Web Crawling and Data Pre-processing

Abdullah Al Foysal 17/04/2020

Machine Learning Steps:

- Data Collection
- Data Preparation / Pre-processing
- Choose model
- Training model
- Testing / Evaluating model
- Parameter Tuning
- Prediction
- Documentation

Data Collection:

In this step we need to gather data. This step is very important because the quality and quantity of data that we gather will directly determine how good our predictive model can be. We can collect data from Kaggle, UCI etc. One of the problem is, we cannot always find a dataset on our topic from Kaggle and UCI. If the data we are looking for is on web pages, however then the solution to all these problems is web crawling. We need to use web crawling to collect or store data from website.

Web Crawling for text data: We can collect and save different information from web pages using web crawler. We can use Python BeautifulSoup and Requests library as well as different frameworks like Scrapy to make a web crawler.Steps to make a web crawler are given below:

1. Select the website that we need to use.
2. Use requests library to access the web page. We request a content of a page from the server.
3. Identifying the URL structure of the website. It is very important when we need to collect data from multiple pages from the selected website. Normally we change the URL parameter and use a loop to control multiple pages.
4. Understand the HTML structure of a single page.

5. Parse the HTML content.A parser is a program that converts a string into a syntax tree.

6. Find all the necessary container like div , section , article , img , p etc. from HTML content that we need and use them to extract information.

7. Use pandas DataFrame to visually show and save the information in a file.
   Github: https://gitlab.com/genie-enterprise/trainees/abudllah
   File: web_crawling.py

Web Crawling for image data: Normally we download images from websites when we mention web crawling for image data. Steps of web crawling to download images are given below:

1. First five steps are similar like web crawling for text data.

2. Find all the img tag from the HTML content.

3. Get src attributes from img tag.

4. Use urlib.request (urllib.request.urlretrieve()) module to download image. We use image src attributes in urlib.request module to download images.
   Github: https://gitlab.com/genie-enterprise/trainees/abudllah
   File: web_crawler_image_data_single_page.py
   File: web_crawler_image_data_multiple_pages.py

So, we can say the basic steps of web crawling for both text and image data are [3]:

1. Document Load.

2. Parsing.

3. Extraction

4. Transformation

Data pre-processing:

Data in the real world is mostly not ready to use. It has inconsistency, noise and missing values. The reason is that it is coming from large set of different sources of data[1][2].

Data pre-processing techniques:

1. **Import libraries**

2. **Read data**

3. **Data quality assessment :**

   - Incomplete data: occupation = ‘’ ‘’

- Noisy data: salary = ‘’-1”
- Inconsistent data: was rating ‘’1,2,3” now rating ‘”A, B, C”
- Duplicate data
- Checking categorical values

4. **Data Normalization:** Data normalization is the process of rescaling one or more attributes to the range of 0 to 1. That means the largest value for each attribute is 1 and the smallest value is 0. The goal of normalization is to change the values of numeric columns in the dataset to a common scale [4]. But for machine learning every dataset does not require normalization. There are different types of normalization:

    - Min-Max Normalization
    - Z Normalization (Standardization)

5. **Dimensionality Reduction:** Dimensionality reduction is a process of reducing the number of variables/features in review. Dimensionality reduction can be divided into two subcategories called Feature Selection and Feature Extraction. Dimensionality Reduction is an important part of data pre-processing where the number of attributes go on increasing. This is a problem due to curse of dimensionality where the model gets tougher and tougher to train as the dimensions of the attributes increases. Dimensionality Reduction can also help for better visualization of data because it is difficult to visualize data in higher dimension. One of the popular techniques of dimensionality reduction is PCA (Principle Component Analysis) transformation. Apart from PCA other techniques are Linear Discriminant Analysis (LDA), and non-negative matrix factorization (NMF).

    - **Feature Selection:** It means selecting relevant features, discarding irrelevant features from dataset. Suppose we need to select feature for predicting mileage of a car. We have Engine capacity, top speed and colour. We will not select colour because it will not help to predict mileage of a car. We can use correlation matrix to see how the different features are related to each other or target variable in our dataset. Correlation can be positive (increase in one value of feature increases the value of the target variable) or negative (increase in one value of feature decreases the value of the target variable). Feature selection is very important because it reduces overfitting, improves accuracy and reduces training time. Some feature selection techniques [5][6]:

        (a) Correlation Matrix with Heatmap

        (b) Univariate Selection: Statistical tests can be used to select those features that have the strongest relationship with the output variable.

        (c) Feature Importance: Feature importance gives you a score for each feature of your data, the higher the score more important or relevant is the feature towards your output variable

- **Feature Extraction:** Feature extraction is the method which converts (M) attributes of data in to (N) attributes. The original number of attributes may be greater (M>N), less (M<N) or equal (M=N) to new attributes. In simple word it means you build a new set of features from the original feature set. Example: Suppose we have a dataset where there is column called country and it has string values. Our algorithm would never understand the string values. So, we can add country names as attributes and use a Boolean value to represent countries. This is a similar method to count vectorizer and one hot encoding. Some feature extraction methods are:

  (a) One-Hot Encoding

  (b) Count Vectorizer

  (c) TFIDF Vectorizer

6. **Train/Validation/Test Split:** In this step we split our dataset into two or sometimes three parts.

   - **Training data set:** Using this data our machine learning algorithms are actually trained to build a model. The model tries to learn the dataset and its various characteristics and intricacies.
   - **Validation data set:** Validation data set actually can be regarded as a part of training set. This set is used to help detect over-fitting and to assist in hyperparameter search
   - **Test data set:** Test data set is used to measure the performance of the model.

References:

[1] https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825

[2] https://towardsdatascience.com/data-pre-processing-techniques-you-should-know- 8954662716d6

[3] https://www.dataquest.io/blog/web-scraping-beautifulsoup/

[4] https://towardsdatascience.com/understand-data-normalization-in-machine-learning- 8ff3062101f0

[5] https://towardsdatascience.com/feature-selection-techniques-in-machinelearning- with-python-f24e7da3f36e

[6] https://machinelearningmastery.com/an-introduction-to-feature-selection/