ACTL3142 Week 7 - Cross Validation and Regularisation

Tadhg Xu-Glassop

2025T2

Motivating Cross Validation

Say we want to make a model to make predictions on unseen data. So, we should directly test how our models would work on unseen data and pick one that performs best!

Motivating Cross Validation

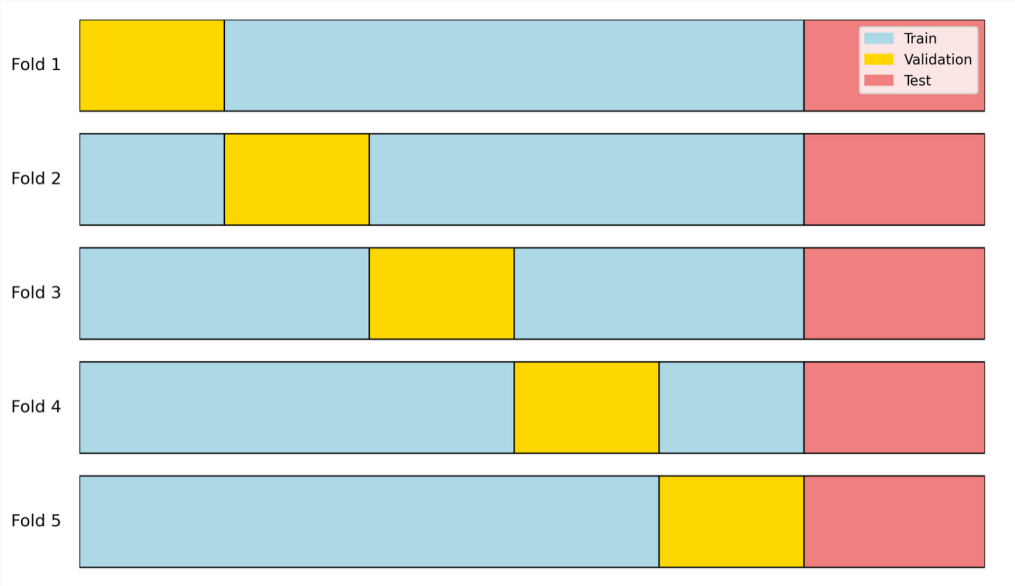
Say we want to make a model to make predictions on unseen data. So, we should directly test how our models would work on unseen data and pick one that performs best!

So far, we've seen splitting the data into train/test set. But, this is quite flawed:

- Comparing lots of models on a single test set puts tons of bias towards a model that happens to perform well on the test set but not necessarily the population; overfitting to the test set.
- If the test set set is small relative to the training, then the above is exacerbated.
- If the test set is large, then we're not using the full potential of our data in training, and so we'll get worse models, leading to biases in training error predictions.

Cross Validation

1. Set aside a test set, and split remaining data into k "folds."
2. For each fold, train the model on the remaining $k-1$ folds and 'validate' the model on that fold wrt to some **score**.
3. Take the average of all k results, and this is the cross-validated score.
4. Pick the model with the best CV score, and train it using all the data and test it using the test set to get a less-biased estimate of future performance.



Choice of k

There is bias-variance tradeoff in the choice of k:

- If k is small, then the folds are very large, so the models aren't being trained on a lot of data and possibly underperforming, leading to more bias, but less variable since our CV scores are done on lots of data.
- If k is large, then the models will be very similar as they are being trained on similar datasets, leading to correlated predictions and increased variance, but less biased since the models are taking more advantage of the whole dataset.

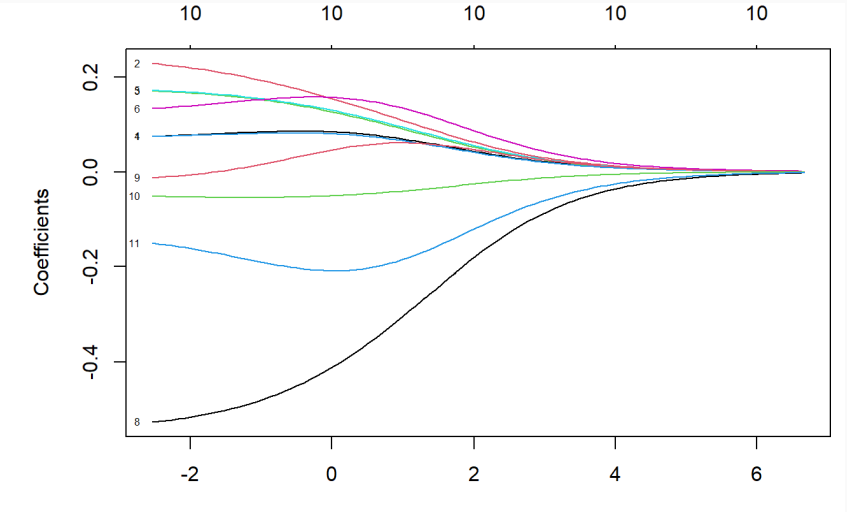
Typically just pick $k = 5$ or $k = 10$.

Ridge Regression

Ridge regression is the optimal solution to

$$\text{RSS} + \lambda \sum_{i=1}^{p} \beta_j^2.$$

λ is a *hyperparameter* that controls the flexibility of the model to achieve an optimal bias-variance tradeoff.

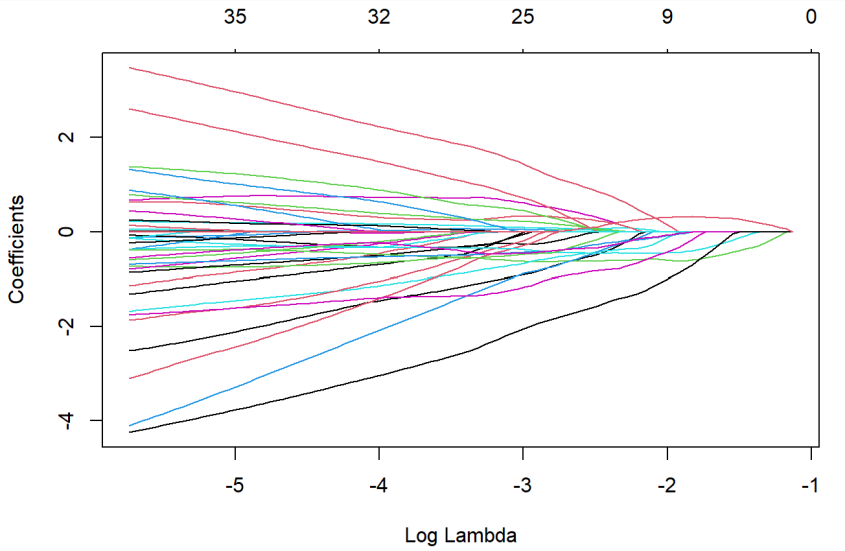


Shrinkage Methods (cont.)

LASSO regression is the optimal solution to

$$\text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|.$$

λ plays an identical role, but LASSO shrinks coefficients to 0.



Notes on shrinkage methods

To find λ, we use CV.

LASSO regularisation does **feature selection** by shrinking less significant predictors to 0.

Can train GLM's with ridge/LASSO penalty, replacing RSS with deviance.