Practical 3

Jumping Rivers

During the lecture we fit a logistic regression model to the breast cancer data for classifying tumors in patients. We are going to fit a KNN classifier to the same data.

- Construct the pipeline ready for fitting the model
- We want to find the best value of K for the classifier when optimising for recall, our motivation is that we want to correctly identify as many of the malignant tumours as possible. Start with a grid search over `k = [1,5,10,20,50,100]`
- Create a plot of the K parameter against the average recall score found in the cross validation grid search
- What region of K looks like it will give the best value?
- Re-run your grid search across that region
- What is the best parameter choice and the corresponding recall score?
- Is this better than the Logistic regression in the notes?