Sentiment Classification via Deep Learning

Anthony Gentile
Sentiment classification

• What is sentiment classification?
  - Determining opinion
  - Polarity
  - Emotions
  - “If they would've focused more on the plot, that would have been a great movie!” => ? [positive, negative, neutral]

• Amazon Reviews Data Set
Judi Dench is so great at this type of comedy. I would have loved to have seen her in something of Neil Simons. Her gift for repartee is impressive. As a long time watcher of "As Time goes By" I was delighted to find this series. Rather bittersweet that it was made with her husband, who she lost after a long and I gather successful marriage.
Previous Work

• Structural Correspondence Learning (SCL) (Blitzer et al 2007)

• Multi-label Consensus Training (MCT) (Li and Zong 2008)

• Spectral Feature Alignment (SFA) (Pan et al 2010)

• Stacked Denoising Autoencoders (SDA) (Glorot et al 2011)

(Baseline: linear SVM trained on raw data.)
Glorot et al (2011)

- Deep Learning
- Stacked Denoising Auto-encoders
- Results show improvement over previous techniques (SFA, MCT, and SCL)
Figure 1. Transfer losses on the Amazon benchmark of 4 domains: Kitchen (K), Electronics (E), DVDs (D) and Books (B). All methods are trained on the labeled set of one domain and evaluated on the test sets of the others. SDA_{sh} outperforms all others on 11 out of 12 cases.

Figure 2. Left: Transfer ratios on the Amazon benchmark. Both SDA-based systems outperforms the rest even if SDA_{sh} is better. Right: Proxy A-distances between domains of the Amazon benchmark for the 6 different pairs. Transforming data with SDA_{sh} increases the proxy A-distance.
Deep Learning

- Subset of Machine Learning, focused on Artificial Intelligence.
- Focused on learning representations rather than handcrafting features.
- Commonly used for tasks related to vision and speech.
- Being used more and more for NLP tasks.
- http://deeplearning.net
Deep Learning

- **Supervised learning** (labeled training data)
  - Most data out there is unlabeled
- **Unsupervised learning**
  - If we have a good model about some supervised data, we can do a good job of working in an unsupervised manner.
- **Multiple layers of abstraction**
  - Each layer is an intermediate representation
Deep Learning

[Lee et al. ICML 2009; Lee et al. NIPS 2009]
Stacked Denoising Autoencoders

- A way to build the representations for deep learning.
- Encoder function $h()$
  - $h(2 \text{ connected points}) = \text{line}$
- Decoder function $g()$
  - $g(\text{line}) = 2 \text{ connected dots}$
- Reconstruction $= g(h(x))$ Did we lose anything?
- Stacked simply means a chained group of encoders
- Denoising means we are taking corrupt input and trying to produce original/clean.
Deep Belief Networks

- Markov Random Fields with multiple layers
- Multiple layer neural networks
  - Stack of logistic regression classifiers.
  - Output of previous layer equals the input of the current layer.
  - More than two layers is considered Deep.
Neural Networks

- Similar to MaxEnt/Logistic Regression

A single neuron
A computational unit with $n$ (3) inputs and 1 output and parameters $W, b$

Bias unit corresponds to intercept term
Glorot Implementation

Preprocess the data

Feature-dictionary.txt

1 i 22956
2 you 8100
3 not 7464
4 was 7462
5 <num> 6183
6 my 5833
...

Data.vec
the preprocessed reviews written in a LibSVM fashion without the labels. Each line corresponds to a
review where each non-zero feature is indicated with <feature_index>:<feature_value>

1:1 2:1 9:1 10:1 14:6 16:2 17:3 18:1 ...
2:2 13:1 21:3 25:1 37:1 58:1 62:1 66:2 ...

Data.lab
Corresponds to .vec lines. 1 = positive, 0 negative

1
0
1
...
Glorot Implementation

Training

- Leverage Theano (Python Deep Learning Library)
- Load up hyperparameters

```python
# Absolute paths
# Rq: The sparse csr matrix should have the shape:
# (number of examples, number of dimensions)
# The data should be binary: 1 for presence of the word

state.PathData = 'path_to_pickled_csr_matrix_of_unsupervised'
state.PathDataSup = 'path_to_pickled_csr_matrix_of_supervised'
state.SavePath = 'path_to_save_model'

state.n_inp = 5000 # number of input to consider
state.n_hid = 5000 # number of hidden units
state.seed = 123456 # seed of the random number generator
state.batchsize = 10 # size of the mini-batch for SGD
state.zeros = 0.8 # binomial corruption noise (zero masking) probability for the DAE
state.ratio = 0.05 # ratio of sampling for reconstruction sampling

# MOST IMPORTANT HYPERPARAMETERS TO CROSS-VALIDATE ON

state.epochs = [1,3,7,15,30,50] # list of number of epochs for which we save the model
state.lr = 0.005 # learning rate
state.regcoef = 0.00001 # L1 penalty coefficient on the activation values
```
Glorot Implementation

- Train the model in a greedy layer-wise fashion
- The hidden layer activation function is the rectifier activation.
- The reconstruction activation function is the sigmoid.
- From one layer to the next we need to scale the parameters in order to ensure that the representation is in the interval [0,1]
Glorot Implementation

...finished training epoch #1
######## RESULTS:
Epoch: 1
Online Reconstruction: 139.599127023
Online Cost: 139.605202193
Online L1: 607.516991975

...finished training epoch #2
...finished training epoch #3
######## RESULTS:
Epoch: 3
Online Reconstruction: 123.509150732
Online Cost: 123.516266995
Online L1: 711.626315309

...finished training epoch #4
...finished training epoch #5
...finished training epoch #6
...finished training epoch #7
######## RESULTS:
Epoch: 7
Online Reconstruction: 113.744708683
Online Cost: 113.752799403
Online L1: 809.071987747
Glorot Implementation

- Problems Encountered
  - Memory intensive
  - CPU intensive
  - Theano errors
Questions & Answers

References:
- http://deeplearning.net/