Practical 12

Jumping Rivers

Alternative Dimension Reduction

We will use the same data set as the previous example so that we can compare the results

a) Create a pipline for fitting a PCR

b) Use a grid search cross validation to find the optimal number of components to use.

c) From your `GridSearchCV` object you can retrieve the best estimator as the `.best_estimator_` attribute. Use this to get a cross validation estimate of the mean squared error for this model.

d) How does this compare to the linear model from the previous practical

e) Why do you think this is?

f) Fit a PLS regression and compare

Diagnosing progression of diabetes

Here we will put together all of the elements we have seen so far to build a model for predicting the progression of diabetes in patients. The data is available as part of the sklearn package.

```
from sklearn.datasets import load_diabetes
import pandas as pd

diabetes = load_diabetes()
X = diabetes.data
y = diabetes.target
feature_names = diabetes.feature_names

full_df = pd.DataFrame(X, columns=feature_names)
full_df['target'] = y
```

a) Start by producing some exploratory graphics and summary statistics. As a hint for producing plots quickly, you can use the **seaborn** library `pairplot()` function.

b) Given what we have discussed I want you to tell me about the diabetes data. What interesting things are there? What are important predictors in disease progression. What is the best model you can come up with for predicting disease progression for a new patient.