Practical 11

Jumping Rivers

Penalised Regression

To really see how the effect of penalised regression we are going to use some synthetic data. The `sklearn.datasets` module has some data generators which are useful for exploring different modelling techniques. The reason this is useful is that by construction we know what we should be expecting to find when using a given technique.

```
from sklearn.datasets import make_regression

X, y, ground_truth = make_regression(
  n_samples=1000, n_features=1000,
  n_informative=100, effective_rank=20,
  random_state=2019, noise=0.2, coef=True
)
```

This code generates some synthetic data that has a large number of useless features and correlation amongst the input variables. This is a situation where we expect standard linear regression to perform poorly.

a) Build a standard linear regression model, remember to include some preprocessing

b) Use 10 fold cross validation to estimate mean squared error

c) In a similar fashion to the notes on Lasso, can you find a lasso regression model that performs better than this linear regression

d) How do the two sets of coefficients compare?

e) How do ridge and elastic net compare?