Contrastive estimation reveals topic posterior information to linear models

Daniel Hsu
*Columbia University*

Joint work with:
Akshay Krishnamurthy (*Microsoft Research*)
Christopher Tosh (*Columbia University*)
Learning useful representations of data

Probabilistic modeling

Deep learning

Image credit: stats.stackexchange.com; bdtechtalks.com
Goal of representation learning

Learned from data

Input Space  Feature Space

Image credit: towardsdatascience.com
Deep neural networks: Already doing it?

Image credit: [Bengio, Courville, Vincent, 2014]
Self-supervised learning

- **Idea**: Learn to solve *self-derived prediction problems*, then introspect.

- **Example: Images**
  - Predict color channel from grayscale channel
    [Zhang, Isola, Efros, 2017]

- **Example: Text documents**
  - Predict missing word in a sentence from context
    [Mikolov, Sutskever, Chen, Corrado, Dean, 2013; Dhillon, Foster, Ungar, 2011]

- **Example: Dynamical systems**
  - Predict future observations from past observations
    [Yule, 1927; Langford, Salakhutdinov, Zhang, 2009]
Example: Text documents

• Learn vector representations for words [or sentences] so that:
  • Co-occurring pairs of words [or sentences] -> corresponding vectors are close
  • Random pairs of words [or sentences] -> corresponding vectors are far

• Recent implementation: "QuickThoughts" [Logeswaran & Lee, 2018]
• Analysis: [Arora, Khandeparkar, Khodak, Plevrakis, Saunshi, 2019]

"The new mascot appears to have bushier eyebrows"

"The S&P 500 fell more than 3.3 percent"

"European markets recorded their worst session since 2016"
What's in the representation?

• **Q**: For what kinds of problems is the learned representation useful?
• [Arora et al, 2019]: Good for multi-class problems where ... 
  • "Co-occurring pairs" = examples from **same class**
  • "Random pairs" = examples from **independently chosen classes**
• **A different approach**: Perspective from probabilistic modeling

"The new mascot appears to have bushier eyebrows"
Our work

• We study a self-supervised representation learning method in the context of topic models

• **Theoretical finding**: Under topic modeling assumptions, learned representations *reveal document's topic posterior to linear models*

• **Experimental finding**: Representations improve over standard bag-of-words and competitive with word2vec in semi-supervised task. Better self-supervised learning -> better downstream supervised learning.
Outline

1. Representation learning method
2. Analysis in a topic model
3. Error transform theorem
4. Experimental study
1. Representation learning method

"Noise Contrastive Embedding"
Self-supervised contrastive learning problem

- Inspired by [Gutmann & Hyvärinen, 2010], but basic idea goes back at least to [Steinwart, Hush, Scovel, 2005; Abe, Zadrozny, Langford, 2006]
- Learn predictor to discriminate between

\[(x, y) \sim P_{X,Y} \text{ [positive examples]}\]

and

\[(x, y) \sim P_X \otimes P_Y \text{ [negative examples]}\]

- Specifically, estimate

\[f^*(x, y) = \Pr[\text{positive} \mid (x, y)]\]

by training a neural network (or whatever) using a loss function like squared loss or logistic loss on random positive & negative examples (evenly balanced: 0.5 \( P_{X,Y} + 0.5 \ P_X \otimes P_Y \)).
Contrastive learning problem with text documents

Positive examples:
Documents from a natural corpus

Negative examples:
First half of a document randomly paired with second half of another document

Can create training data from unlabeled documents!
Constructing the representation

• Given an estimate \( \hat{f} \) of \( f^* \), we construct embedding function for document halves:

\[
\hat{\phi}(x) := \left( \frac{\hat{f}(x, l_i)}{1 - \hat{f}(x, l_i)} : i = 1, \ldots, M \right) \in \mathbb{R}^M
\]

where \( l_1, \ldots, l_M \) are "landmark documents" (selected in a way TBD)
Prior analyses of contrastive learning

• Noise Contrastive Estimation [Gutmann & Hyvärinen, 2010]
  • Train $\hat{f}$ to discriminate samples from $P_\theta$ and samples from $P_{\text{noise}}$
  • Use $\hat{f}$ and data to define likelihood function that can be to (easily) estimate $\theta$

• Density Level Detection [Steinwart, Hush, Scovel, 2005]
  • Same $\hat{f}$ as above, but control ratio of real and noise samples
  • Use $\hat{f}(x)$ to detect if $x$ is in high-density region
  • [Abe, Zadrozny, Langford, 2006] give a "reduction"-style analysis

Q: What does $\hat{f}$ provide when data come from topic models?
2. Analysis in a topic model
Topic model

• [Hofmann, 1999; Blei, Ng, Jordan, 2003; ...]
• \( K \) topics, each specifies a distribution over the vocabulary
• A document is associated with its own distribution \( w \) over \( K \) topics
• Words in document [BoW]: i.i.d. from induced mixture distribution
  • Assume they are arbitrarily partitioned into two halves, \( x \) and \( y \)

E.g.,

\[
\sim \text{iid } \begin{align*}
\frac{1}{5} & \text{ sports} + \frac{2}{5} \text{ science} + \frac{2}{5} \text{ politics} + 0 \text{ business}
\end{align*}
\]
Simple case: One topic per document

• Suppose $w \in \{e_1, \ldots, e_K\}$ (i.e., document is about only one topic)

• **Recall:** $f^*(x, y) = \Pr[\text{positive} \mid (x, y)]$

• **Fact:** Odds ratio = density ratio:

$$g^*(x, y) := \frac{f^*(x, y)}{1 - f^*(x, y)} = \frac{\Pr[\text{positive} \mid (x, y)]}{\Pr[\text{negative} \mid (x, y)]} = \frac{P_{X,Y}(x, y)}{P_X(x)P_Y(y)}$$

Estimated using contrastive learning

Understand via data generating distribution
The density ratio

\[
\frac{P_{X,Y}(x, y)}{P_X(x)P_Y(y)} = \sum_{k=1}^{K} \frac{\Pr(e_k) \Pr(x \mid e_k) \Pr(y \mid e_k)}{P_X(x)P_Y(y)}
\]

Using BoW assumption

\[
= \sum_{k=1}^{K} \frac{\Pr(e_k \mid x) \Pr(y \mid e_k)}{P_Y(y)}
\]

Posterior over topics given \(x\)

\[
= \frac{\pi(x)^T \lambda(y)}{P_Y(y)}
\]

Likelihood of topics given \(y\)
Inside the embedding

• **Embedding:** \( \phi^*(x) = (g^*(x, l_i) : i = 1, \ldots, M) \) where

\[
g^*(x, y) = \frac{f^*(x, y)}{1 - f^*(x, y)} = \frac{\pi(x)^T \lambda(y)}{P_Y(y)}
\]

• Therefore

\[
\phi^*(x) = \begin{bmatrix}
P_Y(l_1)^{-1} \lambda(l_1)^T \\
\vdots \\
P_Y(l_M)^{-1} \lambda(l_M)^T
\end{bmatrix} \pi(x)
\]

(Scaled) likelihoods of topics given \( l_i \)'s

Posterior over topics given \( x \)
Upshot in the simple case

• In the "one topic per document" case, document embedding is a linear transformation of the posterior over topics
  \[ \phi^*(x) = L \pi(x) \]

• As long as \( L \) is full-rank, can express any linear function of topic posteriors using linear functions over \( \phi^*(\cdot) \)

• **Landmark selection**: Under distributional assumptions, letting landmark documents be a large i.i.d. sample from \( P_Y \) suffices.
General case: Exploit bag-of-word structure

- In general, posterior distribution over $w$ (topic distribution) given $x$ is not summarized by just a $K$-dimensional vector.

- If $x$ and $y$ each have $m$ words:
  - Let $p_m(t) := (t^d : |d| \leq m)$ where $t^d = \prod_{i \in [K]} t_i^{d_i}$ for $t \in \mathbb{R}^K$
  - Let $\pi(x) := \mathbb{E}[p_m(w) | x]$ (order $m$ multivariate conditional moments of $w$)
  - There is a corresponding $\lambda(\cdot)$ (that depends on topic model params) such that
    $$g^*(x, y) = \frac{\pi(x)^T \lambda(y)}{P_Y(y)}$$

- **Upshot:** By using landmark documents of varying lengths, $\phi^*(x)$ yields conditional moments of $w$ of varying orders.
3. Error transform theorem
Handling errors in representation learning

• We generally cannot solve the contrastive learning problem perfectly (due to, e.g., sampling error, approximation error).

• Are errors catastrophic when we use an imperfect representation in a downstream supervised learning task? [C.f., learning reductions, à la Beygelzimer, Langford, Zadrozny, 2008]
Setup

• **Ideal**: Our analysis concerned the representation $\phi^*$ derived from $f^*$ that optimally solves the contrastive learning problem (CLP):
  \[
  f^*(x, y) = \Pr[\text{positive} \mid (x, y)]
  \]

• **Reality**: We only solve the CLP approximately to get $\hat{f}$ with some MSE
  \[
  \mathbb{E}_{(X,Y) \sim D} \left[ \left( f^*(X, Y) - \hat{f}(X, Y) \right)^2 \right],
  \]
  where $D$ is data distribution in the CLP, and then form our embedding $\hat{\phi}$.

• **Ideal downstream supervised learning**: Want to learn linear functions of $\pi(x)$ (which are also linear functions of $\phi^*(x)$).

• **Reality**: Instead, we learn linear functions of $\hat{\phi}(x)$. 
Error transform theorem

**Theorem:** Assume $L \in \mathbb{R}^{M \times N}$ has rank minimum non-zero singular value $\Omega(\sqrt{M})$ with high probability (over i.i.d. sample of landmarks), and $f^*$ and $\hat{f}$ have range $[0, c]$ for some constant $c < 1$. Consider any linear function $\theta^* \in \mathbb{R}^N$ of $\pi(x) \in \mathbb{R}^N$. Then w.h.p.,

$$
\min_{\nu \in \mathbb{R}^M} \mathbb{E}_{X \sim P_X} \left[ (\pi(X)^T \theta^* - \hat{\phi}(X)^T \nu)^2 \right] \leq O \left( \|\theta^*\|_2^2 \left( \varepsilon + \frac{1}{\sqrt{M}} \right) \right)
$$

where

$$
\varepsilon = \mathbb{E}_{(X,Y) \sim D} \left[ \left( f^*(X,Y) - \hat{f}(X,Y) \right)^2 \right]
$$

Can also get similar result w.r.t. logistic loss
Main idea of proof

• Ultimately, we bound

$$\frac{1}{M} \mathbb{E}_{X \sim P_X} \left[ \| \phi^*(X) - \hat{\phi}(X) \|_2^2 \right] = \frac{1}{M} \sum_{j=1}^{M} \mathbb{E}_{X \sim P_X} \left[ \left( g^*(X, l_j) - \hat{g}(X, l_j) \right)^2 \right]$$

by some constant times

$$\mathbb{E}_{(X,Y) \sim P_X \otimes P_Y} \left[ \left( f^*(X,Y) - \hat{f}(X,Y) \right)^2 \right] + O \left( \frac{1}{\sqrt{M}} \right)$$

due to random sampling of landmarks $l_1, \ldots, l_M \sim P_Y$ (w.h.p.).

• This can be related to $\varepsilon = \mathbb{E}_{(X,Y) \sim D} \left[ \left( f^*(X,Y) - \hat{f}(X,Y) \right)^2 \right]$ because

$$D = \frac{1}{2} P_{X,Y} + \frac{1}{2} P_X \otimes P_Y$$
4. Experimental study
Practical variant of embedding

• Train $f_X$ and $f_Y$ to approximately minimize

$$
\frac{1}{2} \mathbb{E}_{(X,Y) \sim P_{X,Y}} [\log(1 + \exp(- f_X(X)^T f_Y(Y)))]
+ \frac{1}{2} \mathbb{E}_{(X,Y) \sim P_X \otimes P_Y} [\log(1 + \exp( f_X(X)^T f_Y(Y)))]
$$

so that optimal $f_X^*, f_Y^*$ satisfy

$$
g^*(x,y) = \exp(f_X^*(x)^T f_Y^*(y)) = \frac{\Pr[\text{positive} \mid (x,y)]}{\Pr[\text{negative} \mid (x,y)]}
$$

• This heuristically suggests that using $f_X^*$ as the embedding function will have similar behavior as $\phi^*$
Study dataset and comparisons

- **AG News** [Del Corso, Gulli, Romani, 2005; Zhang, Zhao, LeCun, 2015]:
  Four categories (world, sports, business, sci/tech) of news articles
  - 16,700 words in vocabulary after removing rare words; avg. ~45 words/document
  - Use 4 x 29,000 unlabeled examples for contrastive learning to get $\hat{\phi}$
  - Use (up to) 4 x 1,000 labeled examples to train linear classifier (multi-class logreg)
  - Use 4 x 1,900 labeled examples for test set

- **Our embedding $\hat{\phi}$ (called "NCE" for Noise Contrastive Embedding):**
  - Three-layer ReLU networks with ~300 nodes/layer
  - Dropout regularization, batch normalization, PyTorch initialization
  - Trained using RMSProp

- **Baseline embeddings $\hat{\phi}$:**
  - word2vec [Mikolov et al, 2013], Latent Dirichlet Allocation [Blei et al, 2003], BoW
Accuracy on supervised task vs # sample size

$\hat{\phi}(x) \in \mathbb{R}^N$ for $N = 100$
Varying the number of layers
Performance on contrastive task vs accuracy

![Graph showing performance on contrastive task vs accuracy.](image-url)
t-SNE embeddings of test documents

[Van der Maaten & Hinton, 2008]
Summary

**Broader theme:** Study self-supervised representation learning in the context of probabilistic models

- Under topic modeling assumptions, we prove that learned representations reveal document's topic posterior to linear models
- Error transform theorem and empirical study provides some reassurance that this is not brittle

**Future work:** Other models (e.g., HMMs), use in interactive settings
Acknowledgements

- Thanks to Miro Dudík for initial discussions & suggestions
- Thanks to the NSF for grant CCF-1740833

Thank you!
Synthetic topic model recovery experiments