Machine Learning: Introduction to Scikit-learn

Generally a learning problem consists of a set number of sample data and then attempts to predict the properties of unknown data.

Learning problem categories:

* Supervised learning: data comes with predefined labels
  + Classification: samples belong to two or more classes and we learn from the labelled data how to predict unlabelled data
  + Regression: if the output consists of one or more continuous variables
* Unsupervised learning: algorithm is not provided with any pre-assigned labels

Machine learning is about learning some properties of a data set and then testing those properties against another data set. A common practice in machine learning is to evaluate an algorithm by splitting a data set into two parts. One part is the training set and the other part is the testing set, which the learning properties are tested upon.

Estimator objects

Fitting data: the main API implemented by scikit-learn is that of an estimator. An estimator is any object that learns from data; it may be a classification, regression or clustering algorithm or a transformer that extracts/filters useful features from raw data.  
All estimator objects expose a “fit” method that takes a dataset (usually a 2-d array)  
e.g. estimator.fit(data)

Supervised learning

Generally, supervised learning consists of learning the link between two datasets: the observed data X and an external variable y that we try to predict (“target” or “label”). Usually y is a 1D array of length n samples.  
All supervised estimators in scikit-learn implement a fit(X, y) method to fit the model and a predict(X) method that, given unlabelled observations X, returns the predicted labels y.

k-Nearest neighbours  
The simplest possible classifier is the nearest neighbour: given a new observation X\_test, find in the training set

Linear models

Support Vector Machines

Linear SVMs: Support Vector Machines belong to the discriminant model family: they try to find a combination of samples to build a plane maximising the margin between the two classes. Regularisation is set by the c parameter: a small value for c means the margin is calculated using many or all of the observations around the separating line (more regularisation); a large value for c means the margin is calculated on observations close to the separating line (less regularisation).

Curse of Dimensionality

Dimensions/features/attributes (independent input variables)

As you increase the number of features used to train your model, your prediction accuracy will increase. However, up to a certain point (threshold) your accuracy will begin to decrease as your features increase exponentially. This is because not all of the features contribute to the outcome, so the model starts to get confused thus resulting in a reduced accuracy score.

Classification

Multiclass classification: If you have several classes to predict, an option often used is to fit one-versus-all classifiers and then use a voting heuristic for the final decision.

Shrinkage and sparsity with logistic regression