Machine Learning Nanodegree

Capstone Proposal

Mohamed Elhwary

November 20, 2018

1- Domain Background

The supervised learning is branch of machine learning where we have labeled data (training data), fitting it into an algorithm and gleaning information from it (the function approximation) to learn the mapping function from the input and output. Then it can be able to predict the class label of the new coming data. Such as chatbots and facial recognition.

Natural language processing is the field that focus on making the computer able to understand the natural language of the human so it comes closer to a human-level. Although the computer can not understand well the hide-meaning of words or statements but it can do well job enough to help us.

I uploaded paper " A study of machine learning classifiers for spam detection.pdf "

2- problem statement

In this project I will use supervised learning to learn a classifier to classify the coming messages (e-mails) to spam or not spam. As now days the e-mails and SMS messages ar so common and people receive them every day. I want to prevent the spam messages to be received to users. I have a dataset with 5572 labeled messages.

3- datasets and inputs

I will use dataset from Kaggle => **https://www.kaggle.com/benvozza/spam-classification/data**
Dataset contains 5572 rows with 2 columns. First one is the class label (ham/spam) and the second is the contents of the message. 4825(86.5%) of messages are ham and and 747(13.5% are spam.

4- solution statement

It is classification problem and to solve it I will builed a supervised model. It will contain 4 algorithms Naive Bayes, Logistic Regression, SVM, and DecisionTree. The goal is to predict if a new message spam or not spam.

5- benchmark model

i will use benchmark model to compare the final result to make sure it works. I will use Naive bias classifier , as it is easy and has an initial result without tuning or using complex methods.

6- Evaluation Metrics

nltk.classify.accuracy(model, testing) is the first metric to evaluate. Second one is F-measure to find the precision and recall.

7- project design

The workflow of the project will be like that:
1- Loading the data and exploring it.
2- Regular Expressions to replace url, e-mail address, and phone numbers.
3- Pre-processing:
- Sentence Segmentation.
- Word Tokenization.
- Text Lemmatization.
- Identifying Stop Words.

4- Feature engineering.
5- Building the models.
6- Evaluation.

Resources:

1. https://medium.com/@ageitgey/natural-language-processing-is-fun-9a0bff37854e
2.https://www.researchgate.net/publication/310498804_A_study_of_machine_learning_classifiers_for_spam_detection
3. https://www.nltk.org/book/ch06.html