DWDM Tutorial on toolkits: Scikit, Weka

- Raghavendra
- Shikha

Contents

Classification (Scikit Learn & Weka)

Clustering (Scikit Learn & Weka)

Classification

Classification is a problem of identifying a class to which a new observation belongs, on the basis of a training set of data containing observations.

For example,

- Classify emails into spam / non-spam
- Assigning a diagnosis to a given patient as described by observed characteristics of the patient.

Dataset

Data description - class label

Datatypes can be numerical, ordinal, categorical.

Iris Flower Dataset (3 class classification)

Sepallength	sepalwidth	petallength	petalwidth	class
5.0	3.3	1.4	0.2	Iris-setosa
7.0	3.2	4.7	1.4	Iris-versicolor
8.2	6.5	3.2	7.3.	Iris-virginica

How to handle textual Data ?

Usually categorical values are intentionally encoded into numerical values in order to be used as features. Textual data can be converted into numerical data using below techniques:

1. One-hot encoding
2. Word2Vec

For example, with the vocabulary {UK, US, Germany, India}

One hot encoding: [1,0,0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1]

Word2Vec: [0.23,0.12], [0.45,0.2], [1.43,0.23], [0.32, 0.76]

Data Pre-processing

- Data Cleaning
  - Binning: Smooth a sorted data values by consulting its neighbourhood.
  - Regression: find best line to fit in two attributes.
  - Outliers: Are detected by clustering
- Data Transformation
  - Discretization : The numerical values (eg: age) are replaced by interval range.
  - Normalization (z-score) : Values are scaled to small range from 0 to 1.
  - Generalization : generalizing to higher-level concepts in a concept hierarchy. (street -> city/country)

Classifiers

- Decision Trees
- SVM
- Logistic Regression
- Random Forests
- and many more ....

Scikit Learn

Machine learning library in python

Steps in scikit for classification:

1. Load dataset and preprocessing
2. Feature engineering: DictVectorizer, TfIdfTransformer, Countvectorizer
3. Choose classifier: Decision Tree, Naive Bayes
4. Train classifier (classifier.fit())
5. Test classifier (classifier.predict(X))

Demo.

Take enough time to go through all the contents of below link:

http://scikit-learn.org/stable/user_guide.html

Weka

Data analysis tool written in java.

Steps in Weka

1. Load data
2. Preprocess
3. Feature selection
4. Choose classifier
5. Train the classifier

Demo

Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters).

Clustering methods can be classified into the following categories −

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

Clustering algorithms

Almost all the algorithms are available online.

For example in scikit-learn we have :

- K- Means
- Birch
- DBSCAN
- Hierarchical clustering
- And so on.

Let’s look at the example of using the K-means algorithm from scikit library.

Scikit and Weka Demo

K-Means algorithm

Scikit demo

Weka demo

Useful link : https://www.youtube.com/watch?v=m7kpIBGEdkI