Deep Learning CS583 Fall 2020
Quiz 1 - Section B

October 14th, 2021

Instructor: Jia Xu

Student name: Archana Kalburgi

Student ID: 10469491

Student email address: akalburg@stevens.edu

- **Read these instructions carefully**
- Fill-in your personal info, as indicated above.
- You have 24 hours.
- There are three questions. Each question worths the same (5 points).
- Both computer-typed and hand-writing in the very clear form are accepted.
- This is an open-book test.
- You should work on the exam only by yourself.
- Submit your PDF/Doc/Pages **by 12:30 Oct 15th** on Canvas under Final exam.

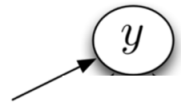
good luck!

1 Question

You are given one or several hidden nodes "h", two inputs x_1, x_2, and the output y. Draw a neural network and assign the weights and bias that performs OR operation:

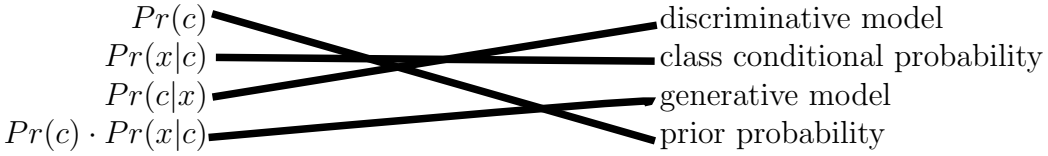
- if $x_1 = 0$, $x_2 = 0$, then $y = 0$
- if $x_1 = 1$, $x_2 = 0$, then $y = 1$
- if $x_1 = 0$, $x_2 = 1$, then $y = 1$
- if $x_1 = 1$, $x_2 = 1$, then $y = 1$

The activation function outputs 1 if the input is greater than zero and outputs 0 otherwise.

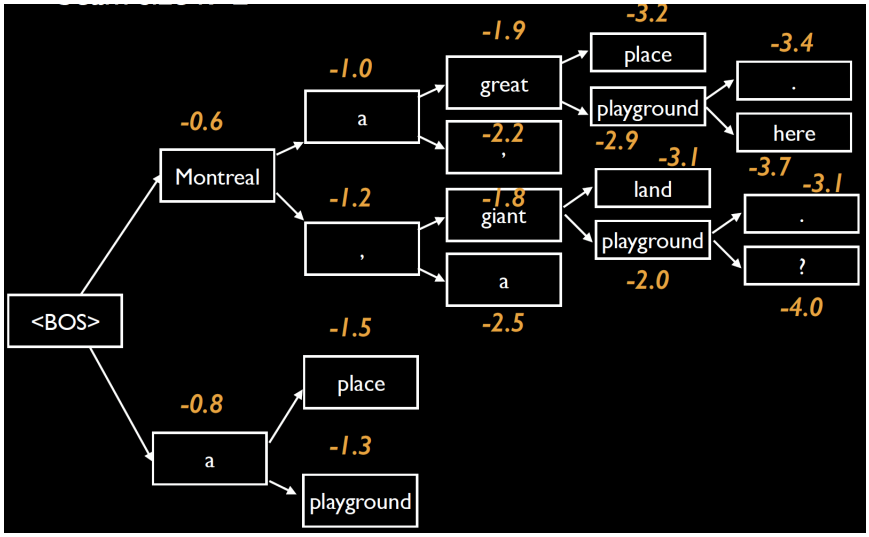


2 Question

- Align each term with its explaination, given c as class, x as input observation.



- In a lecture slide below, the orange color indicates the log-likelihood of the partial path from the beginning of the sentence (BOS) until the current output. What is the prediction output if we change the beam size.



  - beam size k=1, output word sequence:
  - beam size k=2, output word sequence:
  - What are the advantage and the disadvantage of a larger beam size?

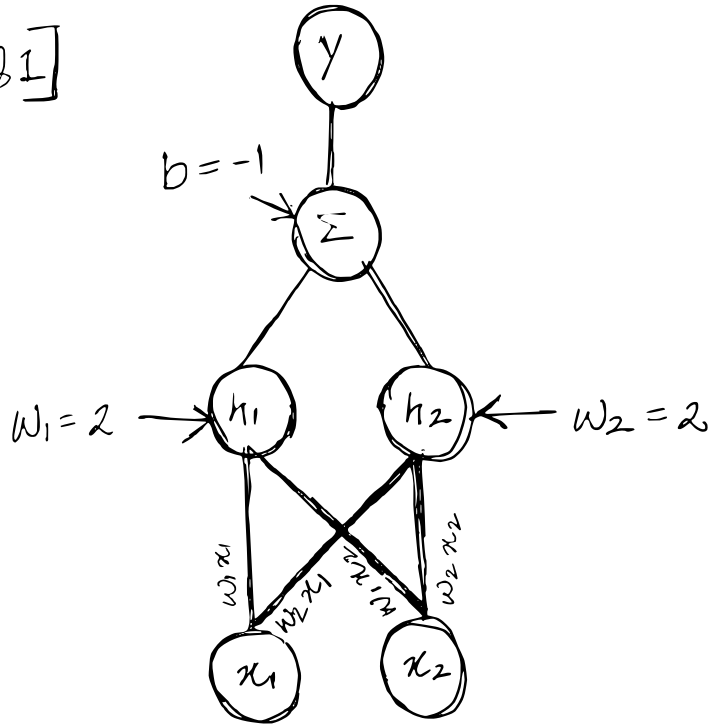
3 Question

- Briefly explain the trigram method of language modeling.

- What is the procedure of 5-fold cross-validation, and what is its advantage over the traditional approach of simply splitting one's available data into a training set and a validation set?

- We have seen that averaging the outputs from multiple models typically gives better results than using just one model. Let's say that we're going to average the outputs from 10 models. Of course, we want 10 good models, i.e. models that also perform well individually. What additional property of a collection of 10 models makes that collection a good candidate for output averaging?

Q1]



- $w_1 \cdot w_2 \cdot b = 2, 2, -1$

- OR operation table

X_1	X_2	Y
0	0	0
0	1	1
1	0	1
1	1	1

1) if $x_1 = 0$ & $x_2 = 0$

$$h_1 = w_1 x_1 + w_1 x_2 + b$$

$$= 2 \times 0 + 2 \times 0 - 1 = -1$$

$$h_2 = w_2 x_1 + w_2 x_2 + b$$

$$= 2 \times 0 + 2 \times 0 - 1 = -1$$

$$z = (-1) + (-1) = -2$$

$$y = \text{activation}(-2) = 0$$

2) if $x_1 = 0$ & $x_2 = 1$

$$h_1 = w_1 x_1 + w_1 x_2 + b$$

$$= 2 \times 0 + 2 \times 1 - 1 = 1$$

$$h_2 = w_2 x_1 + w_2 x_2 + b$$

$$= 2 \times 0 + 2 \times 1 - 1 = 1$$

$$z = 1 + 1 = 2$$

$$y = \text{activation}(2) = 1$$

3) if $x_1 = 1$ & $x_2 = 0$

$$h_1 = w_1 x_1 + w_1 x_2 + b$$
$$= 2 \times 1 + 2 \times 0 + (-1) = 1$$

$$h_2 = w_2 x_1 + w_2 x_2 + b$$
$$= 2 \times 1 + 2 \times 0 - 1 = 1$$

$$z = 1 + 1 = 2$$

$$y = \text{activ fun}(z) = 1$$

4) if $x_1 = 1$ & $x_2 = 1$

$$h_1 = w_1 x_1 + w_1 x_2 + b$$
$$= 2 \times 1 + 2 \times 1 + (-1) = 3$$

$$h_2 = w_2 x_1 + w_2 x_2 + b$$
$$= 2 \times 1 + 2 \times 1 - 1 = 3$$

$$z = 3 + 3 = 6$$

$$y = \text{activ fun}(6) = 1$$

Q2] a)

$Pr(c)$: Prior probability

$Pr(x|c)$: Class conditional probability.

$Pr(c|x)$: Discriminative model

$Pr(c) \cdot Pr(x|c)$: Generative model.

→

Q2 b] i) Word sequence for beam size K=1

Output:

Montreal a great playground.

ii) word sequence for beam size K=2

Output:

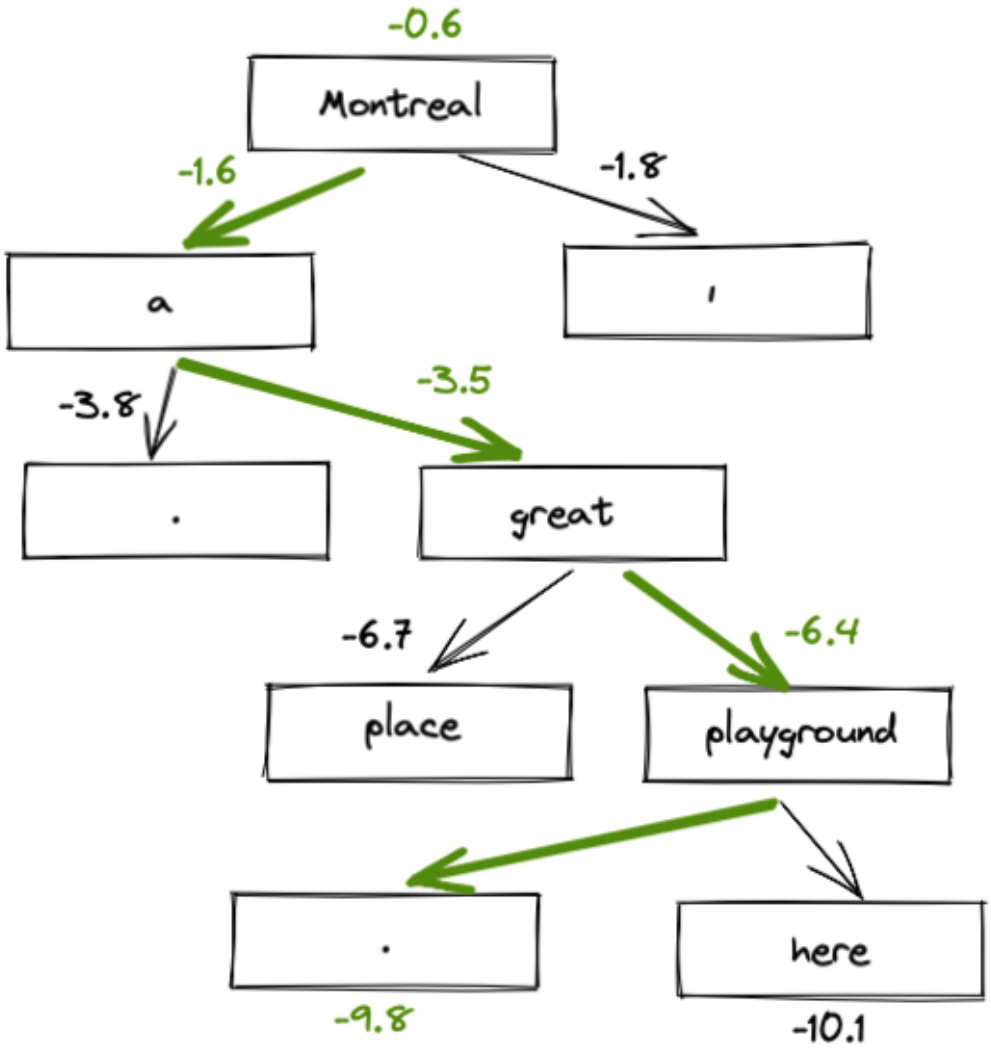
Montreal, giant playground.

iii) Advantages of using larger beam size:-

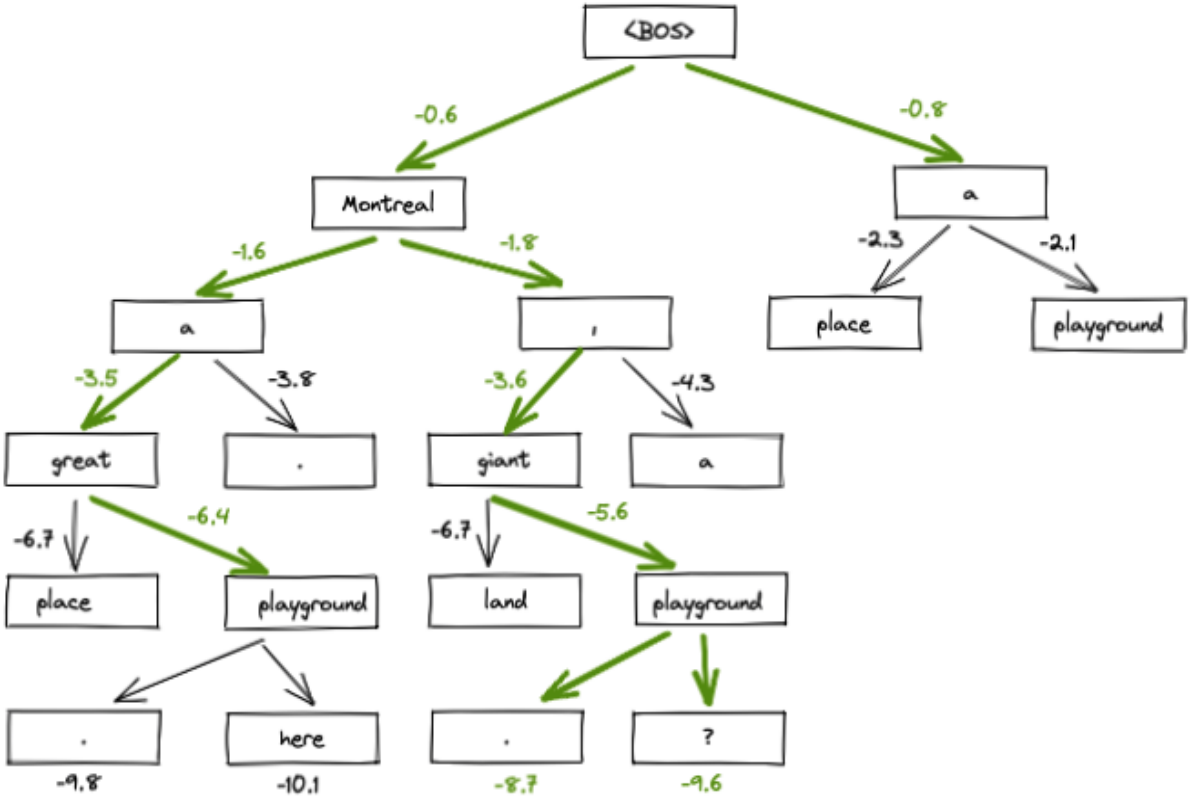
- with larger beam size we can predict a gramatically correct/accurate scentence compared to when K=1, which becomes a greedy search

disadvantage: the complexity of the algorithm increases, since each steap tracks V (vocabulary) words, ∴ complexity becomes $O(V^T)$

Beam search k=1
Output: Montreal , giant playground .



Beam search, k=2:
Output : Montreal , giant playground .



Q3. a] Trigram language model consists of the following:-

1. finite set V,

   V: vocabulary of the language model

2. parameter: $q(w|u,v)$ for each trigram $u.v.w$ such that.

   $w \in V \cup \{STOP\}$ and $u, v \in V \cup \{*\}$

for any sentence $x_1, x_2 \cdots x_n$, where $x_i \in V$ for $i = 1, 2 \cdots (n-1)$, the probability of a sentence under the trigram language model is given by.

$$p(x_1, x_2 \cdots x_n) = \prod_{i=1}^{n} q(x_i | x_{i-2}, x_{i-1})$$

q is estimated using maximum likelihood estimation.

$$q(w_i | w_{i-2}, w_{i-1}) = \frac{count(w_{i-2}, w_{i-1}, w_i)}{count(w_{i-2}, w_{i-1})}$$

- Trigram model are conditioned on the previous two words. rather than just the previous word.
- Intuition of trigram model: instead of computing the probability of a word given its entire history, we can estimate the probability of a word given all the previous words. by using conditional probability of preceding word
- Trigram model computes probability by looking two words into the past. using markov assumption.

Q3 b] 5-fold cross validation procedure:

1) Dataset is shuffled randomly
2) Shuffled dataset is split into 5 groups.
3) Set aside a group as test data
4) Set the remaining groups as training set
5) Fit a model on train data & test on test data.
6) Record the accuracy.
7) Repeat 3, 4, 5, 6 for each group.

Advantages

1. The procedure ensures that each observation is assigned to a test set once and used to train the model 4 times. (K times)
2. Model trained used K-cross validation are less biased compared to traditional train, test and validation set.

Q3 c] Averaging the outputs from multiple model is an ensemble method. Every model contributes an equal amount to the final output.

Limitation:-

- Every model has equal contribution, to the final output, but some model can perform much better or much worse than the other models.

Additional property:

- A weighted ensemble instead of collection of 10 models can increase the performance. Here the contribution of each model to the final output is weighted by the performance of the model.
- Weights indicate the percentage of trust from each model.