Data Mining Tutorial using Python

IS 733: Data Mining

Sreenivasan Ramasamy Ramamurthy – rsreeni1@umbc.edu
PhD Student (Information Systems) and Grader (IS733)

Data Mining

- Identify a problem statement
- Data preparation and preprocessing
- Visualization
- Model Building
- Evaluation

Identify a problem statement

- Determine your hypothesis
- Identify the features and response variable
- Identify the type of analysis

Data preparation and preprocessing

- Missing data
- Filtering – median, low, high, band-pass
- Windowing with/without overlap
- Data formatting – test/train split

Visualization

- Important to understand your data – relationship between features, data distribution, class distribution, imbalanced dataset
- Scatter plots, pair plots
- Dimensionality reduction – PCA, t-SNE

Model Building

- Classification or regression?
- How many features? How many number of data instances?
- Computational complexity
- Shallow learning vs Deep learning

Evaluation

- Comparison of predictions or classifications with ground truth.
- Accuracy, Recall, Precision, F1, Informedness, Markedness, ROC, AUC, confusion matrix, and many more.

Tools

- WEKA
- Python, Pandas, Scikit-learn
- R
- RapidMiner
- MATLAB

Steps to *succeed*

- Do not just feed the data to your classification model.
- Understand and visualize the data.
  - Visible pattern?
  - Imbalanced dataset?
- Find important features.
  - PCA, p-value, correlation coefficient
- Proper class definition
- Validation
  - Precision, Recall, F1-Score, ROC curve, AUC

Common features for different data

- Images
  - Color features such as color histograms which could for instance be in RGB or HSV space
  - histogram approaches, e.g. histogram of oriented gradients (HOG)
  - Texture features such as Tamura's or Haralick's
  - SIFT and SURF features are popular as well
- Text
  - Count Vectors as features
  - TF-IDF Vectors as features – e.g. Word level, N-Gram level, Character level
  - Word Embeddings as features
  - Text / NLP based features
  - Topic Models as features
- Time Series
  - Time, frequency domain statistics, PCA, Moving average, cross & auto correlation

Codes in Python Link

- https://colab.research.google.com/drive/1-L71FKdY30lfJPJIIFO-yvxM8gsC5a75?usp=sharing.