THEORY OF MACHINE LEARNING

LECTURE 23

REPRESENTATION LEARNING

ANNOUNCEMENTS

- **Homework 3 due on Monday April 11**
- **Project presentations:** starting in two weeks! (~18 projects)
  - Dates: April 19, 22, 26 (5 projects /class), couple online
- This week and next: representation learning, robustness

LAST CLASS

- Neural Tangent Kernels
  - Can one analyze dynamics of gradient descent? [can view as Kernel regression for a time-varying kernel]
  - [Jacot et al.] for infinitely wide nets, kernel remains "fixed" – neural tangent kernel; so NN learning == kernel regression
- NNs as Feature Learning or Representation Learning
  - NN transforms inputs -> "feature space embeddings", i.e., new representation
  - **Why?** Representations can have uses beyond classification (e.g., image captions, transfer learning, …)

REPRESENTATION LEARNING

- What makes a good representation?
  - Contrastive (for classification)
  - "Disentangled" or orthogonal
  - Sparse "explanations" for phenomena
  - Hierarchically organized, explanatory
  - Leverage domain knowledge
- Supervised vs Unsupervised
  - Pros: contrastive, better accuracy for *given task*, no special training
  - Cons: may not generalize
  - Pros: generalizes to many tasks, no careful data collection needed
  - Cons: unclear how to learn!

UNSUPERVISED REPRESENTATION LEARNING

- Classic approaches: manual feature engineering, autoencoders and "sparse coding"

- More modern: self-supervised learning, invariances and data augmentation
  - Example: NLP tasks

A Neural Probabilistic Language Model

Yoshua Bengio BENGIOY@IRO.UMONTREAL.CA
Réjean Ducharme DUCHARME@IRO.UMONTREAL.CA
Pascal Vincent VINCENTP@IRO.UMONTREAL.CA
Christian Jauvin JAUVINC@IRO.UMONTREAL.CA
Département d'Informatique et Recherche Opérationnelle
Centre de Recherche Mathématiques
Université de Montréal, Montréal, Québec, Canada

Editors: Jaz Kandola, Thomas Hofmann, Tomaso Poggio and John Shawe-Taylor

Abstract

A goal of statistical language modeling is to learn the joint probability function of sequences of words in a language. This is intrinsically difficult because of the **curse of dimensionality**: a word sequence on which the model will be tested is likely to be different from all the word sequences seen during training. Traditional but very successful approaches based on n-grams obtain generalization by concatenating very short overlapping sequences seen in the training set. We propose to fight the curse of dimensionality by **learning a distributed representation for words** which allows each training sentence to inform the model about an exponential number of semantically neighboring sentences. The model learns simultaneously (1) a distributed representation for each word along with (2) the probability function for word sequences, expressed in terms of these representations. Generalization is obtained because a sequence of words that has never been seen before gets high probability if it is made of words that are similar (in the sense of having a nearby representation) to words forming an already seen sentence. Training such large models (with millions of parameters) within a reasonable time is itself a significant challenge. We report on experiments using neural networks for the probability function, showing on two text corpora that the proposed approach

REPRESENTATION LEARNING IN NLP

- Approaches based on Firth's hypothesis

REPRESENTATION LEARNING IN GRAPHS