Nested cross-validation
Bootstrapping

Outline

- Data leakage
- Nested cross-validation
- Bootstrapping

Data leakage

> *… is when information from outside the training dataset is used to create the model. This additional information can allow the model to learn or know something that it otherwise would not know and in turn invalidate the estimated performance of the model being constructed.*
>
> https://machinelearningmastery.com/data-leakage-machine-learning/

Column-wise leakage

> *If any other feature whose value would not actually be available in practice at the time you'd want to use the model to make a prediction, is a feature that can introduce leakage to your model.*
>
> https://en.wikipedia.org/wiki/Leakage_(machine_learning)

Examples: 'MinutesLate' is included in the training of a model to predict 'IsLate'

Row-wise leakage

> *… is caused by improper sharing of information between rows of data.*
>
> https://en.wikipedia.org/wiki/Leakage_(machine_learning)

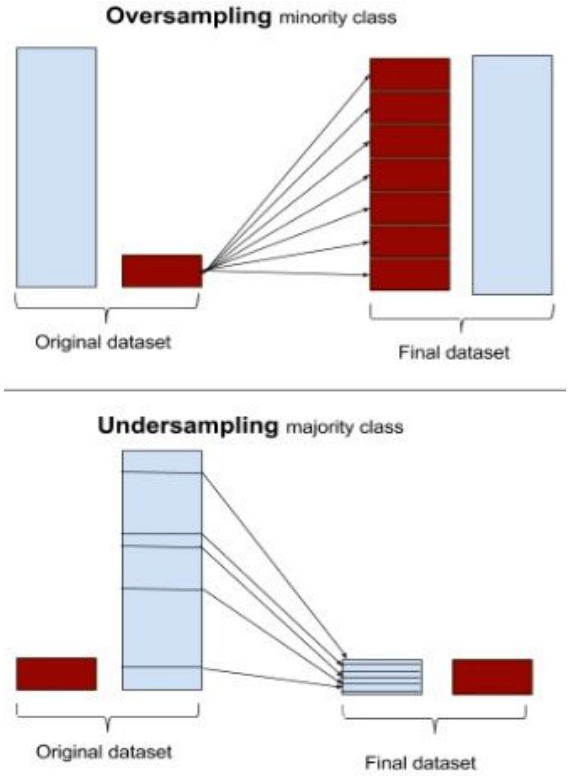
Examples: data rows used for both training and validation/testing or premature featurisation

Data leakage

Minimising risks of leakage

- Perform data preparations within each CV fold
  - Column-wise: centering, standardising, one-hot-encoding, etc. of columns in each fold
  - Row-wise: oversampling and undersampling

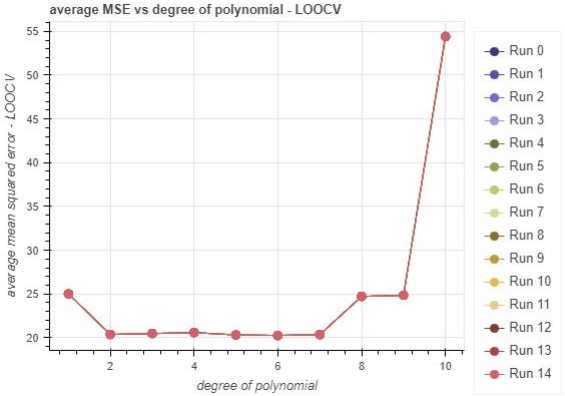
- Completely separate a hold-out dataset for final testing of the model.

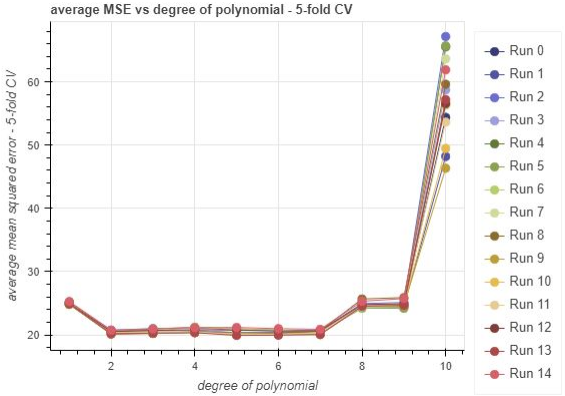


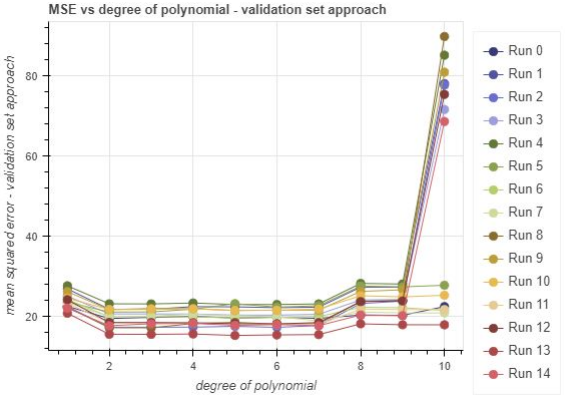
https://stats.stackexchange.com/questions/351638/random-sampling-methods-for-handling-class-imbalance

Data leakage

Recall the last lecture







- We decided that the quadratic model is the simplest model that gives the smallest CV error (model selection)
- Then evaluated the model's MAD 95% confidence interval using 5-fold CV on the same dataset (model evaluation)

Did we commit the sin of data leakage? We did (unintentionally)!

Nested cross-validation

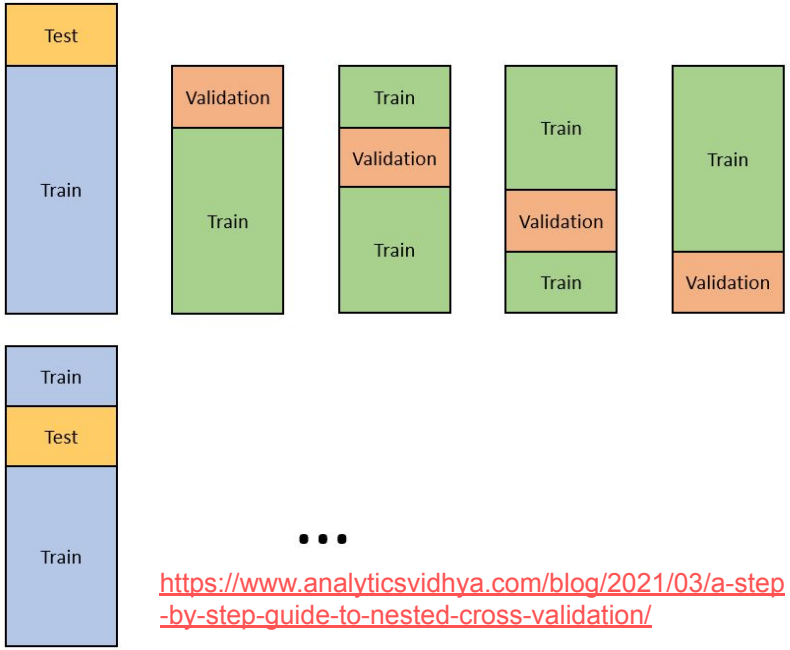
> *In order to overcome the bias in performance evaluation, model selection should be viewed as an integral part of the model fitting procedure, and should be conducted independently in each trial in order to prevent selection bias and because it reflects best practice in operational use.*

https://www.jmlr.org/papers/volume11/cawley10a/cawley10a.pdf

CV used in model selection **or** model evaluation, not both.

To do both, a k-fold CV for model selection is nested inside a k'-fold CV for model evaluation (k can be different to k')

Nested CV keeps a hold-out subset of the data completely separate from the training dataset and any preparations performed on the training set



https://www.analyticsvidhya.com/blog/2021/03/a-step-by-step-guide-to-nested-cross-validation/

Nested cross validation procedure

For each of the k outer folds

- For each of the k’ inner folds, fit candidate models (parameters) on the training set and calculate test error (MSE or MAD) on validation set.

For each candidate model (parameter), we have (k x k’) values of test error. The best model (parameter) is the one with the smallest average of these k*k’ test errors.

We then fit this best model on the whole training-validation set of each outer fold and calculate the test error (on the test set) for **model evaluation**.

Nested cross validation - An alternative procedure

For each of the k outer folds

- For each of the k’ inner folds, fit candidate models (parameters) on the training set and calculate test error (MSE or MAD) on validation set.
- Determine the best model (parameter) for this outer fold, i.e. the one with smallest average of the k’ test errors.

Now it is possible that we may end up with k different best models (one for each outer fold). Which one should we choose?

We can build a model which comprises all these best models by using *ensemble methods*.

We then fit this ensemble model on the whole training-validation set of each outer fold and calculate the test error (on the test set) for **model evaluation**.

Bootstrapping

An extremely powerful statistical tool to quantify the variability of

- a population parameter, e.g. in building the sampling distribution of a population proportion
- a machine learning method, e.g. estimating the standard errors of OLS coefficients

The bootstrapping produces distinct versions of the original data by repeatedly sampling observations *with replacement* from the original data.

Bootstrapping

Example 1. Use bootstrapping to estimate the 95% confidence interval of OLS coefficients of the regression model mpg ~ horsepower (using the Auto dataset).

Solutions. The main steps include

- Generating 1000 versions of the Auto dataset, each has 397 (possibly duplicating) observations randomly picked from the original Auto dataset.
- Fit the regression model mpg ~ horsepower to each of these 1000 datasets.
- Calculate the standard deviation of the coefficients from these 1000 models.

How are the standard deviation results compared to the those calculated using the traditional approach (see results in the Python notebook in lecture 1)?

Exercises

Use bootstrapping to estimate the 95% confidence interval of OLS coefficients of the regression model mpg ~ horsepower + horsepower^2 (using the Auto dataset).

Compare the results against the 95% confidence interval of the coefficients calculated by the traditional method (see the Python notebook in lecture 1).

What conclusion can we make from this exercise and from the above example?