CS 4361/5361 Machine Learning

Fall 2020

Practice Exam 2

1. (20 points) Word similarity. Write a function that receives a word w and an embedding dictionary E and returns the word in E whose embedding is closest to w's (according to Euclidean distance). If w is not in E, your function should return the indicator string '—'. Use the program *find_similar_words_start.py* as starting point.

2. (80 points) Sentence classification. The program *read_sentences_v2.py* generates an array X of size (12245,125,50) containing descriptions of the 12245 sentences found in the dataset and an array Y containing the class each sentence belongs to. Each sentence is described by a sequence of 30 embeddings of size 50. Your task for this exam is the following:

   (a) Compute a representation where each sentence is represented by the average embedding in the sentence (thus your dataset will be of size (12245,50)). Notice that if a particular row in a sentence representation contains all zeros, it means that it was added as padding, thus it should not be included in the average.

   (b) Split the data into training and testing set.

   (c) Compare the performance of multilayer perceptron, decision tree, and random forests to classify the dataset.

   (d) Compress the data to only 5 components using Principal Component Analysis

   (e) Compare the performance of multilayer perceptron, decision tree, and random forests to classify the reduced dataset.