Semisupervised Autoencoder for Sentiment Analysis

Shuangfei Zhai, Zhongfei Zhang.

이종진

Seoul National University

ga0408@snu.ac.kr

July 06, 2018

- Traditional autoencoders suffer from at least two aspects.
  - Scalability with the high dimensionality of vocabulary size.
  - Dealing with task-irrelevant words.
- Proposed are divised to learns highly discriminative feature maps.

- x: n-gram count data, y: label, $\tilde{x}$: reconstruction of x.
- Traditional autoencoder's loss function.

$$D(\tilde{x}, x) = (\tilde{x} - x)^2 \tag{1}$$

  - Reconstruction to be accurate towards frequent words.
- Proposed autoencoder's loss function.

$$D(\tilde{x}, x) = (\theta^T(\tilde{x} - x))^2 \tag{2}$$

  - θ are the weights of the linear classfier for label.
  - Reconstruction to be accurate towards only along directions where the linear classifier is sensitive to.

- $D(\tilde{x}, x) = (\theta^T(\tilde{x} - x))^2$ has rationalized from the perspective of Bregman Divergence
- SVM2

$$L(\theta) = \sum (\max(0, 1 - y_i\theta^T x_i))^2 + \lambda\|\theta\|^2 \tag{3}$$

- θ is fixed.

$$f(x_i) = (\max(0, 1 - y_i\theta^T x_i))^2 \tag{4}$$

- Reconstruct $\tilde{x}_i$ to have small value of $f(\tilde{x}_i) = f(x_i)$
  - we would like to $\tilde{x}_i$ to still be correctly classified by the pretrained linear classifier.
  - Bregman Divergence from $f(x_i)$ and use it as the loss function of the subsequent autoencoder training, the autoencoder should be guided to give rescontruction errors that do not confuse the classifer.

- Bregman Divergence with respect to f.

$$D_f(\tilde{x}, x) = f(\tilde{x}) - (f(x) + \Delta f(x)^T(\tilde{x} - x)). \tag{5}$$

- $f(x_i)$ is a quadratic function of x_i, The Hessian follows as

$$H(x_i) = \begin{cases} (\theta^T(\tilde{x}_i - x_i))^2 & \text{if } 1 - y_i\theta^T x_i > 0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

- Bregman Divergence is simply $(x - \tilde{x})^T H(x - \tilde{x})$ in SVM2

$$D_f(\tilde{x}, x) = \begin{cases} (\theta^T(\tilde{x}_i - x_i))^2 & \text{if } 1 - y_i\theta^T x_i > 0 \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

The Bayesian Marginallization

- Estimate θ using one single classfier can bring bias.
- Bayesian approach, Borrowing the idea of Energy Based Model

$$p(\theta) = \frac{exp(-\beta L(\theta))}{\int exp(-\beta L(\theta)), d\theta} \tag{8}$$

- Rewrite $D(\tilde{x}, x) = \int (\theta^T(\tilde{x} - x))^2 p(\theta) d\theta$, and using sampling method, MCMC.
- Approximate $p(\theta)$ by gaussian $\tilde{p}(\theta) = N(\hat{\theta}, \Sigma)$, then

$$D(\tilde{x}, x) = (\hat{\theta}^T(\tilde{x} - x))^2 + (\Sigma^{\frac{1}{2}}(\tilde{x} - x))^T(\Sigma^{\frac{1}{2}}(\tilde{x} - x)) \tag{9}$$

- $\Sigma = \frac{1}{\beta}(diag(\sum I(1 - y_i\theta^T x_i > 0)x_i^2))^{-1}$

Experiments

- Dataset (IMDB dataset / Amazon review data of five item.)
- Method
  - Bag of Words with uni-gram or bi-gram
  - Normalization:
  $$x_{i,j} = \frac{\log(1 + c_{i,j})}{\max_j \log(1 + c_{i,j})} \tag{10}$$
  - DAE/ DAE with Finetuning / NN / Logistic with Dropout / Semisupervised Bregman Divergence Autoencoder / SBDAE with Finetuning

Experiments

- Book
  - id1: lost credability,quickly!!:chalupa, id2 : 4423
  - asin : 055380121X
  - product name/product type
  - helpful: 12 of 15
  - rating: 2.0
  - title/data/reviewer/reviewer location
  - reviewer text I admit, I haven't finished this book. A friend recommended it to me as I have been having problems with insomnia. I was interested in reading a book about women's health issues and this one sounded intriguing UNTIL she started in with her tarot cards, interest in astrology and angels. Granted, I am not a firm believer in just "the hard facts" but its really hard to believe anything this woman writes after it is clear that common sense isn't alternative enough for her!

Experiments

Table 2: Left: our model achieves the best results on four (large ones) out of six datasets. Right: our model is able to take advantage of unlabeled data and gain better performance.

	books	**DVD**	**music**	**electronics**	**kitchenware**	**IMDB**	**IMDB + unlabled**
BoW	10.76	11.82	11.80	10.41	9.34	11.48	N/A
DAE	15.10	15.64	15.44	14.74	12.48	14.60	13.28
DAE+	11.40	12.09	11.80	11.53	9.23	11.48	11.47
NN	11.05	11.89	11.42	11.15	9.16	11.60	N/A
LrDrop	9.53	10.95	10.90	**9.81**	**8.69**	10.88	10.73
SBDAE	**9.16**	**10.90**	**10.59**	10.02	8.87	**10.52**	**10.42**
SBDAE+	**9.12**	**10.90**	**10.58**	10.01	8.83	**10.50**	**10.41**

Experiments

Table 3: Visualization of learned feature maps. From top to bottom: most activated and deactivated words for SBDAE; most activated and deactivated words for DAE.

nothing cannon outrageously lends teacher	disappointing worst unfortunately terrible predictable	badly disappointing annoying worst poorly	save redeeming awful sucks convince	even attempt unfunny couldn't worst	dull fails stupid worst avoid	excuse had failed rest he	ridiculously dean none ruined attempt
first classic man hard still	tears wonderfully helps awesome terrific	loved finest noir magnificent scared	amazing incredible funniest unforgettable captures	excellent surprisingly beauty unexpected appreciated	perfect ? powerful excellent favorite	years terrific peter cool allows	with best recommended perfect heart
long worst kids anyone trying	wasn't guy music work now	probably fan kind years place	to the and this shows	making give performances least comes	laugh find where before ever	tv might found kids having	someone yet goes away poor
done find before work watching	least book day actors classic	go trying looks everyone performances	kind takes special now someone	recommend instead wife shows night	although everyone anything comes away	ending once wasn't american sense	worth interesting isn't rather around