Chapter8

April 28, 2019

8.1.5. Exercises

1. **Improve the above model.**

   a. **Incorporate more than the past 4 observations? How many do you really need?**

   b. **How many would you need if there were no noise? Hint - you can write sin and cos as a differential equation.**

   c.**Can you incorporate older features while keeping the total number of features constant? Does this improve accuracy? Why?**

   d.**Change the architecture and see what happens.**

2. **An investor wants to find a good security to buy. She looks at past returns to decide which one is likely to do well. What could possibly go wrong with this strategy?**

3. **Does causality also apply to text? To which extent?**

4. **Give an example for when a latent variable autoregressive model might be needed to capture the dynamic of the data.**

8.2.5. Exercises

1. **Suppose there are 100,000 words in the training data set. How many word frequencies and multi-word adjacent frequencies does a four-gram need to store?**

   In English, the Zipf's law in the n-gram data exhibits two regimes: one among words with frequencies above about 0.01% (Zipf's exponent 1) and another (1.4) among words with frequency below 0.0001%.

   By Zipf's law, the normalized frequency of elements of rank k, $f(k; s, N)$ is:

   $$f(k; s; N) = \frac{1}{k^s \sum_1^N \frac{1}{n^s}},$$

   where N = 100,000 is the number of words in the English language, s = 1.07 for unigram.

2. **Review the smoothed probability estimates. Why are they not accurate? Hint - we are dealing with a contiguous sequence rather than singletons.**

   Laplace smoothing allows the assignment of non-zero probabilities to words which do not occur in the sample. Lots of contiguous sequence may not occurs in the bag of word, hence the smoothed probability is not accurate in the right tail.

3. **How would you model a dialogue?**

   First, we can decide which type of dialogue (conversational models) are we going to model (rule-based, retrieval-based, neural generative models, grounded/visual, chit-chat vs. task-based, etc.)?

   Next, choose framework for modelling, such as for neural generative models choosed Semantically Conditioned LSTM-based model. (https://arxiv.org/abs/1508.01745); or Deep Reinforcement Learning for Dialogue Generation (https://aclweb.org/anthology/D16-1127) and so on.

4. **Estimate the exponent of Zipf's law for unigrams, bigrams and trigrams.**

8.3.5. Exercises

1. **If we use an RNN to predict the next character in a text sequence, how many output dimensions do we need?**

   The output dimension should be the length of the dictionary of unique characters from both train and test dataset.

2. **Can you design a mapping for which an RNN with hidden states is exact? Hint - what about a finite number of words?**

3. **What happens to the gradient if you backpropagate through a long sequence?**

   High power of matrices can lead to explode or vanish gradients.

4. **What are some of the problems associated with the simple sequence model described above?**

   a. numerically unstable;
   b. difficulty of long-term information preservation and short-term input skipping in latent variable models

8.4.5. Exercises

1. **What other mini-batch data sampling methods can you think of?**

   Non-uniform mini-batch sampling. (i.e. suppressing the probability of similar data points in the same mini-batch, which will reduce the stochastic gradient noise, leading to faster convergence).

2. **Why is it a good idea to have a random offset?**

   In this way, we can get both coverage(by sequential partitioning strategies) and randomness.

   a. **Does it really lead to a perfectly uniform distribution over the sequences on the document?**

```
Picking just a random set of initial positions is no good either since it does not guarar

        $$(11/n)^n  e^{1}$$
```

b. **What would you have to do to make things even more uniform?**

Offset for the iterator over the data for uniform starts offset = int(random.uniform(0,num_steps))

3. **If we want a sequence example to be a complete sentence, what kinds of problems does this introduce in mini-batch sampling? Why would we want to do this anyway?**

Since a complete sentence is long, it is acceptable to discard half-empty mini-batch. Since these sequences are covered by part of other batches in mini-batch sampling.

8.5.8. Exercises

1. **Show that one-hot encoding is equivalent to picking a different embedding for each object.**

Elementary row and column operations on a matrix are rank-preserving.

2. **Adjust the hyperparameters to improve the perplexity.**

   a. **How low can you go? Adjust embeddings, hidden units, learning rate, etc.**

   b. **How well will it work on other books by H. G. Wells, e.g. The War of the Worlds.**

3. **Run the code in this section without clipping the gradient. What happens?**

4. **Set the pred_period variable to 1 to observe how the under-trained model (high perplexity) writes lyrics. What can you learn from this?**

5. **Change adjacent sampling so that it does not separate hidden states from the computational graph. Does the running time change? How about the accuracy?**

6. **Replace the activation function used in this section with ReLU and repeat the experiments in this section.**

7. **Prove that the perplexity is the inverse of the harmonic mean of the conditional word probabilities.**