Patrick Pegus
Mini Project 1
October 5, 2015
CMPSCI-689
Prof. Sridhar Mahadevan

Theory Question 1

To be an inner product, a function must satisfy the following properties:

$$\begin{aligned}
\textbf{Non-negativity:}\ & f(\vec{x},\vec{x}) \geq 0 \text{ if } 0,\ \vec{x} \text{ must be } \vec{0} \\
\textbf{Linearity:}\ & f(\vec{x}+\vec{y},\vec{z}) = f(\vec{x},\vec{z}) + f(\vec{y},\vec{z}) \\
\textbf{Scalar multiple:}\ & f(\alpha\vec{x},\vec{y}) = \alpha \cdot f(\vec{x},\vec{y}) \\
\textbf{Symmetry:}\ & f(\vec{x},\vec{y}) = f(\vec{y},\vec{x})
\end{aligned}$$

The cosine distance δ does not satisfy the scalar multiple property. For example, let $\alpha = 2$ and $\omega_i^T = [1, 0]$. Then $\delta(\alpha\omega_i, \omega_i) = 1$ and $\alpha \cdot \delta(\omega_i, \omega_i) = 2$.

Theory Question 2

a	b	x	y	$Dim.$	$COSADD_y$	$COSMULT_y$
London	England	Baghdad	Mosul	50	0.648	0.845
London	England	Baghdad	Iraq	50	0.727	0.921
London	England	Baghdad	Mosul	200	0.506	0.766
London	England	Baghdad	Iraq	200	0.574	0.835

Table 1: Given the analogy a is to b as x is to y, let $COSADD_y = \delta(\omega_y, \omega_x - \omega_a + \omega_b)$ and $COSSMULT_y = \frac{\delta(\omega_y,\omega_b)\delta(\omega_y,\omega_x)}{\delta(\omega_y,\omega_a)+\epsilon}$.

As shown in Table 1, both distance measures yield a greater value for the correct answer, Iraq.

Programming Question 3

Dist. Meas.	Dim.	Sem.	Syn.	Tot.
COSADD	50	49.9	44.7	47.0
COSMULT	50	36.5	27.9	31.7
COSADD	100	59.2	61.6	61.3
COSMULT	100	55.6	41.4	47.8
COSADD	200	75.5	65.7	70.1
COSMULT	200	56.9	40.0	47.6

Table 2: Comparison of distance measure accuracies on Google word analogies. The various dimension word embeddings were created by GloVe from the Gigaword5 + Wikipedia2014 6B token corpus.

As shown in Table 2, COSADD significantly outperformed COSMULT in both semantic and syntactic analogy tasks, which is consistent with [1]. Also consistent with [1], accuracy generally increases at a decreasing rate with increased word embedding dimensionality. Finally, accuracy is higher on semantic tasks with both distance measures. This may be caused by GloVe's use of the sum of their model's two ouptut word embedding sets or a larger and more symmetric context window.[1]

Programming Question 4

Dist. Meas.	Dim.	Tot.
COSADD	50	39.2
COSMULT	50	21.2
COSADD	100	55.7
COSMULT	100	38.5
COSADD	200	64.0
COSMULT	200	41.4

Table 3: Comparison of distance measure accuracies on syntactic MSR word analogies. The word embeddings are the same as Table 2.

The trends in accuracy between distance measures and across word embedding dimensionality in Table 3 are similar to those seen with the Google word analogies. One difference is that the accuracy of COSMULT consistently increases with the dimension of the word embeddings.

References

[1] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543, 2014.