3,           # Setup fine tuning
         metric = 'MAE',           # Define what counts as 'good'
         trControl = tr1)
```

Imperial College London

3 Workshop structure

1. Overview of machine learning workflow/single analysis.
2. Describe data and run basic analysis.
3. What is machine learning?
4. Detailed description of each stage in the analysis.
5. More information on caret.
6. What is machine learning bad at?
7. Fuller machine learning workflow.
8. Final details.

4 Machine Learning Workshop: Workflow overview.

```
tr1 <- trainControl(
         method = 'LGOCV',          # Hold out data for testing
         p = 0.75,
         number = 1,
         savePredictions = TRUE)

m1 <-  train(
         time ~ .,                  # Define response and covariates
         data = melanoma,           # Select the data
         method = 'rpart2',         # Choose a model
         tuneLength = 3,            # Setup fine tuning
         metric = 'MAE',            # Define what counts as ‘good’
         trControl = tr1)
```

5 Let's do Machine Learning: Data

Time until death data. See script.

```
data(melanoma, package = "boot")
head(melanoma)

# Remove year and discuss censoring.
melanoma <- melanoma[, -5]

# Overview of data.
featurePlot(melanoma[, -1], y = melanoma$time)
```

Imperial College London

6 Let's do Machine Learning: Basic analysis

```
tr1 <- trainControl(
         method = 'LGOCV',          # Hold out data for testing
         p = 0.75,
         number = 1,
         savePredictions = TRUE)

m1 <-  train(
         time ~ .,                  # Define response and covariates
         data = melanoma,           # Select the data
         method = 'rpart2',         # Choose a model
         tuneLength = 3,            # Setup fine tuning
         metric = 'MAE',            # Define what counts as ‘good’
         trControl = tr1)
```

```
CART

205 samples
  5 predictor

No pre-processing
Resampling: Repeated Train/Test Splits Estimated (1 reps, 75%)
Summary of sample sizes: 156
Resampling results across tuning parameters:

  maxdepth  RMSE      Rsquared   MAE
  1         890.3811  0.2915851  697.8748
  2         856.7927  0.3493889  687.3165
  3         831.3940  0.3999163  645.2901

MAE was used to select the optimal model using the smallest value.
The final value used for the model was maxdepth = 3.
```

8 Any questions?

(This slide will occur many times in this slidedeck.)

9 What is machine learning?

- Focus on prediction.
- Not mechanistic/process based models.
- Not inference of real-world parameters.

10 What is machine learning?

- An analytical aim, rather than a group of models.
- Linear regression is machine learning if you mostly care about prediction!
- Also Neural networks, decision trees, random forest.

11 What is machine learning?

Tasks:

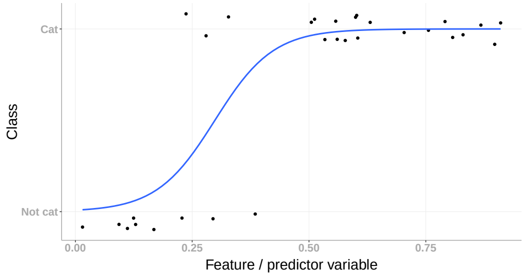
- Supervised learning
  - Covariates and response data. Like most biological models.
- Reinforcement learning
  - Make your own data.
- Unsupervised learning
  - Clustering.

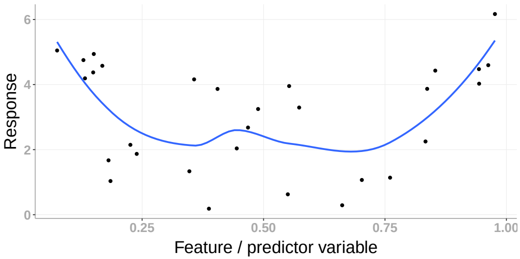
Imperial College London

12 What is machine learning?

Supervised learning.

Classification or regression.





13 Any questions?

14 Out-of-sample validation

```
tr1 <- trainControl(
        method = 'LGOCV',
        number = 1,
        p = 0.75,
        savePredictions = TRUE)

m1 <- train(time ~ .,
            data = melanoma,
            method = 'rpart2',
            tuneLength = 3,
            metric = 'MAE',
            trControl = tr1)
```

15 Out-of-sample validation

- k-fold
  - split into k group. user each group on turn as hold out.
- Repeated k-fold
  - Do k-for multiple times with different random splits.
- Bootstrap
  - Sample N with replacement as training.

16 Out-of-sample validation

- Seperate study
- Spatial
- Temporal
- By covariates
- What question do you want to ask?

Imperial College London

17 Outer validation

- Selecting a model *is part of the model*.
- If we consider many models, taking the best one is a form of overfitting
- Outer cross-validation if this is part of primary research question.
- Unfortunately not implemented in caret. Must do it manually.
- AKA Train, test, validate.

18 Any questions?

19 Let's do Machine Learning: Error metrics.

```
tr1 <- trainControl(
        method = 'LGOCV',
        number = 1,
        p = 0.75,
        savePredictions = TRUE)

m1 <- train(time ~ .,
            data = melanoma,
            method = 'rpart2',
            tuneLength = 3,
            metric = 'MAE',
            trControl = tr1)
```

20 Error metrics: regression

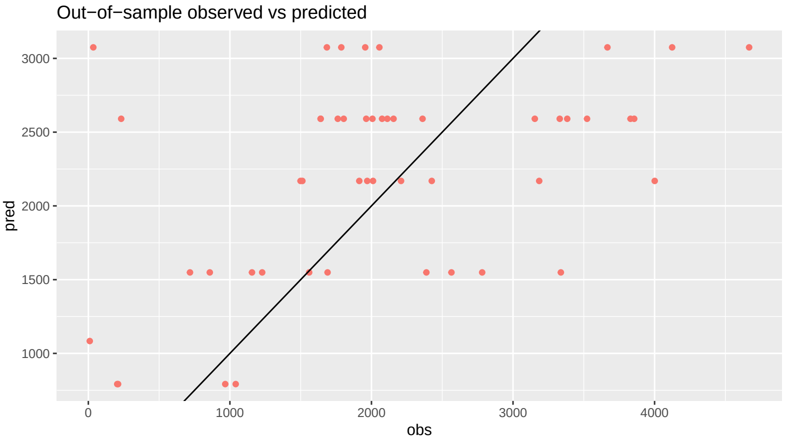
- Match error metric to question.
- Aim for interpretability.
- $MAE = \text{mean}(\text{abs}(\text{pred} - \text{obs}))$
- R^2
- RMSE, correlation are less interpretable
- Correlation isn't quite measuring what you think.
- Scatter plots!

21 Error metrics: regression

- Scatter plots!
- Annoyingly caret doesn't have a function that plots obs vs preds of the hold out data.
- I have written my own plotCV() function. We'll load it later.

Imperial College
London

22 Error metrics: regression



23 Error metrics: classification

- Match error metric to question.
- Aim for interpretability.
- Class balance. Model that never predicts rare class will have high accuracy.
- I can predict COVID infection with $\sim$97% accuracy by saying noone has COVID.
- Kappa, AUC.
- Accuracy does *not* account for imbalance.
- Confusion matrices.

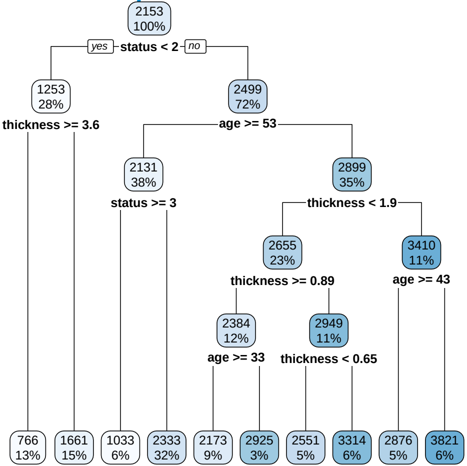
24 Any questions?

25 Chosing a method (surprisingly unimportant)

```
tr1 <- trainControl(
        method = 'LGOCV',
        number = 1,
        p = 0.75,
        savePredictions = TRUE)

m1 <- train(time ~ .,
            data = melanoma,
            method = 'rpart2',
            tuneLength = 3,
            metric = 'MAE',
            trControl = tr1)
```

Imperial College London

26 Chosing a method: rpart2



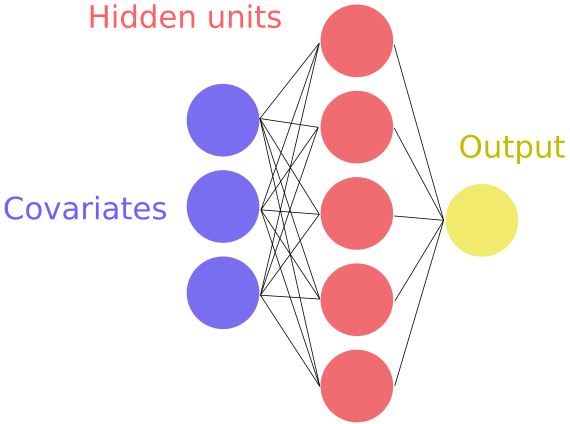
27 No free lunch theorem

- No such thing as a universal, ‘best’ machine learning model.
- So try a few.

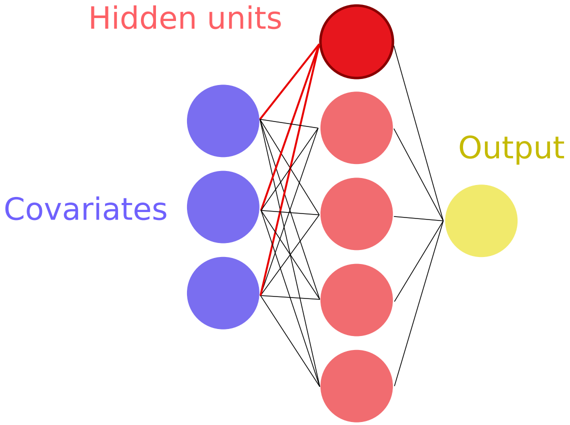
28 What is a RandomForest?

- method = 'rf' or method = 'ranger' in caret.
- Instead of 1 tree, fit many trees and take average prediction.
- For each tree take a random bootstrap of the data.
- For each node consider a random subset of covariates.
  - Consider mtry covariates. Tuning parameter.

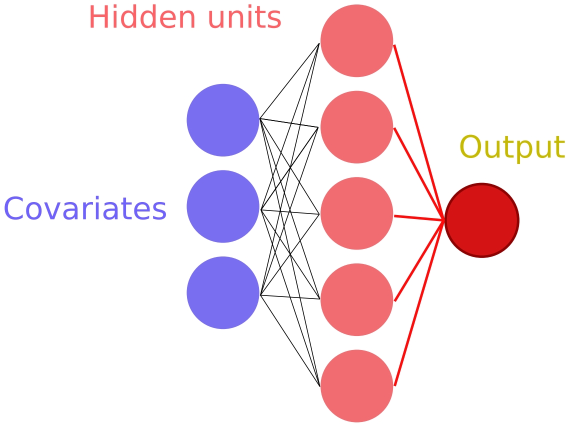
29 What is a Neural Network?



30 What is a Neural Network? Little GLMs.



31 What is a Neural Network? Little GLMs.



32 What is a Neural Network?

- method = 'nnet', method = 'mlpKerasDropout' or many others.
- Optimise the parameters but there are lots of local optima.
- What "architecture"?
  - Hidden units.
  - Extra hidden layers
  - Everything connected to everything else?

Imperial College London

33 Coffee. Any questions?

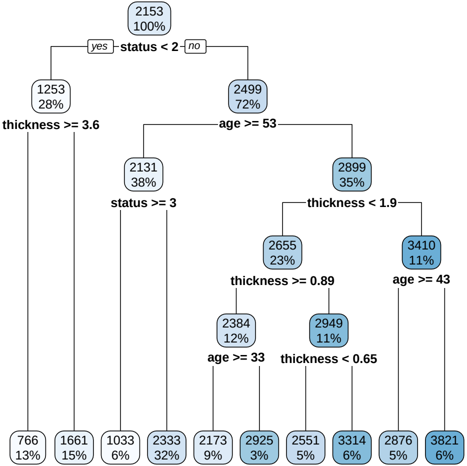
Imperial College London

34 Tuning/hyperparameters

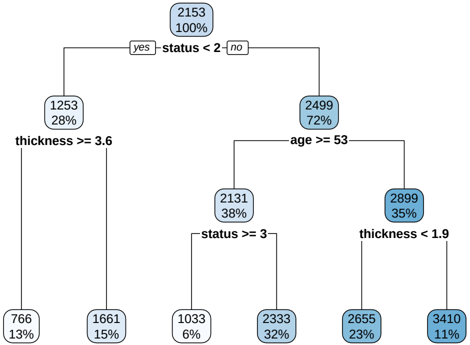
```
tr1 <- trainControl(
        method = 'LGOCV',
        number = 1,
        p = 0.75,
        savePredictions = TRUE)

m1 <- train(time ~ .,
            data = melanoma,
            method = 'rpart2',
            tuneLength = 3,
            metric = 'MAE',
            trControl = tr1)
```

Imperial College
London

35 Maxdepth parameter

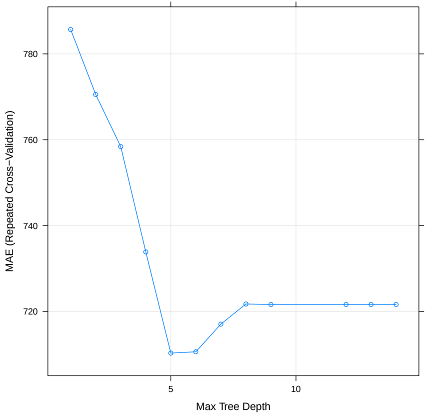


36 Regularisation: forcing a model to be simpler



37 How do we choose? Out-of-sample performance.

Try `tuneLength = 10` values.

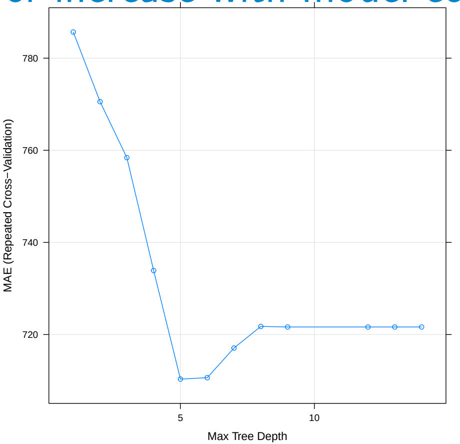


Imperial College London

38 Other tuning parameters.

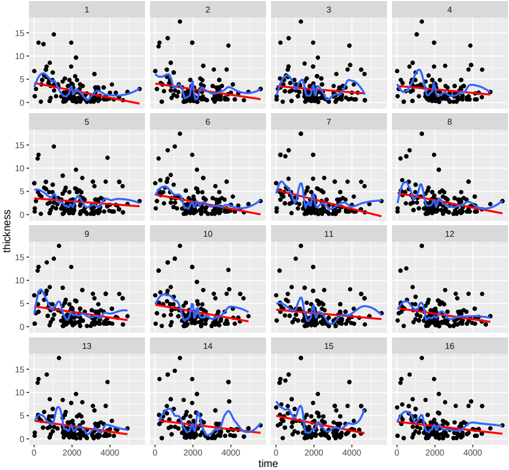
- Stepwise regression cutoff.
- Degree of freedom in GAM.
- Length scale in Gaussian Process.
- Variance of zero-mean Bayesian prior.

39 Why does error increase with model complexity?



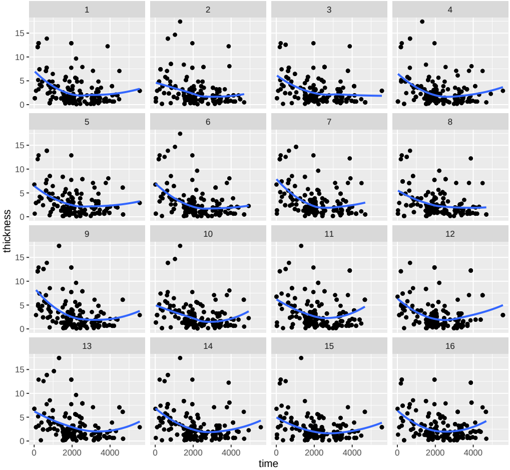
Imperial College London

40 Overfitting and bias/variance.



Imperial College
London

41 Overfitting and bias/variance.



42 Any questions?

43 Caret package

- https://topepo.github.io/caret/model-training-and-tuning.html
- Unified interface to hundreds of models
- Supervised learning
- Full ML workflow
- Excellent documentation

44 Caret details

1. Fit model to all training sets and predict all test sets with all hyperparameter combinations.
2. Select best hyperparameter combination.
3. Fit model using best hyperparameter set to all data.
4. If y is a factor, automatically do classification.
5. Some models are only regression, some only classification, some both.
6. Caret will ask you if you want to install additional packages.
7. Some models don't handle NAs. Error messages not always very useful.
8. Check documentation for parallel computation.

45 Caret documentation

- https://topepo.github.io/caret/index.html
- Really excellent!
- Model list with tuning parameters.
- Models by tag (data weights, tree-based, implicit feature selection).

46 Caret functions and internals

```
plot(m1)

# Uses model trained on full dataset.
# Use this to test on a outer validation dataset.
predict(m1)

m1$results # Validation results.
m1$pred # All validation predictions (all hyperpars)
m1$finalModel # The final fitted model
class(m1$finalModel)
```

47 Caret functions and internals

```
# Load plotCV() and best_tune_preds() functions.

plotCV(m1)
```

48 Grid search for models with many hyperparameters

```
tr_random <- trainControl(
              search = 'random',
              savePredictions = TRUE)

m_random <- train(time ~ .,
            data = melanoma,
            method = 'enet',
            tuneLength = 20,
            metric = 'MAE',
            trControl = tr_random)
```

Imperial College London

49 Grid search for models with many hyperparameters

```
# Give an explit dataframe of parameters
# Need to look up the exact names

gr <- data.frame(lambda = c(1e-4, 1e-5, 1e-6),
                 fraction = c(0.1, 0.5, 0.5))
m_df <- train(time ~ .,
              data = melanoma,
              method = 'enet',
              tuneGrid = gr,
              metric = 'MAE',
              trControl = tr1)
plot(m_df)
```

50 Any questions?

51 What is ML bad at?

```
pl <- read.csv(
  file = 'https://raw.githubusercontent.com/timcdlucas/ml_workshop/master/p

pl1 <- train(g ~ .,
             data = pl,
             method = 'rpart2',
             trControl = tr1)
```

Imperial College London

52 What is ML bad at?

```
pl <- read.csv(
  file = 'https://raw.githubusercontent.com/timcdlucas/ml_workshop/master/p

pl2 <- train(g ~ 0 + I(m1 * m2 / d ^ 2),
             data = pl,
             method = 'lm',
             trControl = tr1)
```

53 Which was better? Any questions?

- Thumbs up: rpart2.
- Smiley face: lm.

54 Short coffee.

55 Fuller ML workflow

```
# Carefully think about hold out data
# This is THE most important part.
tr2 <- trainControl(
        method = 'repeatedcv',
        number = 5,
        repeats = 3,
        savePredictions = TRUE)
```

56 Fuller ML workflow

```
# Carefully choose a metric
my_metric <- 'MAE'
```

57 Fuller ML workflow

```
# Collect data.
# Make new covariates. GDP growth, sum of air pollution last week.
# Covariates are more important than algorithms.

# Log, sqrt, sqaured for linear models.
# Tree models focus on combining covariates.
```

58 Fuller ML workflow

```
# Baseline linear model
m1 <- train(time ~ .,
            data = melanoma,
            method = 'enet',
            tuneLength = 10,
            metric = my_metric,
            trControl = tr2)


# Look at scatter plots for regression
# Look at confusion matrix for classification
plotCV(m1)
```

59 Fuller ML workflow

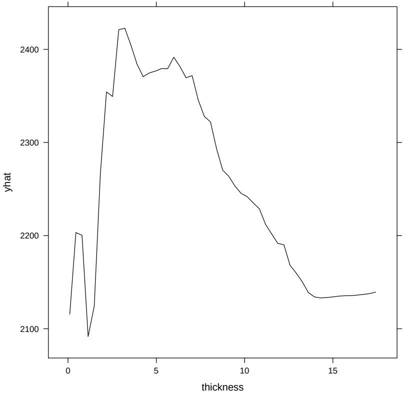
```
# Projection pursuit regression is very fast but nonlinear.
#   Another good baseline.
m2 <- train(time ~ .,
            data = melanoma,
            method = 'ppr',
            tuneLength = 10,
            metric = my_metric,
            trControl = tr2)


# Look at scatter plots for regression
plotCV(m2)
```

60 Fuller ML workflow

```
# Random Forest often peforms very well.
# Easy to tune. Another excellent baseline.
m3 <- train(time ~ .,
            data = melanoma,
            method = 'ranger',
            tuneLength = 5,
            metric = my_metric,
            trControl = tr2)

plotCV(m3)
pdp::partial(m3, pred.var = c('thickness'), plot = TRUE)
```

61 Partial dependence plot.

Fitted relationship between thickness and time.



62 Fuller ML workflow

```
tr2_random <- trainControl(
        method = 'repeatedcv',
        number = 5,
        repeats = 3,
        search = 'random',
        savePredictions = TRUE)

m4 <- train(time ~ .,
            data = melanoma,
            method = 'xgbTree',
            tuneLength = 3,
            metric = my_metric,
            trControl = tr2)
```

63 Fuller ML workflow

```
# Try a few models. No free lunch.
m5 <- train(time ~ .,
            data = melanoma,
            method = 'nnet',
            tuneLength = 10,
            metric = my_metric,
            linout = TRUE,
            trControl = tr2)

# Neural networks need linout = TRUE for regression.
# linear output as apposed to logit/probit.
```

Imperial College London

64 Any questions?

65 Other packages

- mlr3
  - I find it more complicated.
  - Probably better for very complicated pipelines.
- tidymodels
  - Fits into tidyverse.
  - More complicated.
  - Not yet feature complete.
- scikit.learn in Python
- caretEnsemble for ensembles/model averaging.

Imperial College London

66 Extra reading

- Breiman 2001 Statistical Modeling: The Two Cultures.
- Molnar 2020 Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.
- Kuhn 2013 Applied Predictive Modeling.
- Lucas 2020 A translucent box: interpretable machine learning in ecology.
- Bhatt et al. 2017 Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization

Imperial College London

67 Any questions?

tlucas@ic.ac.uk

@StatsForBios

MRC Centre for Environment & Health

Medical Research Council

Imperial College London