NLP Midterm Exam

Álvaro Faúndez

November 2021

Question 1

We have a vocabulary V of size 300. How many parameters will you have to estimate for:

1. an MLE unigram language model
2. a bigram language model with add-one smoothing
3. a maximum likelihood trigram language model

Answer

1. an MLE unigram language model:
   It needs to estimate $|V| = 300$ parameters.
2. a bigram language model with add-one smoothing
   It needs to estimate $|V|^2 = 300^2$ parameters.
3. a maximum likelihood trigram language model
   It needs to estimate $|V|^3 = 300^3$ parameters.

Question 2

Consider the following training corpus T of sentences:

1. `<s> I am Sam </s>`
2. `<s> Sam I am </s>`
3. `<s> I do not like green eggs and ham </s>`

Compute the following maximum likelihood parameters:

- $p(\texttt{I} \mid \texttt{<s>}) =$
- $p(\texttt{</s>} \mid \texttt{Sam}) =$
- $p(\texttt{I} \mid \texttt{do}) =$
- $p(\texttt{ham} \mid \texttt{eggs}, \texttt{and}) =$

Answer

$$p(\texttt{I} \mid \texttt{<s>}) = \frac{count(\texttt{<s> I})}{count(\texttt{<s>})} = \frac{2}{3} \tag{1}$$

$$p(\texttt{</s>} \mid \texttt{Sam}) = \frac{count(\texttt{Sam </s>})}{count(\texttt{Sam})} = \frac{1}{2} \tag{2}$$

$$p(\texttt{I} \mid \texttt{do}) = \frac{count(\texttt{do I})}{count(\texttt{do})} = \frac{0}{1} \tag{3}$$

$$p(\texttt{ham} \mid \texttt{eggs}, \texttt{and}) = \frac{count(\texttt{eggs and ham})}{count(\texttt{eggs and})} = \frac{1}{1} \tag{4}$$

Question 3

We have the following training corpus:

1. `the green book STOP`
2. `my blue book STOP`
3. `his green house STOP`
4. `book STOP`

Assume we have a trigram language model with linear interpolation based on this corpus, with $\lambda_i = \frac{1}{3}$ for all i. Compute the value of the parameter $p(\texttt{book} \mid \texttt{the, green})$ under this model. Assume `STOP` as part of your unigram model.

Answer

The corpus C has 14 tokens. The probabilities needed are:

$$p_{ML}(\texttt{book} \mid \texttt{the green}) = \frac{count(\texttt{the green book})}{count(\texttt{the green})} = \frac{1}{1} \tag{5}$$

$$p_{ML}(\texttt{book} \mid \texttt{green}) = \frac{count(\texttt{green book})}{count(\texttt{green})} = \frac{1}{2} \tag{6}$$

$$p_{ML}(\texttt{book}) = \frac{count(\texttt{book})}{|C|} = \frac{3}{14} \tag{7}$$

Finally, the interpolation

$$\begin{aligned} p(\texttt{book} \mid \texttt{the green}) &= \lambda_3 \times p_{ML}(\texttt{book} \mid \texttt{the green}) + \lambda_2 \times p_{ML}(\texttt{book} \mid \texttt{green}) + \lambda_1 \times p_{ML}(\texttt{book}) \\ &= \frac{1}{3} \times \frac{count(\texttt{the green book})}{count(\texttt{the green})} + \frac{1}{3} \times \frac{count(\texttt{green book})}{count(\texttt{green})} + \frac{1}{3} \times \frac{count(\texttt{book})}{|C|} \\ &= \frac{1}{3}(\frac{1}{1} + \frac{1}{2} + \frac{3}{14}) \\ &= 0.57142857142 \end{aligned} \tag{8}$$

Question 4

Given a vocabulary V of size 1000 and a tag set T of size 30, suppose you wish to train a bigram HMM tagger. How many transition and emission parameters will your model have? Explain.

Answer

The emissions probabilities in a bigram HMM $e(w \mid t)$ are needed for all $winV$ and each $tinT$. That means that there are $|V| \times |T| = 1000 \times 30 = 30000$ possible combinations of words and tags.

The transitions probabilities in a bigram HMM $p(t_i \mid t_k)$ are needed for each possible combination of two transitions. That means there are $|T| \times |T| = 30 \times 30 = 900$ possible combinations of transitions pairs.

In the WSJ corpus example we reviewed in class, the transitions matrix also included the start token `<s>`. That would make the transitions matrix $31 \times 30 = 930$. If we also include the stop token `</s>`, the size would be $31 \times 31 = 961$. The emissions would stay the same.

Question 5

Let $V = \{Karlsson, lives, happily\}$ Let $T = \{N, V\}$

We have a trigram HMM tagger that has the following non-zero parameters (all other parameters are zero):

Transition parameters:

- $p(N \mid START) = 0.5$
- $p(V \mid START) = 0.5$
- $p(N \mid START, N) = 1.0$
- $p(N \mid START, V) = 0.9$
- $p(V \mid START, V) = 0.1$
- $p(STOP \mid N, N) = 1.0$
- $p(STOP \mid V, V) = 1.0$
- $p(STOP \mid V, N) = 1.0$
- $p(s \mid u, v) = 0$ for all other p parameters

Emission parameters:

- $e(Karlsson \mid N) = 0.8$
- $e(happily \mid N) = 0.2$
- $e(lives \mid V) = 0.7$
- $e(happily \mid V) = 0.3$

Under this model, how many pairs of sequences $x_1 \ldots x_n, y_1 \ldots y_n$ have non-zero probability $p(x_1, \ldots x_n, y_1 \ldots y_{n+1}) \geq 0$? Show all of these possible sequences.

Answer

Given the emissions:

$e(w \mid t)$	`Karlsson`	`lives`	`happily`
N	0.8	0	0.8
V	0	0.7	0.7

We can deduce that the possible N are `Karlsson` and `happily` and the possible V are `lives` and `happily`.

And from the transitions, we can deduce:

- then sentences can start either with an N or a V
  - If the sentence starts with an N, the only possibility for the next word is another N
  - If the sentence starts with a V, the only possibility for the next word is another N or V
- Since there are no transitions involving a combination of only N or V, the STOP must come after two words

That means that the only sequences with non-zero probabilities are:

- $(t_1, t_2, t_3) = (N, N, STOP)$
  - $(x_1, x_2, x_3) = (\texttt{Karlsson}, \texttt{Karlsson}, \texttt{STOP})$
  - $(x_1, x_2, x_3) = (\texttt{Karlsson}, \texttt{happily}, \texttt{STOP})$
  - $(x_1, x_2, x_3) = (\texttt{happily}, \texttt{happily}, \texttt{STOP})$
  - $(x_1, x_2, x_3) = (\texttt{happily}, \texttt{Karlsson}, \texttt{STOP})$
- $(t_1, t_2, t_3) = (V, N, STOP)$
  - $(x_1, x_2, x_3) = (\texttt{lives}, \texttt{Karlsson}, \texttt{STOP})$
  - $(x_1, x_2, x_3) = (\texttt{lives}, \texttt{happily}, \texttt{STOP})$
  - $(x_1, x_2, x_3) = (\texttt{happily}, \texttt{Karlsson}, \texttt{STOP})$
  - $(x_1, x_2, x_3) = (\texttt{hapilly}, \texttt{happily}, \texttt{STOP})$
- $(t_1, t_2, t_3) = (V, V, STOP)$
  - $(x_1, x_2, x_3) = (\texttt{lives}, \texttt{lives}, \texttt{STOP})$
  - $(x_1, x_2, x_3) = (\texttt{lives}, \texttt{happily}, \texttt{STOP})$
  - $(x_1, x_2, x_3) = (\texttt{happily}, \texttt{lives}, \texttt{STOP})$
  - $(x_1, x_2, x_3) = (\texttt{happily}, \texttt{happily}, \texttt{STOP})$

That makes a total of 12 possible sequences of words and tags with non-zero probabilities.

Question 6

Find one tagging error in each of the following sentences that are tagged with the Penn Treebank tagset:

1. I/PRP need/VBP a/DT flight/NN from/IN Atlanta/NN
2. Does/VBZ this/DT flight/NN serve/V B dinner/NNS
3. I/PRP have/VB a/DT friend/NN living/VBG in/IN Denver/NNP
4. Can/VBP you/PRP list/VB the/DT nonstop/JJ afternoon/NN flights/NNS

Answer

1. Atlanta should be NNP (proper noun, singular)
2. Dinner should be an NN (noun, singular or mass)
3. have should be a VBP (Verb, non-3rd ps. sing. present)
4. afternoon should be a JJ (adjective)

Question 7

We have a vocabulary $V = \{\texttt{Hello}\}$ and a constant $N \geq 1$. For any $x_1 \dots x_n$ such that $x_i \in V$ for $i = \dots (n-1)$ and $x_n = STOP$, we define

$$p(x_1, \dots, x_n) = \begin{cases} \frac{1}{N} & \text{if } n \leq N \\ 0 & \text{otherwise} \end{cases}$$

Is this a valid language model? Explain.

Answer

In order to be valid, given a corpus C a language model must:

1. $p(s) \geq 0$ for each sentence $s \in C$
2. $\Sigma_{s \in C} p(s) = 1$

Condition 1 : it is satisfied for each sentence because the two options are zero or a positive rational.

Condition 2 : it's not fullfilled. The probability function means that any sentence shorter or equal than N has probability $\frac{1}{N}$ and everything else zero. We can see that choosing $N = 3$. There are only two sentences possible shorter or equal than 3:

1. $x_1, x_2 = \texttt{Hello}, STOP$
2. $x_1, x_2, x_3 = \texttt{Hello}, \texttt{Hello}, STOP$

Those sentences have both a probability of $\frac{1}{3}$ and any other sentence has probability zero because they are longer than $N = 3$. This way, the sum of the probabilities is $\frac{2}{3}$ and not 1 as required to be a valid language model.

Question 8

Now we have a vocabulary $V = \{\texttt{Hello}, \texttt{Goodbye}\}$ and a constant $N \geq 1$. For any $x_1 \dots x_n$ such that $x_i \in V$ for $i = \dots (n-1)$ and $x_n = STOP$, we define

$$p(x_1, \dots, x_n) = \begin{cases} \frac{1}{2} & \text{if } n = 2 \\ 0 & \text{otherwise} \end{cases}$$

Is this a valid language model? Explain.

Answer

The difference with question 8 is in the probability function. Now, the probability function means that every sentence of size 2 has probability $\frac{1}{2}$ and everything else zero. We can see that choosing any N. the are always only two sentences of size 2:

1. $x_1, x_2 = \texttt{Hello}, STOP$
2. $x_1, x_2 = \texttt{Goodbye}, STOP$

Now, the sum of all the probabilities is 1, so it is a valid language model.

Question 3

Consider the task of classifying the word bass using the Naïve Bayes algorithm. The features used are bag-of-word features. Assume the following likelihoods for each word being part of a "fish" and "music" class, and equal prior probabilities for each class.

What class will Naïve Bayes assign to the sentence "I eat fresh bass after music lesson"? Show your work.

	fish	**music**
I	0.09	0.16
eat	0.29	0.06
fresh	0.10	0.05
after	0.07	0.06
music	0.04	0.15
lesson	0.08	0.11

Answer

Classes $C = \{C_{\texttt{fish}}, C_{music}\}$.

Bag of words $\{\texttt{I}, \texttt{eat}, \texttt{fresh}, \texttt{after}, \texttt{music}, \texttt{lesson}\}$

Features $\vec{x} = [1, 1, 1, 1, 1, 1]$

Priors:

$$\begin{aligned} P_{prior}(C_{\texttt{fish}}) &= \frac{1}{2} \\ P_{prior}(C_{\texttt{music}}) &= \frac{1}{2} \end{aligned} \tag{9}$$

$$\begin{aligned} P(C_{\texttt{fish}} \mid \vec{x}) &\propto P(C_{\texttt{fish}}) \times P(\vec{x} \mid C_{\texttt{fish}}) \\ &\propto P(\texttt{I} \mid C_{\texttt{fish}}) \times \\ &\quad P(\texttt{eat} \mid C_{\texttt{fish}}) \times P(\texttt{fresh} \mid C_{\texttt{fish}}) \times P(\texttt{after} \mid C_{\texttt{fish}}) \times \\ &\quad P(\texttt{music} \mid C_{\texttt{fish}}) \times P(\texttt{lesson} \mid C_{\texttt{fish}}) \\ &\propto 0.5 \times 0.09 \times 0.29 \times 0.10 \times 0.07 \times 0.04 \times 0.98 \\ &\propto 2.9232e-7 \end{aligned} \tag{10}$$

$$
\begin{aligned}
P(C_{\texttt{music}} \mid \vec{x}) &\propto P(C_{\texttt{music}}) \times P(\vec{x} \mid C_{\texttt{music}}) \\
&\propto P(\texttt{I} \mid C_{\texttt{music}}) \times \\
&\quad P(\texttt{eat} \mid C_{\texttt{music}}) \times P(\texttt{fresh} \mid C_{\texttt{music}}) \times P(\texttt{after} \mid C_{\texttt{music}}) \times \\
&\quad P(\texttt{music} \mid C_{\texttt{music}}) \times P(\texttt{lesson} \mid C_{\texttt{music}}) \\
&\propto 0.5 \times 0.16 \times 0.06 \times 0.05 \times 0.06 \times 0.15 \times 0.11 \\
&\propto 2.376e-7
\end{aligned}
\tag{11}
$$

Since $P(C_{\texttt{fish}} \mid \vec{x}) > P(C_{\texttt{music}} \mid \vec{x})$, the class assigned is $C_{\texttt{fish}}$.