Predicting Personality Type from Online Forum Text with a Convolutional Neural Network

Low, Daniel Mark (S3120155)
Petre, Bogdan (S3480941)
Xu, Teng Andrea (S3548120)

Machine Learning

November 21, 2017

1 Application

We obtained the data from Kaggle (https://www.kaggle.com/datasnaek/mbti-type/data). This dataset contains over 8600 rows of data. On each row is a person's Personality Type (This persons 4 letter MBTI code/type) and set of each of the last 50 things they have posted. This is a classification problem: we will extract features from the text (e.g., TFIDF, word2vec embeddings, pronouns, sentiment analysis) and use them to predict 1 of 16 personality types.

2 Methods

Convolutional Neural Network (CNN). Optimization method will be mini-batch gradient descent which minimizes an objective function that is written as a sum of differentiable functions, and therefore tries to find the minima every iteration. Support Vector Machine (SVM) with a bag of words and TFIDF scheme. Optimization method is to achieve a hyperplane that has the largest distance to the nearest training-data point of any class.

3 Setup of Experiments

80-20 training-test split. We will use cross-validation to tune the hyperparameters of the models. We will train a CNN for text classification and an SVM and report results and report the test results.

We will also characterize the language features and plot word clouds of each personality type.

4 Programming Language

We will use Python.

5 Planning

Weeks 1-2, 13-26 Nov: Choosing the initial project. Finding dataset. Setting up the environment.

Week 3, 27 Nov - 3 Dec: Read literature on the topic of associating text features with personality and mental traits (e.g., [1]).

Week 4, 4 Dec - 10 Dec Extracting language features such as TFIDF, word2vec embeddings, pronouns, sentiment analysis to train the models.

Week 5, 11 Dec - 17 Dec: Design CNN.

Week 6, 18 Dec - 24 Dec: Tune CNN and ensemble classifier through cross validation.

Week 7, 8 Jan - 14 Jan: Train and test all models (CNN, ensemble and SVM). Characterize language and word clouds for each personality type.

Week 8, 15 Jan - 21 Jan: Write report.

References

[1] H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, 8(9):e73791, 2013.