Statistische Methoden in der Computerlinguistik

3 Combinatorics

Combinatorics is the study of *combinations* of elements – in particular, the number(s) of ways of building lists (sets) of length (size) k out of a set of n possible elements. We have four cases to consider, depending on:

1. whether or not we care in which order items are selected, and

2. whether or not we can re-use previously selected items.

	−Reuse	**+Reuse**
+Order	$(n)_k$	n^k
−Order	$\binom{n}{k}$	$\binom{n+k-1}{k}$

Where:

$$\begin{aligned}
n^k &= \overbrace{n \cdot n \cdot \ldots \cdot n}^{k \text{ factors}} \\
n! &= n \cdot (n-1) \cdot \ldots \cdot 1 \\
(n)_k &= n \cdot (n-1) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!} \\
\binom{n}{k} &= \frac{(n)_k}{k!} = \frac{n!}{(n-k)!k!}
\end{aligned}$$

For proofs, see Krenn and Samuelsson (1997, Section 1.2.7).

Example 1 (Lotto) For a lottery game in which 6 balls are selected from a pool of balls uniquely numbered from 1 to 49, where the balls are not replaced after having been drawn and where order does not matter, there are $\binom{49}{6} = 13,983,816$ possible results.

4 Yet More Probability Theory

4.1 Some Common Probability Distributions

Idea:

- In practical applications, we often have no prior knowledge of the probability distribution(s) for the events we wish to model.
- Although the *relative frequency* $\frac{n(A)}{N}$ can serve as an estimate for $P(A)$, such estimates become quite cumbersome when a large sample space (i.e. natural language data) is involved.
- Instead, we can often use prior knowledge of the domain we wish to model to assign our phenomena to one of a number of common *families of distributions*, and formulating our model in terms of the general properties ("parameters") of the distribution family.
- *Caveat:* such a technique can go horribly, horribly wrong if we choose the wrong underlying distribution, but it's pretty darned handy nonetheless.

4.1.1 Bernoulli Distribution

A stochastic experiment with exactly two possible outcomes is called a *Bernoulli trial*, and is described by a parametric distribution known as a *Bernoulli distribution*, which is completely specified by the probability p of one of the two possible outcomes. Random variables for Bernoulli trials are typically mapped to the values 0 ("failed trial") and 1 ("successful trial").

For a random variable X and a probability p, we write $X \sim b(1,p)$ to indicate that X has a Bernoulli distribution with $p_X(1) = p$. For such a random variable, the following properties hold:

$$\begin{array}{rrcrcl}
\textbf{Success Probability:} & b(1;1,p) & = & p_X(1) & = & p \\
\textbf{Failure Probability:} & b(0;1,p) & = & p_X(0) & = & 1-p \\
\textbf{Mean:} & E(b(1,p)) & = & E(X) & = & p \\
\textbf{Variance:} & Var(b(1,p)) & = & Var(X) & = & p \cdot (1-p)
\end{array}$$

Example 2 (Fair Coin) A random variable $X = \{heads \mapsto 1, tails \mapsto 0\}$ describing a single coin toss has a Bernoulli distribution:

$$X \sim b(1, P(heads))$$

Example 3 (Lotto) Let X be a random variable describing a lottery game as in Example 1, where for a 6-number ticket $t = \{x_1, x_2, x_3, x_4, x_5, x_6\}$, let $X(t) = 1$ iff all of the numbers x_i are drawn, and let $X(t) = 0$ otherwise. If the lottery game is fair, then X has a Bernoulli distribution with probability $\frac{1}{13{,}983{,}816} \approx 7.15 \times 10^{-8}$:

$$X \sim b\,(1, 0.0000000715)$$

4.1.2 Binomial Distribution

A finite series of Bernoulli trials in which each trial is independent from all other trials is described by a *Binomial distribution*, and is completely specified by the number n of trials in the series and the probability p of success in any trial.

For a random variable X, a natural number n, and a probability p, we write $X \sim b(n,p)$ to indicate that X has a binomial distribution over n independent trials with $p_X(1) = p$ in any of those trials, and for a natural number $k \leq n$, we write $b(k;n,p)$ to indicate the probability of k successes in n such trials – but how do we find it?

- Given an $X \sim b(n,p)$ and a $k \leq n$, the probability of any specific (ordered) sequence containing k successes and $n-k$ failures is $p^k(1-p)^{n-k}$, since the individual trials are independent.

- We are interested only in the number k of successful trials, ignoring the order in which they occur, so we must count every sequence containing k successes as a single event.

- There are exactly $\binom{n}{k}$ such sequences – this is just the number of ways of picking k successful trial indices from a "bag" of n trial indices *without* regard to order and *without* re-using previously selected indices.

Therefore, for a random variable $X \sim b(n,p)$, the following properties hold:

$$\begin{array}{rrcl}
\textbf{k-Success:} & b(k;n,p) & = & \binom{n}{k} p^k (1-p)^{n-k} \\[2ex]
\textbf{Mean:} & E(X \sim b(n,p)) & = & \sum\limits_{k=0}^{n} k \cdot b(k;n,p) \\
 & & = & np \\[2ex]
\textbf{Variance:} & Var(X \sim b(n,p)) & = & \sum\limits_{k=0}^{n} (k-np)^2 \cdot b(k;n,p) \\
 & & = & np(1-p)
\end{array}$$

Example 4 (2 Fair Coins) Let X be a random variable modelling the toss of 2 fair coins which maps each experiment to the number of coins which come up "heads". Then, X has a binomial distribution:

$$\begin{aligned} X &\sim b(2, 0.5) \\ \\ b(0; 2, 0.5) &= 1 \cdot 0.5^0 \cdot (1-0.5)^{2-0} = 0.25 \\ b(1; 2, 0.5) &= 1 \cdot 0.5^1 \cdot (1-0.5)^{2-1} = 0.5 \\ b(2; 2, 0.5) &= 1 \cdot 0.5^2 \cdot (1-0.5)^{2-2} = 0.25 \\ \\ E(X) &= 2 \cdot 0.5 = 1 \\ Var(X) &= 2 \cdot 0.5 \cdot (1-0.5) = 0.5 \end{aligned}$$

Example 5 (Lotto) Let X be a random variable describing a sequence of fair lottery drawings as in Example 1 whose values are the number of full (6-position) matches ("winning tickets") in the sequence of lottery drawings. Suppose that Lotto drawings occur weekly and that X describes Lotto participation once a week since the birth of Socrates (ca. 470 BC), or roughly 128648 consecutive weeks. Then, X has a binomial distribution:

$$\begin{aligned} X &\sim b(n = 128648, p \approx 7.15 \times 10^{-8}) \\ \\ b(1; n, p) = E(X) &\approx 128648 \cdot 7.15 \times 10^{-8} \approx 0.0092 \end{aligned}$$

4.2 Parameter Estimation

Idea:

- Often, we can't observe probabilities $P(X)$ directly, but we can perform a series of experiments and tally the observed values $\vec{x}$ of a random variable X.
- If we can additionally sort our phenomena into a parameterized model, we can express the distribution function for X as conditioned on the (unknown) parameters $\theta \in \Xi$ of that model: $P(x) = P(x|\theta)$. For a binomial distribution $X \sim b(n, p)$, $\theta = p$.
- Given an observed set of values and an expression of their dependance on one or more model parameters, we can attempt to find the model parameters which best account for the observed data.
- Having thus estimated a model's parameters, we can effectively fix the (imperfect) model, using it to predict the outcomes of further experiments.

4.2.1 Maximum Likelihood Estimation

Definition 1 (Likelihood Function) *A likelihood function for parameters $\theta \in \Xi$ and an observed sample $X = x$ of a random variable X is a function $L : \Xi \to$*

$\mathbb{R}$ *which can be expressed as a (linear function of a) conditional probability function*[1] *with its first argument held fixed to the observation x:*

$$L(\theta) = aP(X = x|\theta) + b \qquad a, b \in \mathbb{R}$$

Example 6 (Likelihood: Coin Toss) A coin is tossed 10 times, resulting in 8 heads and 2 tails. By prior knowledge, we assume that the trials are independent, and thus a random variable X modelling the process has a binomial distribution: $X \sim b(10, p)$. The model's parameters θ are given by p, so $\Xi = [0, 1]$ and:

$$L(p) = \binom{10}{8} p^8 (1-p)^2$$

Definition 2 (Maximum Likelihood) *Given a likelihood function $L : \Xi \to [0, 1]$ the parameters $\theta \in \Xi$ which maximize $L(\theta)$ are called the* maximum likelihood estimate *for θ:*

$$MLE(\theta) = \arg\max_{\theta} L(\theta)$$

Example 7 (MLE: Coin Toss) Continuing Example 6, we find (unsurprisingly) that:

$$MLE(p) = \arg\max_{p} \binom{10}{8} p^8 (1-p)^2 = 0.8$$

4.2.2 Bayesian Statistics

Parameter estimation in Bayesian statistics takes into account *prior belief* in the acceptability of model parameters θ. Formally, this involves an additional probability distribution on Ξ and the application of Bayes' law in order to revise prior beliefs in the face of new evidence, thus computing a *posterior distribution*:

$$P(\theta|x) = \frac{P(\theta)P(x|\theta)}{P(x)}$$

Although the prior probability $P(x)$ of x is unknown, we can ignore it if we just want to find the parameters which maximize the posterior distribution – that is, if we want to compute a new "best" value for θ given x:

$$MAP(\theta) = \arg\max_{\theta} P(\theta|x)$$

Given our (possibly new) parameter estimate θ, we can turn our attention to new data. Iteration of this process is known as *Bayesian updating*.

[1] Despite the formulation of likelihood functions in terms of probabilities, likelihood functions are not in general probability functions themselves.

Example 8 (Bayesian Updating: Coin Toss) Consider the coin toss experiment from Example 6. Suppose our prior belief is modelled by by the function $G : \mathbb{R} \rightarrow \mathbb{R}$, where:

$$G(p) = \begin{cases} 6p(1-p) & \text{if } 0 \leq p \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Then for the observed sequence of 8 heads and 2 tails, we have:

$$\begin{aligned} P(p|x) &= \frac{P(p)P(x|p)}{P(x)} \\ &= \frac{p^8(1-p)^2 \cdot 6p(1-p)}{P(x)} \\ &= \frac{6p^9(1-p)^3}{P(x)} \end{aligned}$$

so that:

$$\begin{aligned} MAP(p) &= \arg\max_p P(p|x) \\ &= \arg\max_p \frac{6p^9(1-p)^3}{P(x)} \\ &= 0.75 \end{aligned}$$

References

B. Krenn and C. Samuelsson. *The Linguist's Guide to Statistics*. URL `http://www.coli.uni-sb.de/~krenn/stat_nlp.ps.gz`. Unpublished manuscript, 1997.