Estimation of quantities related to the multinomial distribution with unknown number of categories

Marcin Kuropatwiński

Introduction

In this white paper we aim to demonstrate the usefulness and simplicity of a set of formulas for estimation of quantities related to the *multinomial distribution with an unknown number of categories*. We will not give a derivation of the formulas, which is covered in our patent application [1], but instead we will discuss the main formulas and show how they can be applied in practice.

To start we provide a definition of the *multinomial distribution*. The multinomial distribution is defined by a *probability mass function* (pmf) over a set of *categories*, which is a finite, countable set. We will denote the cardinality of the set of categories S by K. We will denote by p_i the probability (which we call *proportions*) of the category $i \in S = \{1,\ldots,K\}$. We note that probabilities are larger than zero and sum to unity. We will call the *experiment* the process of drawing an element from the set S of categories randomly according to the probability mass function $\{p_i\}_{i \in S}$. In the typical setup known from the literature K is known. However, in some applications we do not know the number of categories. An example of such an application is speech or speaker recognition with discrete models. In such cases we observe some symbols (categories) but we are unaware of how many distinct categories we could observe if the number of experiments approaches infinity. This poses a serious obstacle as known estimators for p_i make use of known K. To tackle the indicated obstacle, we treat K as a hidden random variable and estimate it along with the probability mass function. This involves solving certain combinatorial problems, which, fortunately, have closed-form analytical solutions.

Known estimators for multinomial proportions

Let us review the two known methods of estimation of the multinomial proportions. We will assume that the reader is familiar with the concepts of *maximum likelihood* (ML) and *minimum mean square error* (MMSE) estimation [2].

We estimate K from samples (training sets). Assume that we performed M experiments. We will denote the results of experiments by $o_j \in S$, $j \in \{1,\ldots,M\}$. Further the results of the experiments will be called *observations* and by O we denote the sequence of observations. We also need define the counts of the observation. The count P_i is the number of occurrences of the category i in the sequence of observations. Formally, it is equal to $P_i = |\{o_j : o_j = i\}|$, where $|A|$ denotes cardinality of the set A. Two prevalent methods of estimation of the multinomial proportions are the ML and MMSE methods.

The ML estimator leads to the formula:

$$p_i^{\mathrm{ML}} = \frac{P_i}{M}. \tag{1}$$

The disadvantage of this equation is that it assigns zero probability to unseen categories so it is not used in applications where we know that we have not seen all categories in the training set.

The MMSE method is an alternative to the ML method. It leads to the formula:

$$p_i^{\mathrm{MMSE}} = \frac{P_i + 1}{M + K}. \tag{2}$$

This formula was first introduced by Pierre-Simon Laplace so it is called also the Laplace estimator.

The MMSE method assigns probability equal to $\frac{1}{M+K}$ for all unseen categories and thus it does better than the previous formula in practical application. However it makes use of K, which is usually not known. Thus we spent some efforts to find methods of estimating K.

Estimation of the number of categories

In the course of the work on the patent, we found ML and MMSE estimators for K, the number of categories. We also found a general formula for $p(K \mid O)$, which is quite complex and, therefore, not shown here (it can be found in the patent application). The estimators make use of quantity Z, which we call the *diversity index*. The diversity index Z is equal to $|\,\mathrm{UNIQUE}(O)\,|$, where $\mathrm{UNIQUE}(A)$ is the set of unique elements in the multiset A.

We divide the presentation into two subcases. The first case assumes that all proportions are equal, so we deal with the uniform distribution. The second case relaxes the assumption of equal proportions.

Uniform distribution

The conditional probability function of K given O can take three distinct forms. In the first form it is monotonically increasing. That means that the maximum of this likelihood function is attained at infinity. This form is associated with our ignorance about true K. A condition for this form is that the diversity index equals the number of experiments. In other words, as long as there are as many distinct observed categories as experiments we can say nothing about true K.

For the second form, the likelihood function has single maximum in the range $[Z,\infty)$. This happens when the diversity index is less than the number of experiments. It is convenient to introduce another quantity, which plays an important role in our theory. This quantity is the *generalization coefficient* N, which is equal by definition:

$$N \equiv \frac{M}{Z} \tag{3}$$

The ML estimate of K can then be obtained by solving the following equation:

$$\frac{1}{v}\ln\left(\frac{1}{1-v}\right) = N\,, \tag{4}$$

where:

$$v = \frac{K_{\mathrm{ML}}}{Z}\,, \tag{5}$$

is the fraction of the saturated/learned support.

The last form is when the likelihood is monotonically decreasing. The condition for this form is following:

$$M > \log_{\frac{Z+1}{Z}}(Z+1)\,. \tag{6}$$

If this condition holds, the ML estimate for K is equal Z.

In our patent application a figure shows how large M must be to learn the given fraction of categories.

This concludes presentation of dependencies for the uniform distribution. Next, we discuss the case of the non-uniform distribution, which is more relevant for real-world scenarios.

Non-uniform distribution

In this section we present the ML and MMSE estimators for K for the case of a non-uniform distribution. As in the case of uniform distribution, the likelihood function of K given O can take three forms:

- it is monotonically increasing if and only if:

$$M = Z\,. \tag{7}$$

- it is monotonically decreasing if and only if:

$$M > Z^2\,. \tag{8}$$

- it has single maximum for K in the range $[Z,\infty)$ if and only if none of the above conditions hold.

The case with a single-maximum form is associated with the ML estimator for K. This estimator is very simple:

$$K_{\mathrm{ML}} = \frac{ZN}{N-1}\,. \tag{9}$$

Another estimator is the MMSE estimator:

$$K_{\mathrm{MMSE}} = \mathrm{Mean}[K] = \mathrm{E}[K \mid O] = \frac{Z(M-1)}{M-Z-2} = \frac{ZN-1}{N-\frac{2N}{M}-1} \tag{10}$$

As can be seen the formula is reasonable if $M > Z+2$. We have chosen to present the variance of the this MMSE estimator, which is equal to:

$$\mathrm{Var}[K] = Z(Z+1)\left[\frac{\Gamma(M-1)\Gamma(M)\,{}_3F_2(2,Z+1,Z+2;1,M+Z+1;1)}{\Gamma(M-Z-1)\Gamma(M+Z+1)} - \frac{Z(Z+1)}{(M-Z-2)^2}\right]. \tag{11}$$

This formula does not have the simplicity of the solutions shown thus-far. However, the formula is easily computable using modern mathematical software (like *Mathematica*). In the above formula Γ is the gamma function and F is the generalized hypergeometric function defined in terms of infinite power series.

The sunrise problem

The sunrise problem can be expressed as follows:

"What is the probability that the sun will rise tomorrow?"

This problem has been first formulated by Laplace and it illustrates the difficulty of probability theory to deal with such questions. Suppose that we want to compute the probability of the sunrise based on the history of this event. We may use the Laplace estimator given by formula (2). If we allow the possibility that the sun will not rise tomorrow our K is equal two. This causes that the probability of the sun not rising tomorrow is non-zero and is inversely proportional to the number of sunrise events in the known history. We would expect a probability that is zero or almost zero but the Laplace estimator returns probability, which is relatively large. The estimators, which are considered to solve the sunrise problem correctly, return in this case a probability, which is inverse proportional to the *square* of the number of sunrise events in the history [3]. In the following section we propose such estimator.

Proposed estimators for multinomial proportions

In this section we present two flavors of multinomial proportions estimators. One is a single formula solution, which, however, does not solve the sunrise problem properly and a second estimator that involves a piecewise function (which poses a minor complication) but solves the sunrise problem correctly.

The first estimator results from computing the expectation of the Laplace estimator with respect to the unknown K. We will call this estimator MMSE1 estimator. It can be expressed as:

$$p_i^{\mathrm{MMSE1}} = \frac{(P_i + 1)(M - Z - 1)}{M(M-1)}. \tag{12}$$

The advantage of this estimator over the Laplace estimator is that is accounts implicitly for the unknown K. In other words, it only depends on the observed quantities: counts, diversity index and the number of experiments.

The second estimator we call the MMSE2 estimator. We first define two sets of categories. The first set, denoted B, is the set of categories that have been observed in the training set. The second set, denoted A, is the set of unobserved categories. We now compute the probabilities of those sets:

$$\Pr(A) = (K_{\text{MMSE}} - Z)\frac{M - Z - 1}{M(M-1)} = \frac{(M - Z - 1)(Z-1)Z}{(M - Z - 2)(M-1)M} \approx \frac{1}{N^2} \tag{13}$$

and

$$\Pr(B) = 1 - \Pr(A)\,. \tag{14}$$

Now we can compute probabilities $\Pr(i \mid A)$ and $\Pr(i \mid B)$. First, the $\Pr(i \mid A)$ is equal:

$$\Pr(i \mid A) = \begin{cases} \frac{1}{K_{\text{MMSE}} - Z} & \text{if} \quad \frac{1}{K_{\text{MMSE}} - Z} \leq 1 \\ 1 & \text{otherwise} \end{cases}, \tag{15}$$

where we assume that all categories in the set A are equally probable and there is at least one such category (otherwise the probability $\Pr(i \mid A)$ could be greater than one, which is nonsense). Second, we provide the equation for $\Pr(i \mid B)$:

$$\Pr(i \mid B) = \frac{P_i + 1}{M + Z}\,. \tag{16}$$

Now we are in the position to formulate the MMSE2 estimator:

$$p_i^{\text{MMSE2}} = \begin{cases} \Pr(i \mid A)\Pr(A) & \text{if} \quad i \in A \\ \Pr(i \mid B)\Pr(B) & \text{otherwise} \end{cases}. \tag{17}$$

It can be shown that the estimator solves correctly the sunrise problem in the sense the probability of unseen categories goes to zero with the square of the number of observations.

Applications

We believe that the presented theory has many applications in data analysis, many of which we likely cannot imagine yet. Apart from predicting the number of categories, and estimating the multinomial proportions, the theory can be used obviously for:

- predicting the number of observations needed to learn a given percent of the total mass of the multinomial distribution (that means we learn such a set of categories that sum of the proportions over the set is equal to some chosen fraction). To do that we compute the MMSE estimate for K and substitute it for Z to the equation for $\Pr(B)$. Then we equate the resulting expression to the desired percent of mass and solve for M;
- low-computational-cost estimation of entropy or upper bounds on entropy with $O(M)$ computational complexity and $O(K)$ worst case memory complexity;

- adjusting complexity of the discrete models – this is described in detail in our patent application;
- testing random variables for statistical independence;
- estimation of the number of extrema of multimodal functionals.

Conclusions

We derived a set of simple but non-trivial formulas that can be used to analyze the multinomial distribution in a new way. The advantage of the proposed methods is that they do not assume a known number of categories and are distribution-free (depend only on the diversity index, counts and the number of experiments) and thus can be used in practical applications like speech or speaker recognition.

References

[1] M. Kuropatwiński, "Method For Adjusting Discrete Model Complexity in an Automatic Speech Recognition System," USA Patent, 2012.

[2] W. Feller, *An introduction to probability theory and its applications*, 1957.

[3] E. S. Ristad, "A Natural Law of Succession," Princeton1995.