Lab 3 – Feature Selection and Classification using Python

3.1 Feature Selection Python

Feature Selection

- Removes distracting or noisy features to improve classification accuracy and model build time
- Filter Method
  - information gain
  - chi-square test
  - fisher score
  - correlation coefficient
  - variance threshold
  - …
- Wrapper methods
  - recursive feature
  - elimination sequential feature selection algorithms
  - genetic algorithms
  - …
- Embedded methods
  - recursive feature
  - L1 (LASSO) regularization
  - decision tree
  - …

Feature Selection

- Sklearn -> feature_selection :

https://scikit-learn.org/stable/modules/feature_selection.html#l1-based-feature-selection

3.2 Classification using Python

Data Frames

- Way to store data in rectangular grids that can be easily overviewed.
- Each row corresponds to values or an instance while each column contains data for a specific variable.
- Pandas
  - Popular python package for Data Science.
  - Offers powerful, expressive and flexible data structures that make data manipulation and analysis easy.
  - Pandas DataFrame is one such structure.

Pandas DataFrame

- Three main components:
  - Data
  - Index
  - Columns
- Can specify the index and column names.
- Index indicates the rows and column names indicate difference in columns.
- Can select an index using either .loc or .iloc
  - df.iloc[row][column] will use the position. .iloc[2] will look for values for dataframe that are at index 2.
  - df.loc[row][column] will use the label. .loc[2] will look for values for dataframe that have an index labelled 2.

NumPy Arrays

- Provides an efficient storage and better way handling of data for mathematical operations.
- Creates homogeneous n-dimensional arrays. All elements of NumPy array should be of same type.
- Advantages:
  - Dimensions can be changes at runtime if multiplicity factor produces the same number of elements. Example, 2*5 matrix can be converted to 5*2 and 1*4 into 2*2 by using .reshape() function.
  - Can create single dimensional array from any multi-dimensional array using .ravel() function.
  - Can perform mathematical operation on array like addition, subtraction, multiplication and division. np.array([1,2,3])*2.
  - Can also multiply two numpy arrays. np_array1*np_array2.
  - Some inbuilt functions like sum(), min(), max() amongst others.
  - Find shape of numpy array using nparray.shape()

NumPy Arrays

- Indexing
  - a[2:3] retrieves 3rd row and 4th column as the indexing starts at 0.
  - a[2,:] returns all columns of the 3rd row.
  - a[:,2] returns column 3 for all rows.

Scikit-learn

- Easy and clean Machine Learning library.
- Provides wide selection of supervised and unsupervised learning algorithms.
- Built on top of several common data and math Python libraries. Hence, can pass numpy arrays and pandas data frames directly to Machine Learning algorithms of scikit.
- Some of the libraries:
  - NumPy: Matrices and math operations
  - SciPy: Scientific and technical computing
  - Matplotlib: Data visualisation
  - Pandas: Data handling, manipulation, and analysis.
- Focuses on Machine learning and data modelling. Not concerned with loading, handling, manipulation and visualising data.

Scikit-learn

- Some robust algorithms include:
  - Regression: Fitting linear and non-linear models
  - Clustering: Unsupervised classification
  - Decision Trees: Tree induction and pruning for both classification and regression tasks
  - Neural Networks
  - SVMs
  - Naïve Bayes
  - Ensemble methods
  - Feature manipulation
  - Outlier detection
  - Model selection and validation

Tasks today

- Access Python either through VM or on your own laptop.
- Use IDE (PyCharm or Jupyter) for running the scripts.
- Study the code provided and configure the learning algorithms.
- Optional
  - Identify set of parameters for each learner to tune
  - Create a table or graph that shows relationship between classification accuracy with different values of parameter that is tuned.
  - For each learner, identify two parameters the accuracy is sensitive to.
  - Identify a range of values and examine the accuracy of each learner for each value of the parameter.

References

- https://www.datacamp.com/community/tutorials/pandas-tutorial-dataframe-python
- https://www.tutorialspoint.com/python_pandas/python_pandas_dataframe.htm
- https://pandas.pydata.org/pandas-docs/stable/getting_started/dsintro.html
- https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html
- https://www.tutorialspoint.com/numpy/numpy_array_attributes.htm
- https://towardsdatascience.com/a-hitchhiker-guide-to-python-numpy-arrays-9358de570121
- https://docs.scipy.org/doc/numpy/reference/generated/numpy.array.html
- https://towardsdatascience.com/an-introduction-to-scikit-learn-the-gold-standard-of-python-machine-learning-e2b9238a98ab
- https://scikit-learn.org/stable/

Thank you and Have fun!