Statistical Methods in Natural Language Processing (NLP)

Class 2: Probability Theory

Charalambos (Haris) Themistocleous

Department of Philosophy, Linguistics and Theory of Science, Centre for Linguistic Theory and Studies in Probability

Previous Class

1. Introduction to the course.
2. History of using Statistics and Probability Theory in Computational Linguistics.
3. Python (and optionally R).
4. Combinatorics and Permutations.

Note: If you have laptops please have them with you to run the code we will be discussion in class.

Let us start with a problem

In English, 5 letters (A, E, I, O, U) represent vowels and 21 letters (B, C, D, F, G, H, J, K, L, M, N, P, Q, R, S, T, V, W, X, Y and Z) represent consonants (Y sometimes is used as a vowel: compare *yes* vs. *cry*, but let ignore this for now.).

How many possible **CVCV** words (where C = Consonant, V= Vowel) are possible with these letters?

Answer

$$21 \times 5 \times 21 \times 5 = 11025$$

Another Problem: Task

How many **CVCV** words are possible, if vowels and consonants can be used only once?

Permutations

How many possible arrangements are possible with the letters c, a, t of the word *cat*?

Answer

To find how many possible arrangements are possible with the three letters c, a, t then there are 6 possible permutations: $1 \times 2 \times 3 = 6$.

In general to determine the number of permutations in set of n things that does not include repetitions, then this number is $n!$, i.e.,

$$n(n1)(n2)3 \times 2 \times 1 = n!$$

The $n!$ is also know as factorial and it is equal to the number of rearrangements of n things.

Problem

How many different arrangements or *permutations* are possible for the letters of the word *letters*?

Solution

there are 7! possible combinations for the word *letters*. However, note that in this word there are 2Es and 2Ts $(2! \times 2!)$. So, the result is $7!/2! \times 2! = 1260$.

Overall, when there are cases with repetitions we employ the following:

$$\frac{n!}{n_1!n_2!\dots n_k!}$$

Python function for factorials

The Python function for calculating the factorial is factorial(). You can find the factorial function in numpy.math.factorial:

```
import numpy as np
```

so to find 7! in Python:

```
>>>> np.math.factorial(7)
5040
```

R function for factorials

to find 7! in R:

```
factorial(7)
[1] 5040
```

Combinations

There can be n ways to select an item from S; to select a second item there are $n-1$ ways, to select a third item there are $n-2$ ways etc. So, to specify the number of different sets of k items that could be formed from a total of n items, we apply the following:

$$\frac{n(n1)\dots(nk+1)}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

when $0 \leq k \leq n$, and let it equal 0 otherwise.

Example

Problem: How many 3 letter words could be formed from the 5 letters: i, n, a, m, and e.

Solution

To this purpose, we need to work as follows.

- There are 5 ways to choose the first letter; 4 ways to chose the second letter, and 3 ways to select the final letter. Therefore, there are 5 × 4 × 3 ways of selecting the group of 3 letters. Note that in this case there no constraints for the order of these letters.
- So, to get all the possible combinations of 3 letters in a group, then each group of three letters should be counted 6 times (i.e., 1 × 2 × 3).

So the total number of 3 letters that could be created is:

$$\frac{5 \times 4 \times 3}{3 \times 2 \times 1} = 10$$

The result provided by 12 is $\binom{n}{k}$, that is the $\binom{n}{k}$ is the number of possible combinations of n items selected k at a time.

Another example

Example. 4 students should be selected from a class of 15 students. How may different groups of students are possible?

Solution There are 1365 combinations calculated as follows:

$$\binom{n}{k} = \frac{15 \times 14 \times 13 \times 12}{4 \times 3 \times 2 \times 1} = 1365$$

Example

Example. There are 20 consonants and 6 vowels in a language, how many different letter combinations are possible consisting of 3 vowels and 4 consonants.

Solution There are $\binom{20}{4}$ combinations of 4 consonants and $\binom{6}{3}$ combinations of 3 vowels. Therefore, by applying the fundamental principle there are $\binom{20}{4} \times \binom{6}{3} = \binom{20\times19\times18\times17}{4\times3\times2\times1} \times \binom{6\times5\times4}{3\times2\times1} = 96900$.

Python function

```
import operator as op
def choose(n, r):
    r = min(r, n-r)
    if r == 0: return 1
    numer = reduce(op.mul, xrange(n, n-r, -1))
    denom = reduce(op.mul, xrange(1, r+1))
    return numer//denom
```

Let us test this

```
>>> print(choose(5,3))
10
```

R Function for choose

```
choose(5,3)
[1] 10
```

There also the R function combn(), which enumerates all possible combinations: try this combn(5,3).

Set Theory: Basics

There are many phenomena that concern probabilities.

- A **set** is a collection of distinct objects.
- **Empty sets** are sets that contain nothing and are denoted with the symbol $\emptyset$.

Sample Spaces

The set of all possible outcomes of the experiment is the *sample space* and it is usually denoted using the set notation as S or Ω (the universal sets are indicated with U).

Think of an experiment, that its result cannot be predicted with certainty, yet might know that the set of all *possible* outcomes is given. For example, if this experiments is the tossing of a coin, then if we toss coin once, then the outcome will be a head or a tail:

$$S = \{head, tail\}$$

If the experiment involves the tossing of a dice, then in this case the typical sample space is the number of pips that face up.

$$S = \{1, 2, 3, 4, 5, 6\}$$

This suggests that if a person throws the dice once, the facing up pip may be one of the six numbers in 22.

Union

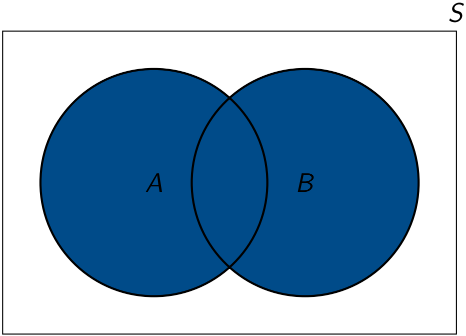
The outcome of an experiment is an *event*. The event is simply a subset of the sample space. So, for the example in 21, an event can be all the head H = head or T = tail. Therefore, the sample space is the *union*—denoted by ∪—of these two events:

$$H \cup T = \{head, tail\}$$

Therefore, the union of two sets A and B is the set of elements which are in A, in B, or in both A and B, namely

$$A \cup B = \{x : x \in A \text{ or } x \in B\}.$$

In the case of the dice experiment, if A = 1, 3, 5 and B = 2, 4, 6 then A ∪ B = 1, 2, 3, 4, 5, 6. The union is often schematically represented with a Venn Diagram.



Intersection

Let us consider the tossing of the coin experiment again; what is the sample space if we toss the coin twice? In this case, sample space is:

$$S = \{(head, head), (head, tail), (tail, head), (tail, tail)\}$$

The event (A) that a head appears in the first coin is thus:

$A = (head, head), (head, tail)$

If there are two events on that the head appears in the first coin (A) and an event that the head appears in the second coin, the second event will be

$B = (head, head), (tail, head)$

An event that contains the results of both A and B is known as the *intersection* of A and B and it is denoted with the symbol ∩.

$$A \cap B = (head, head)$$

So,

$$A \cap B = \{x : x \in A \wedge x \in B\}$$

that is

$x \in A \cap B$ if and only if $x \in A$ and $x \in B$.

So [illegible] the dice experiment the intersection of the sets 1, 2, 3 and 2, 3, 4 is 2, 3. The intersection is represented using a Venn diagram in

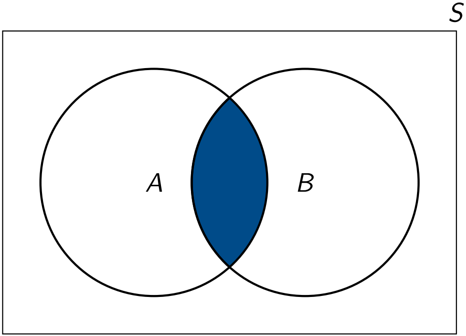


Figure: The shaded region represents intersection of the two events: A ∩ B.

Mutually Exclusive

Let us consider this example: What is the sample space of a dice rolled twice:

$$S = \{(i, j) : i, j = 1, 2, 3, 4, 5, 6\}$$

So, the outcome (i,j) can occur if *i* is the first dice and *j* the second dice. Therefore, the sample space consists of 36 cases. These cases are visualized in Table 1, where in each cell represents (i, j) pair.

Table: Sample space of a dice thrown twice

First Roll/Second Roll	1	2	3	4	5	6
1	(1,1)	(2,1)	(3,1)	(4,1)	(5,1)	(6,1)
2	(1,2)	(2,2)	(3,2)	(4,2)	(5,2)	(6,2)
3	(1,3)	(2,3)	(3,3)	(4,3)	(5,3)	(6,3)
4	(1,4)	(2,4)	(3,4)	(4,4)	(5,4)	(6,4)
5	(1,5)	(2,5)	(3,5)	(4,5)	(5,5)	(6,5)
6	(1,6)	(2,6)	(3,6)	(4,6)	(5,6)	(6,6)

More than two events

If have two events, the first includes all the results from the dice that are equal to 5 and all the second contains all the results that are equal to 6. A = (1,4), (2,3), (3,2), (4,1) B = (1,5), (2,4) (3,3), (4,2), (5,1) When the $AB = 0$, the A and B are known as mutually exclusive events.

When we have more than two events $A_1, A_2, \ldots$, then $\bigcup\limits_{i=1}^{\infty} A_i$ is the union of these events, that is the event that consists of all the results in A_i, for $i = 1 \ to \ \infty$ and the $\bigcap\limits_{i=1}^{\infty} A_i$ is the intersection of these events, that is the event that consists of all the results in A_i, for $i = 1 \ to \ \infty$.

Complement

A^c is the complement of A, which includes all the results in a sample space S, that are not part of A ($A = S\ A$). The complement is represented by the Venn Diagram in Figure 2.

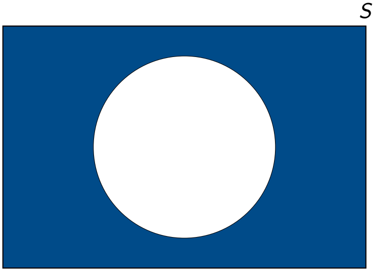


Figure: The shaded region represents the complement E^c of the union of two events A and B.

For two events A and B where all the results of B are also in A, then the B is contained in A or B is a subset of A, and this is denoted with the symbol $\subset$. So the $B \subset A$ stands for the B is a subset of A. Also, the symbol $\supset$ in **A $\supset$ B** means that A is a proper superset of B. The $B \subset A$ is represented by the Venn Diagram in the figure:

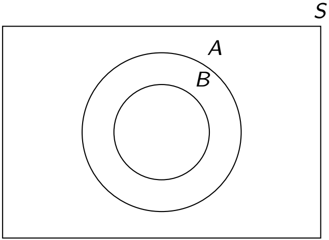


Figure: $B \subset A$..

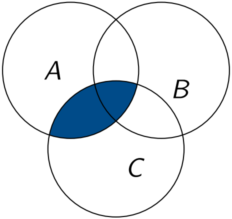
The following are true.

$$\begin{aligned} A \cup A^c &= S \\ A \cap A^c &= \emptyset \\ P(A) &= 1 - P(A^c) \end{aligned}$$

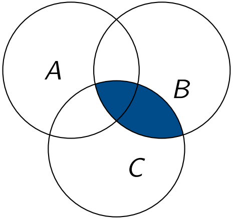
Operations

Commutative property	$\mathrm{A} \cup \mathrm{B} = \mathrm{B} \cup \mathrm{A}$	$\mathrm{AB} = \mathrm{BA}$
Associative property	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$	$\mathrm{(AB)C} = \mathrm{A(BC)}$
Distributive property	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	$\mathrm{AB} \cup \mathrm{C} = (\mathrm{A} \cup \mathrm{C})\ (\mathrm{B} \cup \mathrm{C})$

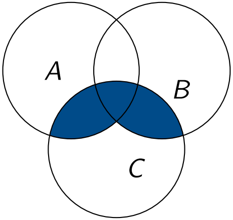
The distributive property is defined using Venn Diagrams:



Shaded region: $A \cap C$



Shaded region: $B \cap C$.



Shaded region: $A \cup (B \cap C)$

DeMorgans Law

In propositional logic and boolean algebra, De Morgan's laws are a pair of transformation rules that are both valid rules of inference. The rules express conjunctions and disjunctions purely in terms of each other via negation. These rules are defined in plain language as follows:

1. the negation of a conjunction is the disjunction of the negations, and
2. the negation of a disjunction is the conjunction of the negations

We may define these rules as follows, in logic:

$$\neg(P \wedge Q) \iff (\neg P) \vee (\neg Q),$$

and

$$\neg(P \vee Q) \iff (\neg P) \wedge (\neg Q),$$

where
P and Q are propositions,

$\neg$ is the negation logic operator (NOT),

$\wedge$ is the conjunction logic operator (AND),

$\vee$ is the disjunction logic operator (OR),

$\iff$ is the symbol for if and only if (IFF).

In a similar manner to logic, the DeMorgans Law can be determined in propabilistic terms as follows:

- the complement of the union of two sets is the same as the intersection of their complements (see 42); and
- the complement of the intersection of two sets is the same as the union of their complements (see 42).

These two laws can be expressed in set theory as follows:

$$(A \cup B)^c = A^c \cap B^c$$

$$(A \cap B)^c = A^c \cup A^c$$

The de Morgan laws can be defined in a more generized form as follows

$$(\bigcap_{i\in I} A_i)^c \equiv \bigcup_{i\in I} A_i^c,$$
$$(\bigcup_{i\in I} A_i)^c \equiv \bigcap_{i\in I} A_i^c,$$

where I is an indexing set.

Kolmogorov axioms

The axiomatization of probability theory has been established by the Russian mathematician Andrey Kolmogorov in his now classic *Foundations of the Theory of Probability* (1933), who proposed a set of axioms,(known as the Kolmogorov axioms or Axioms of Probability. An axiom is a proposition that is assumed without proof as correct. Based on an axiom other conclusions may be drawn.

In probability theory we can assume that for an event A in a sample space S, there is a value $P(A)$, which is the probability of A. The $P(A)$ follows the following three axioms:

First axiom: Non-negativity. The probability of an event is a non-negative real number:

$$0 \leq P(A) \leq 1$$

So, the probability that the outcome of the experiment is an outcome A, which is a number between 0 and 1. In more formal terms this axiom is defined as:

$$P(A) \in \mathbb{R}, P(A) \geq 0 \qquad \forall A \in F$$

where

F is the event space. In particular, $P(A)$ is always finite.

Second axiom: Normalization. The probability that at least one of the events in the entire sample space will occur is 1.

$$P(S) = 1$$

Third axiom: Finite additivity. Any countable sequence of disjoint sets (synonymous with mutually exclusive events) $A_1, A_2, \ldots$ satisfies

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

This suggests that, in a sequence of mutually exclusive events, the probability of at least one of these events to occur is equal to the sum of their probabilities.

Next

- *Conditional Probability* $P(A|B) = P(A, B)/P(B)$, this is simply the probability of A, given that B occurred and is fundamental for understanding probability theory.
- The Conditional Probability *is* Probability $P(A|B)$ is a probability function for any fixed B.
- Any theorem that holds for probability also holds for conditional probability. We use this as the definition of conditional probability.
- We will introduce the *Bayes' Rule*, which unites marginal, joint, and conditional probabilities.