Word2Vec

Michael Collins, Columbia University

Motivation

- We can easily collect very large amounts of unlabeled text data
- Can we learn useful representations (e.g., word embeddings) from unlabeled data?

Bigrams from Unlabeled Data

- Given a corpus, extract a training set $\{x^{(i)}, y^{(i)}\}$ for $i = 1 \ldots n$, where each $x^{(i)} \in \mathcal{V}$, $y^{(i)} \in \mathcal{V}$, where $\mathcal{V}$ is the vocabulary
- For example,

  Hispaniola quickly became an important base from which Spain expanded its empire into the rest of the Western Hemisphere .

  Given a window size of $+/-3$, for $x =$ base we get the pairs

$$(\texttt{base}, \texttt{became}), (\texttt{base}, \texttt{an}), (\texttt{base}, \texttt{important}),$$
$$(\texttt{base}, \texttt{from}), (\texttt{base}, \texttt{which}), (\texttt{base}, \texttt{Spain})$$

Learning Word Embeddings

- Given a corpus, extract a training set $\{x^{(i)}, y^{(i)}\}$ for $i = 1 \ldots n$, where each $x^{(i)} \in \mathcal{V}$, $y^{(i)} \in \mathcal{V}$, where $\mathcal{V}$ is the vocabulary
- For each word $w \in \mathcal{V}$, define word embeddings $\theta'(w) \in \mathbb{R}^d$ and $\theta(w) \in \mathbb{R}^d$
- Define Θ', Θ to be the two matrices of embeddings parameters
- Can then define

$$p(y^{(i)}|x^{(i)}; \Theta, \Theta') = \frac{\exp\{\theta'(x^{(i)}) \cdot \theta(y^{(i)})\}}{Z(x^{(i)}; \Theta, \Theta')}$$

where $Z(x^{(i)}; \Theta, \Theta') = \sum_{y \in \mathcal{V}} \exp\{\theta'(x^{(i)}) \cdot \theta(y)\}$

Learning Word Embeddings (Continued)

- Can define

$$p(y^{(i)}|x^{(i)};\Theta,\Theta') = \frac{\exp\{\theta'(x^{(i)})\cdot\theta(y^{(i)})\}}{Z(x^{(i)};\Theta,\Theta')}$$

  where $Z(x^{(i)};\Theta,\Theta') = \sum_{y\in\mathcal{V}}\exp\{\theta'(x^{(i)})\cdot\theta(y)\}$

- A first objective function that can be maximized using stochastic gradient:

$$\begin{aligned} L(\Theta,\Theta') &= \sum_{i=1}^{n}\log p(y^{(i)}|x^{(i)};\Theta,\Theta') \\ &= \sum_{i=1}^{n}\left(\theta'(x^{(i)})\cdot\theta(y^{(i)})\} - \underbrace{\log\sum_{y\in\mathcal{V}}\exp\{\theta'(x^{(i)})\cdot\theta(y)\}}_{\text{Expensive!}}\right) \end{aligned}$$

An Alternative: Negative Sampling

- Given a corpus, extract a training set $\{x^{(i)}, y^{(i)}\}$ for $i = 1 \dots n$, where each $x^{(i)} \in \mathcal{V}$, $y^{(i)} \in \mathcal{V}$, where $\mathcal{V}$ is the vocabulary
- In addition, for each i sample $y^{(i,k)}$ for $k = 1 \dots K$ from a "noise" distribution $p_n(y)$. E.g., $p_n(y)$ is the unigram distribution over words y
- A new loss function:

$$
\begin{aligned}
L(\Theta', \Theta) &= \sum_{i=1}^{n} \log \frac{\exp\{\theta'(x^{(i)}) \cdot \theta(y^{(i)})\}}{1 + \exp\{\theta'(x^{(i)}) \cdot \theta(y^{(i)})\}} \\
&\quad + \sum_{i=1}^{n} \sum_{k=1}^{K} \log \frac{1}{1 + \exp\{\theta'(x^{(i)}) \cdot \theta(y^{(i,k)})\}}
\end{aligned}
$$