2. Maximum Likelihood Estimation

Extension Note: Homework 5 due date has been extended by 1 day to **August 17 23:59UTC** .

Consider a general multinomial distribution with parameters θ. Recall that the likelihood of a dataset $\mathcal{D}$ is given by:

$$P(\mathcal{D};\theta) = \prod_{i=1}^{|\theta|} \theta_i^{c_i}$$

where c_i is the occurrence count of the i-th event.

The MLE of θ is the setting of θ that maximizes $P(\mathcal{D};\theta)$. In lecture we derived this to be

$$\theta_i^* = \frac{c_i}{\sum_{j=1}^{|\theta^*|} c_j}$$

Unigram Model

4/4 points (graded)

Consider the sequence:

A B A B B C A B A A B C A C

A unigram model considers just one character at a time and calculates $p(w)$ for $w \in \{A, B, C\}$.

What is the MLE estimate of θ? Give your result to three decimal places.

θ_A^* 6/14 ✔ **Answer:** 0.4285714286

θ_B^* 5/14 ✔ **Answer:** 0.3571428571

θ_C^* 3/14 ✔ **Answer:** 0.2142857143

Using the MLE estimate of θ on $\mathcal{D}$, which of the following sequences is most likely?

- ABC
- BBB
- ABB ✔
- AAC

Solution:

We calculate the MLE as $\frac{\text{count}(w)}{N}$ where $N = 14$ and the counts are 6, 5, and 3.

For comparing probabilities in part two, we simply multiply. We only need to compare the numerators: $6 \times 5 \times 3$, 5^3, 6×5^2, and $6^2 \times 3$.

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

Bigram Model 1

1/1 point (graded)
A bigram model computes the probability $p(\mathcal{D}; \theta)$ as:

$$p(\mathcal{D}; \theta) = \prod_{w_1, w_2 \in \mathcal{D}} p(w_2 | w_1)$$

where w_2 is a word that follows w_1 in the corpus.

This is also a multinomial model. Assume the vocab size is N. How many parameters are there?

Grading note: The formula above contains an error: the probability $p(\mathcal{D}; \theta)$ in a bigram model is generally:

$$p(\mathcal{D}; \theta) = p(w_0) \prod_{w_1, w_2 \in \mathcal{D}} p(w_2 | w_1)$$

where w_0 is the first word, and (w_1, w_2) is a pair of consecutive words in the document. In this case, the number of parameters is $(N-1) + (N^2 - N) = N^2 - 1$. However, with the model as written above, there are only parameters $N^2 - N$.

The grader is now fixed to accept both as correct and regrading is happening.

N^2 - 1 ✔ **Answer:** N^2 - 1

STANDARD NOTATION

Solution:

Recall the likelihood of D in bigram model is (though this is not what written):

$$p(\mathcal{D}; \theta) = p(w_0) \prod_{w_1, w_2 \in \mathcal{D}} p(w_2 | w_1)$$

where w_0 is the first word, and (w_1, w_2) is a pair of consecutive words in the document.
Denote the set of all N words by V. The set of parameters is

$$\{p(w_0) : w_0 \in V\} \cup \{p(w_1 | w_2) : w_1 \in V, w_2 \in V\}$$

and the only constraints on these parameters are

$$\begin{aligned} \sum_{w_0 \in V} p(w_0) &= 1 \\ \sum_{w_1 \in V} p(w_1 | w_2) &= 1 \quad \text{for all } w_2 \in V. \end{aligned}$$

Hence, the number of parameters is $(N-1) + (N^2 - N) = N^2 - 1$. (Note that this is also the number of parameters $p(w_1, w_2)$ where $w_1 \in V, w_2 \in V$, which determine the joint distribution.

Solution to the problem as written:
The likelihood of D in bigram model was given as

$$p(\mathcal{D};\theta) = \prod_{w_1,w_2\in\mathcal{D}} p(w_2|w_1)$$

without taking into account the likelihood $p(w_0)$ of the first word. In this case, the parameters are

$$\{p(w_1|w_2) : w_1 \in V, w_2 \in V\}$$

where $\sum_{w_1\in V} p(w_1|w_2) = 1$ for all $w_2 \in V$. Hence, the number of parameters is $N^2 - N$.

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

Bigram Model 2

1/1 point (graded)
Which of the following represents the MLE for the **conditional probability** $p(w_2 \mid w_1)$?

- $\frac{\text{count}(w_1,w_2)}{\sum_{w_1',w_2'\in\mathcal{D}}\text{count}(w_1',w_2')}$
- $\frac{\text{count}(w_1,w_2)}{\sum_{w_1',w_2\in\mathcal{D}}\text{count}(w_1',w_2)}$
- $\frac{\text{count}(w_1,w_2)}{\sum_{w_1,w_2'\in\mathcal{D}}\text{count}(w_1,w_2')}$ ✔
- $\frac{\sum_{w_1',w_2\in\mathcal{D}}\text{count}(w_1',w_2)}{\sum_{w_1,w_2'\in\mathcal{D}}\text{count}(w_1,w_2')}$

Solution:

This is a simple application of Bayes Rule:

$$p(w_2|w_1) = \frac{p(w_1,w_2)}{p(w_1)}$$

To compute $p(w_1)$, we marginalize out w_2.

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

Bigram Model 3

1/1 point (graded)
Consider the same sequence from the unigram model:

A B A B B C A B A A B C A C

If you estimate θ on this, what probability will be assigned to the following test sequence? Assume the starting probabilities of all characters $p(w|\text{null})$ is uniform. Give your answer to three decimal places.

A A B C B A B

0

Answer: 0

Solution:

There is no need to compute the actual probability. Since the transition $C \to B$ does not appear in $\mathcal{D}$, the probability assigned to this new sequence will be 0. This is why techniques like smoothing are important in practice for small datasets.

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

Discussion

Show Discussion

Topic: Unit 4 Unsupervised Learning (2 weeks) :Homework 5 / 2. Maximum Likelihood Estimation