Similarity Learning for Text Classification

R Vamosi

Overview

1. Why similarity
2. My use case
3. Method
4. Current state

Why Similarity

- Similar objects are objects with common properties
- Vector v,w in a feature space. Similarity can be defined by
  - cos(angle(v,w)) = $v^T w/|v||w|$
- Vectors look similar:
  - principal components (from PCA method, principal component analysis)
  - number of dimensions.

What about words?

Towards Similarity: Words

Inter-relation (model) between words: summer ~

- heat, temperature, sun
- -winter, -cold

Towards Similarity: Words

Similarity measure is done by the sub-space projection method, e.g. word2vec

The output space, neural latent space, poses following characteristics:

- similarity
- some meaning in the vector space:
  e.g.: winter - rain + sun + heat ~ summer

Some “magic” of unsupervised learning

Towards Similarity: Words

What do we need?

- $\text{word}_k \rightarrow w_n=(k_1, k_2, \ldots, k_{dim})$ *(word vectors)*
- Some rule F:

  $F(w_n) = p_n$,a point in the output space

  $\text{word}_n, \text{word}_k$ similar $\longleftrightarrow$ $\text{metric}(p_n, p_k)$ small
- dimensionality dim low compared to cardinality of input set

Towards Similarity: Text

- n-gram = (token$_k$, token$_{k+1}$, …, token$_{k+n}$)
  = (this, is, a, meaningless, sentence, but, an, example, for, now, .)

Similarity within classes:

- documents
- books
- collections, journals

Similarity Learning: Text

Estimator for $F(w_n) = p_n$ with $\text{metric}(p_n, p_k)$

$w_k = (\text{word}_k, \text{word}_{k+1}, \ldots, \text{word}_{k+n})$

- common and uncommon words
- permutation of these words
- length
- “style” = pattern

Similarity Learning

Metric space with

- n-grams together within class: similar
- different classes apart

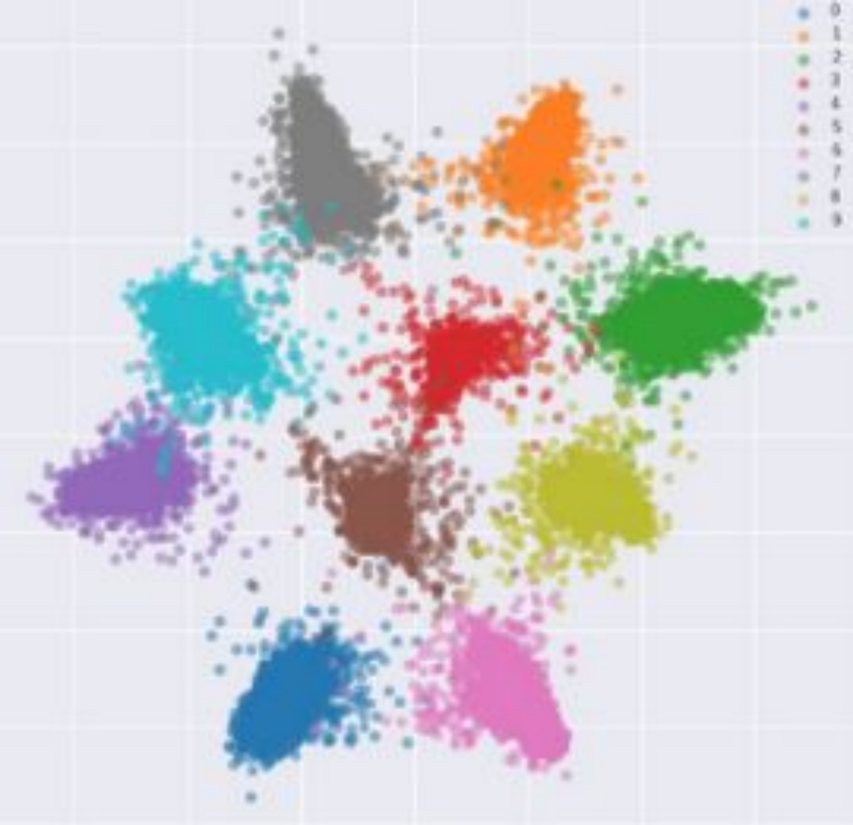

Fig.: 2D projection of the projection space (output)

Similarity Learning

Estimators:

- short-range vs long-range
- short-sequence vs long-sequence
- corpus -> dictionary -> encoding

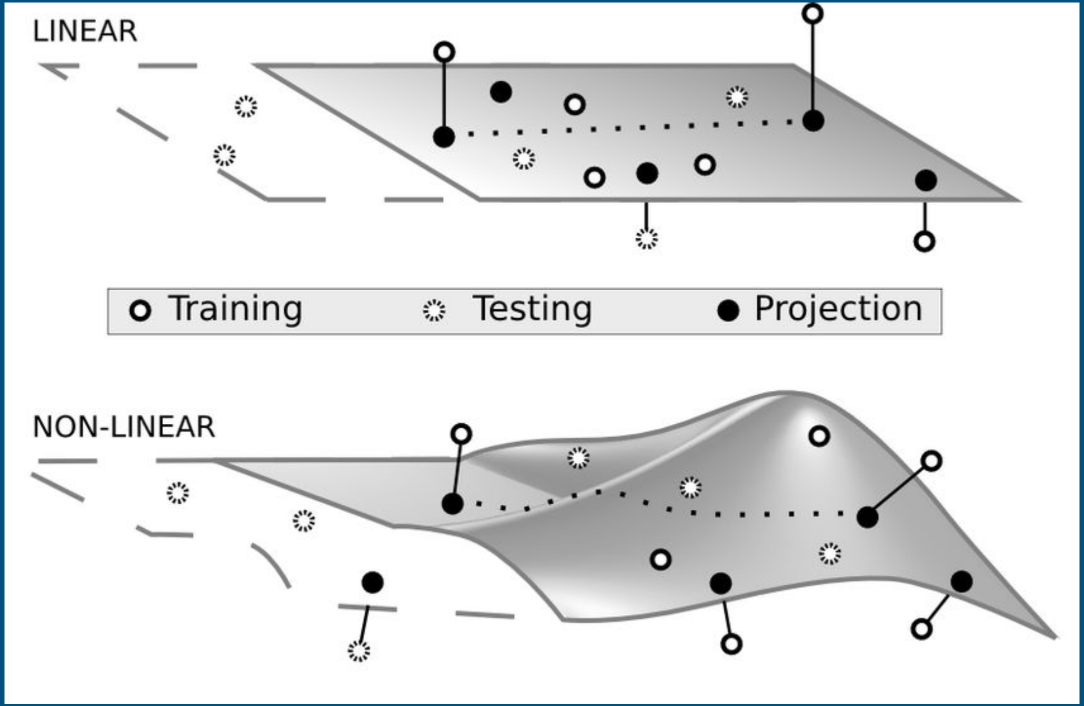

Fig.: linear and non-linear manifold in the output space

Similarity Learning: : Model

Estimator oftentimes
artificial neural network (ANN):

- Feedforward
- CNN
- LSTM

Learning by triplets (xa,x+,x-) only:

- xa and x+ are n-grams of the same class
- x- is a sentence of a different class

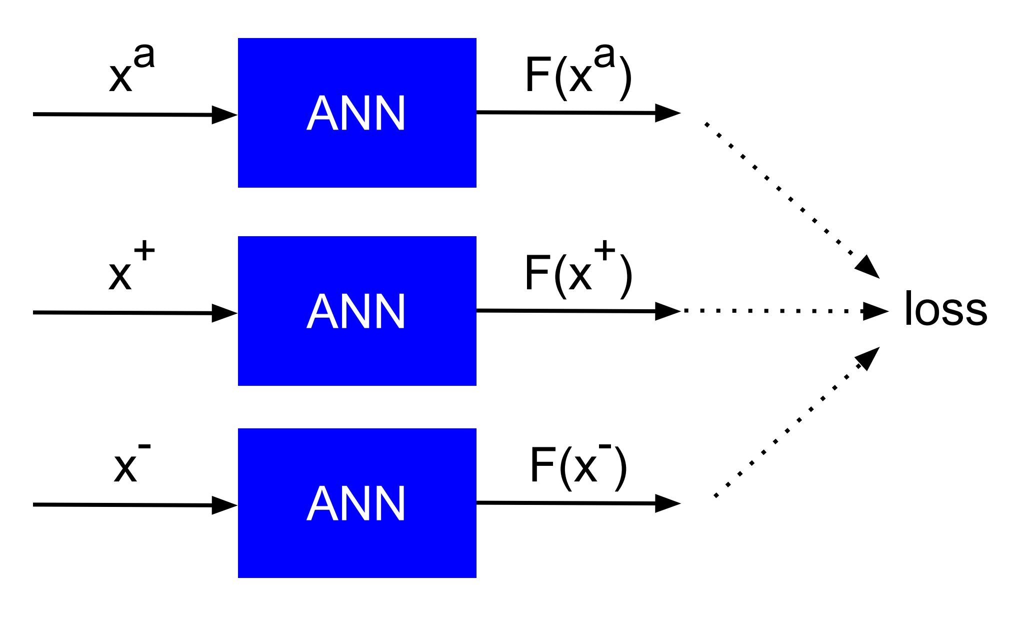

Fig.: ANN transforms input x into F(x)

Overview

short-range window -> prediction of the document class such as:

- economics, technical, scientific, gossips, art, beauty, ...
- publication / publisher / author
- ?

BACKUP

-