CT5141 Lab Week 7

James McDermott

In the well-known Credit Approval dataset (part of the famous UCI machine learning dataset repository), the task is to classify loan applications as positive or negative (i.e. application should be granted, or should not) according to their features.

1. **Data**. Visit https://archive.ics.uci.edu/ml/datasets/Credit+Approval to get the dataset. Read it into your python session, and take a look. Check that all columns are the type you expect. Check for unexpected columns or rows. Drop any rows with missing data. Extract `X` and `y`.
2. **Feature engineering**. Many of the features are categorical: how can we encode them? Program this to produce a new dataset.
3. **Train-test split**. Make a 80-20 train-test split with a fixed random seed of 0 (so that we all have exactly the same split).
4. **Classification**. Try several classification algorithms, ideally in a loop. Print out both the training and test performance.
5. **Feature selection**. Try `SelectKBest` to choose a subset of features. Notice its effect on both training and test performance.
6. **Pipeline**. Put the one-hot encoder, the feature selection, and a classifier together in a pipeline.