CS 107, Probability, Spring 2020
Lecture 16

Michael Poghosyan
mpoghosyan@aua.am

AUA

24 February 2020

Content

- Repeated, Independent Trials: Multinomial Distribution
- Applications of Conditional Probability: Simple Recommender System
- Applications of Conditional Probability: Language Models (*n*-gram Models)

Last Lecture ReCap

Last time we were talking about Repeated Trials Model, in particular, about Multinomial Model:

- Assume we have a Simple Experiment (Trial), and several disjoint Events in that Experiment, say, $A_1, \dots, A_m$. We assume that in one Trial $\mathbb{P}(A_k) = p_k$. We repeat our Trial n times, independently, and let X_k be the number of times A_k will appear in that n Trials. Then

$$\mathbb{P}(X_1 = k_1, X_2 = k_2, \dots, X_m = k_m) = \binom{n}{k_1, k_2, \dots, k_m} \cdot p_1^{k_1} \cdot p_2^{k_2} \cdot \ldots \cdot p_m^{k_m},$$

where

$$\binom{n}{k_1, k_2, \dots, k_m} = \frac{n!}{k_1! \cdot k_2! \cdot \ldots \cdot k_m!}.$$

Example:

Problem: Assume that my waiting time for the Metro Train at the Baghramyan Station is a uniform random number from $[0, 10]$ (in minutes). What is the probability that for 10 days, 3 times I will wait for the train for more than 8 min, for 5 days my waiting time will be between 3 and 8, and for 2 days, I will wait no more than 3 min?

Solution: Multinomial Case!

- Simple Experiment: Considering the waiting time for one day, $\Omega_{Simple} = [0, 10]$;
- Complementary Simple Events: $A_1 = (8, 10]$, $A_2 = (3, 8]$, $A_3 = [0, 3]$, so $A_1 \cup A_2 \cup A_3 = \Omega_{Simple}$;
- The probabilities of Simple Events: $p_1 = \mathbb{P}(A_1) = \frac{2}{10} = \frac{1}{5}$, $p_2 = \mathbb{P}(A_2) = \frac{5}{10} = \frac{1}{2}$, $p_3 = \mathbb{P}(A_3) = \frac{3}{10}$, and

$$p_1 + p_2 + p_3 = 1$$

Example:

Problem: Assume that my waiting time for the Metro Train at the Baghramyan Station is a uniform random number from $[0, 10]$ (in minutes). What is the probability that for 10 days, 3 times I will wait for the train for more than 8 min, for 5 days my waiting time will be between 3 and 8, and for 2 days, I will wait no more than 3 min?

- We repeat the Simple Experiment $n = 10$ times, the Trials are independent;
- We are interested in the Probability of having exactly $k_1 = 3$ times A_1, exactly $k_2 = 5$ times A_2 and exactly $k_3 = 2$ times A_3;
- And we have $k_1 + k_2 + k_3 = n$
- Rest: OTB

Example:

Problem: Assume we have 132 balls in a box, 88 white balls, 26 blue balls and 18 green ones. 15 times we are choosing a ball at random from the box, with replacement. What is the probability that

- Exactly 10 times we will have white balls? OTB
- 7 times we will have white ones, and 5 times we will have green ones? OTB
- We will not have any green balls shown? OTB

Some Applications of the Conditional Probabilities

Simple Recommender System

Example: Simple Recommender System

Hopefully, you know what is a recommender (or recommendation) system. Let us consider the following problem:

Problem: (Movie Recommendation) Assume that the person watched and liked the movie A. Which movies to recommend him/her for watching?

Assumption: We can use only the database of users, where we have the list of movies watched and liked/disliked by users.

Basic Idea: we will recommend the movie B with the highest probability:

$$\mathbb{P}(\text{like B}|\text{like A}).$$

Now, it remains to calculate this conditional probability.

Example: Simple Recommender System, Cont'd

Now, how to calculate the probabilities? -

$$\mathbb{P}(\text{like B}|\text{like A}) = \frac{\#\text{ persons who liked A that liked B}}{\#\text{ persons who liked A}}$$

Explanation: for $\mathbb{P}(\text{like B}|\text{like A})$, our "universe", new Sample Space consists of all persons who liked A. Among them we want to measure the probability of liking B.

Example: Simple Recommender System, Cont'd

Some ideas for improvement:

- do several recommendations (say, recommend 5 movies)
- sometimes randomly recommend movies that do not have high conditional probability, even with 0 probability (say, from the set of persons who watched A, nobody watched C - that doesn't mean that our person who watched and liked A will dislike C)
- Take into consideration the probability $\mathbb{P}(\text{like A})$ - if this is small, then the recommendation can be not effective (think about the diagnosis case!). So put a threshold on this probability, make predictions if $\mathbb{P}(\text{like A})$ is larger than some fixed value

Extra: Some Applications

Some Applications of the Conditional Probabilities

Simple Language Modeling

Example: Simple Language Modeling

Hopefully, you know what is a Language $\ddot\smile$ Well, one of the challenges in the computer science is the language modeling, which is used in speech recognition, machine translation, part-of-speech tagging, parsing, Optical Character Recognition, handwriting recognition, information retrieval and other applications (from Wiki).

Say, when we ask Google some question, say, ”what is the age of the universe”, it gives `https://www.google.com/search?&q=what+is+the+age+of+the+universe`

Example: Simple Language Modeling

In language modeling, one is modeling the probabilities of the sentences to happen (in the totality of all possible sentences). Say, for example, we can have (some not-so-realistic, fictitious numbers):

$\mathbb{P}(\text{The art and science of asking questions is the source of all knowledge.}) = 10^{-23}$

and

$$\mathbb{P}(\text{I love teaching and doing Mathematics.}) = 10^{-32}$$

or

$$\mathbb{P}(\text{She loves you yeah yeah yeah}) = 10^{-25}$$

But what about

$$\mathbb{P}(\text{you yeah loves yeah she yeah}) = ?$$

Example: Simple Language Modeling

or

$$\mathbb{P}(\text{doing love and teaching Mathematics I}) = ?$$

or

$\mathbb{P}$(questions the and of of asking the is science source all art knowledge.) $= ?$

Example: Simple Language Modeling

To give the simple model, we assume that $w_1, w_2, ..., w_N$ are all words, so

$$V = \{w_1, w_2, ..., w_N\}$$

is our vocabulary. Now, each sentence S is an ordered set of words, say,

$$S = w_1, w_2, ..., w_n,$$

where n is the word-length of the sentence. And the Probability of having the sentence S is

$$\mathbb{P}(S) = \mathbb{P}(w_1, w_2, ..., w_n).$$

Here the order matters, so we calculate the Probability of having the first word w_1, then followed by w_2 and so on.

Example: Simple Language Modeling

Now, using the Conditional Probabilities, we can write

$$\mathbb{P}(S) = \mathbb{P}(w_1, w_2, ..., w_n) =$$

$$= \mathbb{P}(w_1) \cdot \mathbb{P}(w_2|w_1) \cdot \mathbb{P}(w_3|w_1, w_2) \cdot ... \cdot \mathbb{P}(w_n|w_1, w_2, ..., w_{n-1}).$$

Calculation of Conditional Probabilities is computationally expensive, so one simplifies the model in different ways:

- **Model 1: Unigram Model** - we assume that the words are independent. In this case

  $$\mathbb{P}(S) = \mathbb{P}(w_1, w_2, ..., w_n) = \mathbb{P}(w_1) \cdot \mathbb{P}(w_2) \cdot ... \cdot \mathbb{P}(w_n).$$

  Here the Probabilities can be calculated by (if we have a large corpus, large dataset of sentences):

  $$\mathbb{P}(w_k) = \frac{\#\text{Sentences containing } w_k}{\#\text{All Sentences}}.$$

Example: Simple Language Modeling

- **Model 2: Bi-gram Model** - we assume that each words depends only on the preceding one, the previous word. In this case
$$\mathbb{P}(S) = \mathbb{P}(w_1, w_2, ..., w_n) =$$
$$= \mathbb{P}(w_1) \cdot \mathbb{P}(w_2|w_1) \cdot \mathbb{P}(w_3|w_2) \cdot ... \cdot \mathbb{P}(w_n|w_{n-1}).$$
Here the Probabilities can be calculated easily. We calculate
$$\mathbb{P}(w_1) = \frac{\#\text{Sentences starting with } w_1}{\#\text{All Sentences}}.$$
and
$$\mathbb{P}(w_2|w_1) = \frac{\#\text{Sentences containig } w_1 \text{ followed by } w_2}{\#\text{Sentences containig } w_1}.$$

Example: Simple Language Modeling

Now, having this model, and, say, having an incomplete sentence $S = w_1, w_2, ..., w_k$, you can predict the next word w_{k+1}! Just choose the one with the highest conditional probability $\mathbb{P}(w_{k+1}|w_k)$.
Surely you are familiar with Google Mail's sentence completion feature $\ddot\smile$ No miracle here, you can do it now as a Conditional Probability homework exercise $\ddot\smile$

Of course, I am oversimplifying things $\ddot\smile$

And you can consider 3-gram, n-gram models too!

Example: Simple Language Modeling

By the way, in the Bi-gram Model, we assume that each words depends only on the preceding one, the previous word. Mathematically, we assume that

$$\mathbb{P}(w_k|w_1, w_2, ..., w_{k-1}) = \mathbb{P}(w_k|w_{k-1}),$$

and this is the celebrated **Markov Chain Model**, which we will cover a little bit at the end of our course!