Cross Validation Scoring

Note: cross\_val\_score uses the k-fold cross validation method.

In cross validation, k refers the number of groups the data is split into.

The general procedure is as follows:

1. Shuffle dataset randomly (although default settings are shuffle=false, so the data will be split in its current state).
2. Split the data into k number of groups (denoted by the ‘cv’ value selected e.g. cv=5)
3. For each unique group:

- Take the group as a hold out (test) data set

- Take the remaining groups as a training data set

- Fit the model on the training set and evaluate it on the test set

- Retain the evaluation score and discard the model

1. Summarise the skill of the model using the sample of model evaluation scores

Configuration of k value

An uninformed chosen value of k may result in a mis-representative idea of the skill of the model, such as a score with high variance (the score changes a lot based on the data used to fit the model), or high bias, (such as an overestimate of the skill of the model).

Three ways of selecting a k value:

Representative: the k value is chosen so each group of data is large enough to be statistically representative of the broader dataset.

k=10: The value of k is fixed to 10, a value of that has been found through experimentation to generally result in a model skill estimate with low bias a modest variance.

k=n: The value for k is fixed to n, where n is the size of the dataset to give each test entry an opportunity to be used in the hold out dataset. This approach is called leave-one-out cross-validation.

If a value of k is chosen that does not evenly split the dataset, then one group will contain the remainders. It is preferable to split the dataset into k groups with the same number of entries, such that the sample of model skill scores are all equivalent.