Neural networks

Natural language processing - one-hot encoding

NATURAL LANGUAGE PROCESSING

Topics: one-hot encoding

- From its word ID, we get a basic representation of a word through the one-hot encoding of the ID
  - the one-hot vector of an ID is a vector filled with 0s, except for a 1 at the position associated with the ID
    - ex.: for vocabulary size $D=10$, the one-hot vector of word ID $w=4$ is

$$\mathrm{e}(w) = [\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ ]$$

  - a one-hot encoding makes no assumption about word similarity
    - $||\mathrm{e}(w) - \mathrm{e}(w')||^2 = 0$ if $w = w'$
    - $||\mathrm{e}(w) - \mathrm{e}(w')||^2 = 2$ if $w \neq w'$
    - all words are equally different from each other
  - this is a natural representation to start with, though a poor one

NATURAL LANGUAGE PROCESSING

Topics: one-hot encoding

- The major problem with the one-hot representation is that it is very high-dimensional
  - the dimensionality of $\mathrm{e}(w)$ is the size of the vocabulary
  - a typical vocabulary size is $\approx 100\ 000$
  - a window of 10 words would correspond to an input vector of at least $1\ 000\ 000$ units!
- This has 2 consequences:
  - vulnerability to overfitting
    - millions of inputs means millions of parameters to train in a regular neural network
  - computationally expensive
    - not all computations can be sparsified (ex.: reconstruction in autoencoder)