MACHINE LEARNING USING SCIKIT LEARN

TOPICS

- Types of Machine Learning
- Key Algorithms in Machine Learning
- Practical Applications of Machine Learning
- Various frameworks/libraries popular for ML
- Why Scikit Learn?
- Code Walkthrough

MACHINE LEARNING

- ML is the study of computer algorithms that can improve automatically through experience and by the use of data.
- It is part of artificial intelligence (AI).
- ML algorithms build a model based on sample data, known as "training data", in order to make predictions or decisions without being explicitly programmed to do so.
- ML algorithms are used in a wide variety of applications, such as medicine, email filtering, speech recognition, and computer vision.

TYPES OF LEARNING

- Supervised learning
  - Supervised machine learning can take what it has learned in the past and apply that to new data using labelled examples to predict future patterns and events. It learns by explicit example.
- Unsupervised learning
  - Supervised learning tasks find patterns where we have a dataset of “right answers” to learn from. Unsupervised learning tasks find patterns where we don’t. This may be because the “right answers” are unobservable, or infeasible to obtain, or maybe for a given problem, there isn’t even a “right answer” per se.
  - Unsupervised learning is used against data without any historical labels.

ML ALGORITHMS

- Binary Classification
- Regression
- Anomaly Detection
- Clustering
- Matrix Factoring

BINARY CLASSIFICATION

- Supervised learning algorithm.
- Task involves classifying the elements of a set into two groups (true/false) on the basic of a classification rule.
- Typical scenarios;
  - Medical testing to determine if a patient has certain disease or not.
  - Quality control in industry, deciding whether a specification has been met.
  - In information retrieval, deciding whether a page should be in the result set of a search or not.

REGRESSION

- Supervised learning algorithm.
- A set of statistical processes for estimating the relationships between a dependent variable (outcome/label) and one or more independent variables (predictors/features).
- Regression analysis is widely used for prediction and forecasting.
- Example scenarios;
  - Weather forecasting
  - Sales performance analysis
  - Stock market prediction
  - House pricing

ANOMALY DETECTION

- The identification of unexpected or rare items, events or observations which raise suspicions in data.
- Typically the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in text.
- Anomalies (spiked/change points) also referred to as outliers, noise, deviations and exceptions.

CLUSTERING

- Unsupervised learning algorithm.
- An algorithm that look for patterns in data, such as groups of customers based on their behaviour.
- During training, data is grouped based on the features, and then during the prediction, the closest match is chosen.
- Example;
  - Sorting music files.
  - Predicting customer preferences.

MATRIX FACTORIZATION

- An algorithm to provide recommendation.
- This algorithm is tailored to problems where historical data is available and the problem to solve is predicting a selection from that data.
- Example: Netflix recommendations.

VARIOUS FRAMEWORKS/LIBRARIES POPULAR FOR MACHINE LEARNING

Arguably, TensorFlow, PyTorch, and scikit-learn are the most popular ML frameworks. Still, choosing which framework to use will depend on the work you're trying to perform. These frameworks are oriented towards mathematics and statistical modeling (machine learning) as opposed to neural network training (deep learning).

- TensorFlow and PyTorch are direct competitors because of their similarity. They both provide a rich set of linear algebra tools, and they can run regression analysis.
- Scikit-learn has been around a long time and would be most familiar to R programmers, but it comes with a big caveat: it is not built to run across a cluster.
- Spark ML is built for running on a cluster, since that is what Apache Spark is all about.

SCIKIT-LEARN

- Scikit-learn was initially developed by David Cournapeau as a Google summer of code project in 2007.
- Later Matthieu Brucher joined the project and started to use it as apart of his thesis work. In 2010 INRIA got involved and the first public release (v0.1 beta) was published in late January 2010.
- The project now has more than 30 active contributors and has had paid sponsorship from INRIA, Google, Tinyclues and the Python Software Foundation.

WHAT IS SCIKIT-LEARN?

- Scikit-learn provides a range of supervised and unsupervised learning algorithms via a consistent interface in Python.
- It is licensed under a permissive simplified BSD license and is distributed under many Linux distributions, encouraging academic and commercial use.
- The library is built upon the SciPy (Scientific Python) that must be installed before you can use scikit-learn. This stack that includes:
  - NumPy: Base n-dimensional array package
  - SciPy: Fundamental library for scientific computing
  - Matplotlib: Comprehensive 2D/3D plotting
  - IPython: Enhanced interactive console
  - Sympy: Symbolic mathematics
  - Pandas: Data structures and analysis

WHAT ARE THE FEATURES?

The library is focused on modeling data. It is not focused on loading, manipulating and summarizing data. For these features, refer to NumPy and Pandas.

Some popular groups of models provided by scikit-learn include:

- Clustering: for grouping unlabeled data such as KMeans.
- Cross Validation: for estimating the performance of supervised models on unseen data.
- Datasets: for test datasets and for generating datasets with specific properties for investigating model behavior.
- Dimensionality Reduction: for reducing the number of attributes in data for summarization, visualization and feature selection such as Principal component analysis.

CONT.

- Ensemble methods: for combining the predictions of multiple supervised models.
- Feature extraction: for defining attributes in image and text data.
- Feature selection: for identifying meaningful attributes from which to create supervised models.
- Parameter Tuning: for getting the most out of supervised models.
- Manifold Learning: For summarizing and depicting complex multi-dimensional data.
- Supervised Models: a vast array not limited to generalized linear models, discriminate analysis, naive bayes, lazy methods, neural networks, support vector machines and decision trees.

ACTIVITY

THE END