UIC Web Search Engine

Context Pseudo-Relevance Feedback and Page Rank approach

Mirko Mantovani  
 Computer Science department  
 University of Illinois at Chicago  
 Chicago, Illinois  
 [mm](mailto:email@email.com)anto2@uic.edu

ABSTRACT

This document is a report for the final project of CS 412 Machine Learning at University of Illinois at Chicago.

It consists in a binary classification prediction problem, where, given the [Young People Survey dataset](https://www.kaggle.com/miroslavsabo/young-people-survey/) containing about 1000 examples (people) and 150 features about their personal tastes in music, movies, and personality, we have to train and build a model that is able to predict if a person is Very Empathetic (if Empathy is 4 or 5) or not very Empathetic (Empathy level is 1, 2 or 3), to help select suitable volunteers to help people with Alzheimer.

The repository containing the python notebook can be accessed through GitHub at:

[mirkomantovani/Machine-Learning-empathy-prediction](https://github.com/mirkomantovani/Machine-Learning-empathy-prediction)

1 DATA EXPLORATION - PREPROCESSING

1.1 Missing values

The first thing I noticed by looking at some statistics about the dataset, is that there were missing values in almost every feature. A very important observation was that there were also missing values in the Class (Empathy). I dropped the examples with missing Empathy because imputing them in any way would not have made any sense since it’s what we are predicting.

Many other features had from about 0.1% to 2% of missing values, randomly distributed in the dataset. I decided to impute all these values using the mode.

1.2 One Hot Encoding

Since many algorithms cannot deal with categorical values, and the dataset presented 11 categorical features, I decided to encode them, based on the type of features and their values, I divided these categorical features in three groups: ordinal, binary, OHE.

Ordinal attributes are the ones whose values can be ordered by using some kind of logic. For me these features were:

* Punctuality
* InternetUsage
* Education

Binary features are the ones whose value can be encoded to be either true or false, for me these features were:

* Gender
* LeftRightHanded
* OnlyChild
* VillageTown
* HouseFlats

I used OHE on the remaining categorical values because I thought that they could not really be ordinal, and some kind of information could have been lost in considering them as ordinal. These features were:

* Smoking
* Lying
* Alcohol

Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00