# word2vec Explained: Deriving Negative Sampling

Yoav Goldberg and Omer Levy

February 14, 2014

# Skip-gram Model

We want to maximize the probability of contexts given words.

$$\arg\max_{\theta} \prod_{w \in Text} \left[ \prod_{c \in C(w)} p(c \mid w; \theta) \right]$$

# Skip-gram Model

We want to maximize the probability of contexts given words.

$$\arg\max_{\theta} \prod_{w \in Text} \left[ \prod_{c \in C(w)} p(c \mid w; \theta) \right]$$

$$\arg\max_{\theta} \prod_{(w,c) \in D} p(c \mid w; \theta)$$

# Parameterization with Softmax

We model conditional probabilities with softmax:

$$p(c \mid w; \theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}$$

# Parameterization with Softmax

We model conditional probabilities with softmax:

$$p(c \mid w; \theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}$$

Take logs and switch to sum:

$$\arg\max_{\theta} \sum_{(w,c) \in D} \log p(c \mid w)$$

# Parameterization with Softmax

We model conditional probabilities with softmax:

$$p(c \mid w; \theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}}$$

Take logs and switch to sum:

$$\arg\max_{\theta} \sum_{(w,c) \in D} \log p(c \mid w)$$

Expanded form:

$$\sum_{(w,c) \in D} \left( v_c \cdot v_w - \log \sum_{c'} e^{v_{c'} \cdot v_w} \right)$$

# Negative Sampling Setup

We ask: did $(w, c)$ come from the data?

$$p(D = 1 \mid w, c; \theta) = \frac{1}{1 + e^{-v_c \cdot v_w}} \equiv \sigma(v_c \cdot v_w)$$

# Negative Sampling Setup

We ask: did $(w, c)$ come from the data?

$$p(D = 1 \mid w, c; \theta) = \frac{1}{1 + e^{-v_c \cdot v_w}} \equiv \sigma(v_c \cdot v_w)$$

Objective with negatives:

$$\arg\max_{\theta} \sum_{(w,c) \in D} \log \sigma(v_c \cdot v_w) + \sum_{(w,c) \in D'} \log \sigma(-v_c \cdot v_w)$$

# Negative Sampling Derivation

We want to maximize the probability that observed pairs are from the data.

$$\arg\max_{\theta} \prod_{(w,c)\in D} p(D = 1 \mid w, c; \theta)$$

# Negative Sampling Derivation

We want to maximize the probability that observed pairs are from the data.

$$\arg\max_{\theta} \prod_{(w,c)\in D} p(D = 1 \mid w, c; \theta)$$

$$= \arg\max_{\theta} \log \prod_{(w,c)\in D} p(D = 1 \mid w, c; \theta)$$

# Negative Sampling Derivation

We want to maximize the probability that observed pairs are from the data.

$$\arg\max_{\theta} \prod_{(w,c)\in D} p(D = 1 \mid w, c; \theta)$$

$$= \arg\max_{\theta} \sum_{(w,c)\in D} \log p(D = 1 \mid w, c; \theta)$$

# Negative Sampling Derivation

We want to maximize the probability that observed pairs are from the data.

$$\arg\max_{\theta} \prod_{(w,c)\in D} p(D=1 \mid w,c;\theta)$$

$$= \arg\max_{\theta} \sum_{(w,c)\in D} \log p(D=1 \mid w,c;\theta)$$

Using the sigmoid:

$$p(D=1 \mid w,c;\theta) = \sigma(v_c \cdot v_w)$$

# Negative Sampling Derivation

We want to maximize the probability that observed pairs are from the data.

$$\arg\max_{\theta} \prod_{(w,c) \in D} p(D = 1 \mid w, c; \theta)$$

$$= \arg\max_{\theta} \sum_{(w,c) \in D} \log p(D = 1 \mid w, c; \theta)$$

Using the sigmoid:

$$p(D = 1 \mid w, c; \theta) = \sigma(v_c \cdot v_w)$$

So the objective becomes:

$$\arg\max_{\theta} \sum_{(w,c) \in D} \log \sigma(v_c \cdot v_w)$$

# Adding Negative Samples

To prevent trivial solutions, introduce negative pairs $D'$.

$$\arg\max_{\theta} \prod_{(w,c)\in D} p(D=1 \mid w,c;\theta) \prod_{(w,c)\in D'} p(D=0 \mid w,c;\theta)$$

# Adding Negative Samples

To prevent trivial solutions, introduce negative pairs $D'$.

$$\arg\max_{\theta} \prod_{(w,c)\in D} p(D=1 \mid w,c;\theta) \prod_{(w,c)\in D'} p(D=0 \mid w,c;\theta)$$

$$= \arg\max_{\theta} \sum_{(w,c)\in D} \log \sigma(v_c \cdot v_w) + \sum_{(w,c)\in D'} \log \sigma(-v_c \cdot v_w)$$