Quiz Submissions - Quiz 3

WENXIN ZHANG (username: wz2164)

Attempt 1

Written: Oct 7, 2021 2:55 PM - Oct 8, 2021 4:16 PM

Submission View

Your quiz has been submitted successfully.

Question 1 **1 / 1 point**

Remember our definition for perplexity:

$$Perplexity = 2^{B_2}$$

where

$$\mathbf{B_x = -\frac{1}{\sum_{n=1}^{N} T^N} \sum_{n=1}^{T} \sum_{t=1}^{T^n} \log_x p(w_t^n \mid w_{<t}^n)}$$

Show that perplexity is identical regardless of what base we use. In other words, that:

$$Perplexity = d^{B_d} = c^{B_c}$$

for all positive d, c.

quiz3-1.png (56.32 KB)

Question 2 **3 / 3 points**

Suppose we are computing perplexity for a single token in a single example (this is pretty unusual, but let's just take this as an exercise).

Answer the following questions in order:

a. We have a vocabulary of size 4, and our conditional probability distribution

$$P(y_t \mid y_{<t})$$

is [0.25, 0.25, 0.25, 0.25]. The true token ID is 0 (i.e. the first token in the vocab). What is the "perplexity" for this dataset?

b.

$$P(y_t \mid y_{<t})$$

is [0.25, 0.00, 0.00, 0.75]. What is the perplexity now?

c.

$$P(y_t \mid y_{<t})$$

is [0.00, 0.25, 0.00, 0.75]. What is the perplexity now?

Answer for blank # 1: 4

Answer for blank # 2: 4

Answer for blank # 3: Infinity

Question 3 **1 / 1 point**

We compare two language models (assume same architecture, vocabulary, but trained on different corpora) A and B. Is it possible for A to have a higher accuracy in predicting the correct tokens in language modeling, while B has a lower perplexity? Explain why/why not.

Yes. Considering A has a much richer corpora to train itself, its perplexity can be higher.

Question 4 **1 / 1 point**

What is the space complexity for storing an RNN-based language model (excluding input embedding and output embedding layers)?

Assume a naive transition function, such as the one in the lecture:

$$h_{t+1} = R(h_t, w_{t+1}) = \sigma(W_r h_t + W_x w_{t+1} + b)$$

Let D=hidden dimension size, V=vocabulary size, N=number of training examples, T=max sequence length.

Hint:

If you are unfamiliar with what "space complexity" means, read it as "at what rate will the memory required to store the model grow, in the worst case scenario". For instance, in the case of a simple linear regression with D features and N examples, the space complexity is O(D). In the case of k-nearest-neighbors with D features and N examples for predicting a scalar, it would be O(ND).

- O(DV)
- O(D^2)
- O(D^2T)
- O(ND + DT)

Question 5 **3 / 3 points**

Respond Yes/No to each of the following questions in order.

a. We would like to compare two language models. In model A we convert the training, validation and test sets to lowercase before tokenizing, building a vocabulary from the top most frequent 10,000 tokens, and training and evaluating our model. In model B, we use the same architecture, but we did not do the lowercasing. Can we directly compare the perplexities of these models?

b. We would like to compare two language models. We use the same vocabulary for both models, but we use Wikipedia as our training corpus for model A, and IRC chat logs as our training corpus on model B.

Both models are evaluated on the same held-out test set. Can we directly compare the perplexities of these models?

c. We would like to compare two language models. We use the same vocabulary for both models. For model A, we use an 8-layer RNN with dropout, trained with an Adam optimizer. For model B, we just compute the unigram distribution of tokens, and use that as our "conditional" probability distribution for every token. Can we directly compare the perplexities of these models?

Answer for blank # 1: No

Answer for blank # 2: Yes

Answer for blank # 3: Yes

Question 6 **1 / 1 point**

Explain in your own words why we might want to tie the weights of the input and output embeddings of a language model.

1. Embedding space maps the token that we have in vocabulary into dense representation (in embedding dimension). It makes sense that we always have the same representation for a token in the vocabulary, since all we want to do is to optimize this vectorized token in embedding dimension to fit our training data well. The meaning of this embedding at the input part and the output part are the same here.

2. Weight tying can reduce the size of neural translation models to less than half of their original size without harming their performance.

Question 7 **1 / 1 point**

Explain in your own words the relationship between hidden states h_t, input tokens x_t, predictions y_t, and the Markov property in the context for RNNs.

At each step t, the predictions

$$y_t$$

are determined by input tokens

$$x_t$$

and hidden states

$$h_{t-1}$$

; after which this predicted

$$y_t$$

would act as

$$h_t$$

to feed into next step of RNNs for predicting

$$y_{t+1}$$

. RNNs have Markov property. From the property of RNNs:

$$p(y_{t+1} \mid y_t, y_{t-1}, \ldots, y_0) = p(y_{t+1} \mid x_t, h_{t-1}, y_{t-1}, \ldots, y_0) = p(y_{t+1} \mid x_t, h_{t-1}) =$$

, it's obvious that for RNNs, the conditional probability distribution of the next state, conditioned on

both the past states and the current state, is equal to the conditional probability of the next state given the current state only.

Question 8 **0.1 / 0.1 points**

How long did you spend to complete this quiz (in hours)?

1 Hour.

Done