CHAPTER 24

More Supervised Machine Learning Techniques with Scikit-learn

This chapter will cover using Scikit-learn to implement machine learning models using techniques such as

- Feature engineering
- Resampling methods
- Model evaluation methods
- Pipelines for streamlining machine learning workflows
- Techniques for model tuning

Feature Engineering

Feature engineering is the process of systematically choosing the set of features in the dataset that are useful and relevant to the learning problem. It is often the case that irrelevant features negatively affect the performance of the model. This section will review some techniques implemented in Scikit-learn for selecting relevant features from a dataset. The techniques surveyed include

- Statistical tests to select the best *k* features using the **SelectKBest** module
- Recursive feature elimination (RFE) to recursively remove irrelevant features from the dataset

E. Bisong, *Building Machine Learning and Deep Learning Models on Google Cloud Platform*,
https://doi.org/10.1007/978-1-4842-4470-8_24

- Principal component analysis to select the components that account for the variation in the dataset
- Feature importances using ensembled or tree classifiers

Statistical Tests to Select the Best *k* Features Using the SelectKBest Module

The following list is a selection of statistical tests to use with **SelectKBest**. The choice depends if the dataset target variable is numerical or categorical:

- ANOVA F-value, **f_classif** (classification)
- Chi-squared stats of non-negative features, **chi2** (classification)
- F-value, **f_regression** (regression)
- Mutual information for a continuous target, **mutual_info_regression** (regression)

Let's see an example using chi-squared test to select the best variables.

```
# import packages
from sklearn import datasets
from sklearn.feature_selection import SelectKBest
from sklearn.feature_selection import chi2

# load dataset
data = datasets.load_iris()

# separate features and target
X = data.data
y = data.target

# display first 5 rows
X[0:5,:]

# feature engineering. Let's see the best 3 features by setting k = 3
kBest_chi = SelectKBest(score_func=chi2, k=3)
fit_test = kBest_chi.fit(X, y)
```