Statistical Natural Language Processing
Assignment 1

October 21, 2024

Victor Pekkari

Question 2

2 (a)

I had three layers all containing 300 neurons. My learning rate was set to 0.001. I had a weight decay of 1e-5. I drop out percentage of 20% after layer 2 and layer 3. And I used the embedding layer with 300 dimensions. These hyperparameters gave me over 80% accuracy on the development set. I used drop out and weight decay to prevent overfitting.
Weight decay prevents overfitting because it penlaises large weights which is common in overfitted networks. Without dropout a network can more easily fit outliers, but you need many neurons working together to fit outliers, thats why drop out prevents this. I also didn't want my network architecture to become too complex as it increases the risk of overfitting.

2 (b)

Using random initialized weights for the embedding layer made training a lot slower (timewise), and made it converge to a "stable" accuracy score in more epochs. It also made the model overfit to the training data more, than with the pre-trained embedding layer. I think the perfect combination would have been to start with the pre-initialized embedding layer, and then fine tuning it.

Question 3

3 (a)

$$P(dog|the) = \frac{1}{2} \tag{1}$$

$$P(cat|the) = \frac{1}{2} \tag{2}$$

3 (b)

$$w_{dog} = w_{cat} \implies v_{the}^t \cdot w_{dog} = v_{the}^t \cdot w_{cat} = z \tag{1}$$

$$P(X = \text{the} \mid Y = \text{dog}) = \frac{\exp(v_{\text{the}}^t \cdot w_{\text{dog}})}{\sum_{y'} \exp(v_{\text{the}}^t \cdot w_{\text{y'}})} \tag{2}$$

$$= \frac{\exp(z)}{\sum_{y'} \exp(v_{\text{the}}^t \cdot w_{\text{y'}})} = P(X = \text{the} \mid Y = \text{cat}) \tag{3}$$

$$\implies \forall v_{the} \in \mathbb{R}^2 \colon\ P(cat|the) = P(dog|the) = \frac{1}{2}$$

3 (c)

$$\{(the, dog), (dog, the), (the, cat), (cat, the), (a, dog), (dog, a), (a, cat), (cat, a)\}$$

3 (d)

our training examples tell us that the:

- "a" occurs in the same way with "dog" and "cat"
- "the" occurs in the same way with "dog" and "cat"
- "dog" occurs in the same way with "the" and "a"
- "cat" occurs in the same way with "the" and "a"

We want to treat "dog" and "cat" similarly since they always occur in the same context. We also want to treat "the" and "a" similarly because of the same reason.

Our desired probabilites are:

$$P(dog|the) \approx 0.5$$
$$P(cat|the) \approx 0.5$$
$$P(dog|a) \approx 0.5$$
$$P(cat|a) \approx 0.5$$

By using the similar vector-encodings, and context-encodings for similar words we get:

$$v_{the} = v_a = (1, 0)$$
$$v_{dog} = v_{cat} = (0, 1)$$

and the following context vectors:

$$w_{the} = w_a = (1, 0)$$
$$w_{cat} = w_{dog} = (0, 1)$$