SENTIMENT ANALYSIS
USING SEMI-SUPERVISED RECURSIVE AUTOENCODER

By: Vinay Kumar, CSE, IIT Kanpur
Guided by: Prof. Amitabha Mukerjee
Sentiment Analysis: What is it and why are people doing something about it?
Sentiment Analysis

Definition:
“It is the field of study that analyzes people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics and their attributes (from sources like text).”

-Pang, Bo and Lillian Lee
Sentiment Analysis (Applications)

- Stock Prediction
- Social Media Monitoring
- Twitter Sentiment
- Movie/product reviews
Methodology

Dataset
- Phrases and Sentences (movie reviews)

Recursive Autoencoder
- Creates a N-dimensional vector representation for each phrase

Multinomial regression
- Multiclass classifier

Sentiment
- (based on all probabilities)
Recursive Autoencoder

N-dim vector - every word

Min reconst. error

Vector rep for phrase
Multinomial Regression (Softmax):

Cost Function:

\[
J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{k} 1\{y^{(i)} = j\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}
\]

Notice that this generalizes the logistic regression cost function, which could also have been written:

\[
J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \left(1 - y^{(i)}\right) \log \left(1 - h_\theta(x^{(i)})\right) + y^{(i)} \log h_\theta(x^{(i)})
\]

\[
= -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=0}^{1} 1\{y^{(i)} = j\} \log p(y^{(i)} = j|x^{(i)}; \theta)
\]

\[
J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{k} 1\{y^{(i)} = j\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}} + \frac{\lambda}{2} \sum_{i=1}^{k} \sum_{j=0}^{n} \theta_{ij}^2
\]

:Weight decay

Class label probabilities:

\[
p(y^{(i)} = j|x^{(i)}; \theta) = \frac{e^{(\theta_j - \psi)^T x^{(i)}}}{\sum_{l=1}^{k} e^{(\theta_l - \psi)^T x^{(i)}}}
\]

\[
= \frac{e^{\theta_j^T x^{(i)}} e^{-\psi^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}} e^{-\psi^T x^{(i)}}}
\]

\[
= \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_l^T x^{(i)}}}.
\]
Dataset

Consists of 8544 sentences which is converted to 156060 English phrases from movie reviews.

Issues and Resolutions

• **Randomized N-dimensional vector**: Using *word2vec* to map similar words nearby to exploit the similarity of meaning/context.

• **Belong to more than one class -&gt; Unnecessary errors**: Using **Softmax regression** instead of logistic regression.

• **Altering the values of random parameters as N, α, the parameter that controls the relative weighting between the reconstruction error, error of the unsupervised part, and the cross-entropy error, error of the supervised learning part.**
Source code for Recursive Auto-encoder:

http://www.socher.org/index.php/Main/Semi-SupervisedRecursiveAutoencodersForPredictingSentimentDistributions

Thanks to Richard Socher (Phd fellow at Stanford University) and Stanford CoreNLP for source code of Recursive Autoencoder
References:


Questions?