Jessica Braganza
BE Comp 2
F17112151

PAGE NO :
DATE :

Machine Learning

Assignment 1

1) Explain how ML works for the following common supervised learning applications:

I. Spam detection

i) In modern email system, you must have encountered a spam filter.

ii) This spam filter is a supervised learning system.

iii) Several spam filtering methods are used these days by email clients and other applications.

iv) ML algorithms are programmed well to ensure the security and updation of spam filters continuously.

v) Many email examples and their labels (spam/not spam) are fed into these systems, which in turn find out how to pre-emptively filtrate malicious emails in order that their user isn't troubled by them.

vi) Several of those additionally behave in such a way that a user can give new labels to the system and it can learn user preference

vii) The false spammers can simply be detected by observing specific patterns and by rule-based spam filtering.

viii) Examples of some spam filtering techniques are Perceptron and C4.5 Decision Tree.

II Natural Language Processing

i) To understand meaning of text documents machine learning algorithms are made for natural language processing.

ii) These documents can be any text : social media comments, online reviews and survey responses, even financial, medical, legal and regulatory documents.

iii) The intention behind this is to improve, accelerate and automate the text analytics functions and NLP that turn unstructured text into usable data and insights.

iv) Machine learning for NLP and text analytics includes a set of statistical methods for classifying parts of speech, entities, sentiment and other aspects of text.

v) The techniques are expressed as a model that is then applied to other text, also recognized as supervised machine learning.

vi) It can be a set of algorithms that can work on large sets of data to extract meaning, called as unsupervised machine learning.

vii) It's significant to understand the difference between supervised and unsupervised learning and to find the best of both in one system.

III Sentiment analysis

i) Sentiment analysis is the classification of subjective opinions or emotions (positive, negative or neutral) within text data using natural language processing.

ii) It helps gauge public opinion, conduct market research, monitor brand or product reputation, analyze social media sentiment and understand customer experiences.

iii) Using Sky.ai AI Platform for NLP you can quickly build and deploy a high quality custom Sentiment Analysis ML model.

iv) Good model choices include SVMs, Random Forests and Naive Bayes.

2) What are the different methods for managing missing values?

Ans. i) In the real world, most datasets consists of missing data i.e. incorrectly encoded data, or such type of data that is inappropriate for modelling.

ii) Sometimes such missing data is just that - missing.

iii) The actual value in given field is absent, for example: an empty string in a csv file, or sometimes it is encoded with a speical keyword or string.

iv) To work with this case of missing value is dependent on the nature of dataset.

v) Some common ways of dealing with missing values:

a) Casewise deletion of missing data: In this, the cases or row which consist of missing values are deleted or dropped permanently from the dataset. For which those have very few missing values in large dataset, this approach works well.

b) Replace missing values with mean/median value of the feature in which they occur. This is used in case of numerical. The mean/median is totally dependent on form of distribution of data. For uneven or asymmetrical data, the median may be more suitable, while for symmetrical or even

and more normally distributed data, the mean could be a better choice.

e) Replace with some constant value outside fixed value range.

In this method, missing values are grouped seperately in a category which is represented by a constant value. This option is usually preferred when other ways prove inappropriate to predict missing values. The disadvantage is that it may affect the performance of linear models. Global constant values are reused to fill missing values.

vi) Addition of new feature 'is null', indicates the rows with missing values.

vii) This feature helps the tree based models to understand that missing values are present.

viii) The disadvantage is that we twin the number of features.

ix) Run predictive models that assign the missing data.

x) This should be done in combination with some kind of cross-checking scheme in order to avoid loss.

xi) This may prove to be very effective with the final model.