RPTU
Department of Computer Science
Neural Networks for NLP
Mayank Nagda

Rheinland-Pfälzische Technische Universität Kaiserslautern Landau

Neural Networks for NLP - Sheet 1

30.10.2023

Deadline: 13.11.2023 - 13:15, on Olat as **pdf/.py/.ipynb** and/or after the lecture

---

Task 1: Linear- and Nonlinear Models — 6 Points

Assume we are given two lists - one containing 1,000 randomly selected Spanish words and another containing 1,000 randomly selected English words. Given a new word that is either English or Spanish we want to use those lists to predict its language. We will first vectorize all words, i.e. find a sensible mapping ϕ that maps any word to some vector in $\mathbb{R}^n$.

(a) Describe a vectorization mapping for words for this problem.

(b) When given a new word, w, we use vectorization to obtain $\phi(w)$ and compare it to the other 2,000 vectors in order to determine its language. For this purpose, should we use a linear model (such as an SVM), or should we use a non-linear approach (such as neural networks)? Explain.

(c) Suppose we use a neural network to obtain such a classifier. Describe the input and output of such a network as well as its training procedure using the given two lists.

Task 2: Conditional probabilities and language modelling — 6 Points

Consider the following scenario. We are given the sentence "Today, my friend and I will go to the ... and tomorrow we will visit the ... " and we are aiming to complete the sentence. Naturally, there are many ways to finish this sentence but to make it simple we will only consider the options "cinema", "zoo" and "bar" for both blanks and we assume that each option is equally likely and that both blanks are independent.

(a) What is the chance that they are going to the zoo twice?

(b) We are now given the information that they are going to the zoo at least once. What is the chance that they are going to the zoo twice?

(c) The two are now reconsidering their plans. Both of them dislike one of the options and they don't want to do the same thing twice. Since both are shy they don't want to say their preferences. In the end they agree on the following: One person chooses the activity for today and the other will **then** choose the activity for tomorrow. Would you rather choose first or second?

Task 3: Preprocessing — 12 Points

In the Materials folder you will find tutorial-1.ipynb.

(a) Remove stopwords and special tokens from the text (BoW). How did the accuracy change?

(b) Apply tf-idf. How did the accuracy change? (You have to apply this manually. Don't use packages)

(c) Find and report the minimal sufficient vocabulary size based on the accuracy.