Introduction to Probability Theory

L645
Advanced NLP
Autumn 2009

Probability Theory

To start out the course, we need to know something about statistics and probability

- This is only an introduction; for a fuller understanding, you would need to take a statistics course

Probability theory = theory to determine how likely it is that some outcome will occur

Principles of Counting

- **multiplication principle**: if there are two independent events, P and Q, and P can happen in p different ways and Q in q different ways, then P **and** Q can happen in $p \cdot q$ ways
- **addition principle**: if there are two independent events, P and Q, and P can happen in p different ways and Q in q different ways, then P **or** Q can happen in $p + q$ ways

Principles of Counting – Examples

- example 1: If there are 3 roads leading from Bloomington to Indianapolis and 5 roads from Indianapolis to Chicago, how many ways are there to get from Bloomington to Chicago?

  answer: $3 \cdot 5 = 15$
- example 2: If there are 2 roads going south from Bloomington and 6 roads going north. How many roads are there going south or north?

  answer: $2 + 6 = 8$

Exercises

- How many different 7-place license plates are possible if the first two places are for letters and the other 5 for numbers?
- John, Jim, Jack and Jay have formed a band consisting of 4 instruments.
  - If each of the boys can play all 4 instruments, how many different arrangements are possible?
  - What if John and Jim can play all 4 instruments, but Jay and Jack can each play only piano and drums?

Probability spaces

We state things in terms of an **experiment** (or trial)—e.g., flipping three coins

- **outcome**: one particular possible result

  e.g. first coin = heads, second coin = tails, third coin = tails (HTT)
- **event**: one particular possible set of results, i.e., a more abstract idea

  e.g., two tails and one head ({HTT, THT, TTH})

The set of basic outcomes makes up the **sample space** (Ω)

- Discrete sample space: countably infinite outcomes (1, 2, 3, ...), e.g., heads or tails
- Continuous sample space: uncountably infinite outcomes (1.1293..., 8.765..., ...), e.g., height

We will use $\mathcal{F}$ to refer to the set of events, or **event space**

Sample space

Die rolling

If we have a 6-sided die

- Sample space $\Omega = \{\text{One, Two, Three, Four, Five, Six}\}$
- Event space $\mathcal{F} = \{\{\text{One}\}, \{\text{One, Two}\}, \{\text{One, Three, Five}\} \ldots\}$
  - With 6 options, there are $2^6 = 64$ distinct events

Probability functions

Probabilities are numbers between 0 (impossible) and 1 (certain)

A **probability function (distribution)** distributes a probability mass of 1 over the sample space Ω

- A probability function is any function $P : \mathcal{F} \rightarrow [0, 1]$, where:
  - $P(\Omega) = 1$
  - For disjoint sets $A_j \in \mathcal{F}$: $P(\bigcup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} P(A_j)$ (countable additivity)

    disjoint: $A_j \cap A_k = \emptyset$ for $j \neq k$

In other words: the probability of any of the events A_j happening is the sum of the probabilities of any of the individual events happening

- e.g., $P(roll = 1 \cup roll = 2) = P(roll = 1) + P(roll = 2)$

Example

Toss a fair coin three times. What is the chance of exactly 2 heads coming up?

- Sample space $\Omega = \{\text{HHH,HHT,HTH,HTT,THH,THT,TTH,TTT}\}$
- Event of interest $A = \{\text{HHT,HTH,THH}\}$

Since the coin is fair, we have a uniform distribution, i.e., each outcome is equally likely (1/8)

- $P(A) = \frac{|A|}{|\Omega|} = \frac{3}{8}$

Probabilty function

Die rolling

Sample space $\Omega = \{\text{One, Two, Three, Four, Five, Six}\}$

- For a fair coin, probability mass is evenly distributed
- $P(\{One\}) = P(\{Two\}) = \ldots = \frac{1}{6}$

Event $B =$ divisible by 3

- $P(B) = P(\{Three, Six\}) = \frac{2}{6} = \frac{1}{3}$

Useful fact

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- The probability of unioning A and B requires adding up their individual probabilities
- ... then subtracting out their intersection, so as not to double count that portion

Conditional probability

The **conditional probability** of an event A occurring given that event B has already occurred is notated as $P(A|B)$

- Prior probability of A: $P(A)$
- Posterior probability of A (after additional knowledge B): $P(A|B)$

  (1) $P(A|B) = \frac{P(A \cap B)}{P(B)}$
  - In some sense, B has become the sample space

For example, the probability of drawing a king, given that it's a red suit is:

(2) $P(king|red) = \frac{P(king \cap red)}{P(red)} = \frac{2/52}{26/52} = \frac{1}{13}$

Conditional probability

Die rolling

Let A be outcome of rolling a number divisible by two and C be outcome of rolling a number divisible by four

(3) $P(A|C) = \frac{P(A \cap C)}{P(C)} = \frac{P(\{Four\})}{P(\{Four\})} = 1$

(4) $P(C|A) = \frac{P(A \cap C)}{P(A)} = \frac{P(\{Four\})}{P(\{Two, Four, Six\})} = \frac{1}{3}$

The chain rule

First, we have the multiplication rule, a restatement of $P(A|B) = \frac{P(A \cap B)}{P(B)}$:

(5) $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$

More generally, we have the chain rule, which we will use in Markov models:

(6) $P(A_1 \cap ... \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)...P(A_n|\bigcap_{i=1}^{n-1} A_i)$

i.e., to obtain the probability of a number of events occurring:

- select the first event
- select the second event, given the first
- and so on: select the i'th event, given all of the previous ones

Conditioning E on F and F^c

- Let E and F be events.

  $$E = EF \cup EF^c$$

  where EF and EF^c are mutually exclusive.

- $$\begin{aligned} P(E) &= P(EF) + P(EF^c) \\ &= P(E|F)P(F) + P(E|F^c)P(F^c) \\ &= P(E|F)P(F) + P(E|F^c)P(1 - P(F)) \end{aligned}$$

Example

- An insurance company groups people into two classes: Those who are accident prone and those who are not.
- Their statistics show that an accident prone person will have an accident within a fixed 1-year period with prob. 0.4. Whereas the prob. decreases to 0.2 for a non accident prone person.
- Let us assume that 30 percent of the population are accident prone.
- What is the prob. that a new policyholder will have an accident within a year of purchasing a policy?

Example (2)

- Y = policyholder will have an accident within one year

  A = policyholder is accident prone

- look for P(Y)

- $$\begin{aligned} P(Y) &= P(Y|A)P(A) + P(Y|A^c)P(A^c) \\ &= 0.4 \cdot 0.3 + 0.2 \cdot 0.7 = 0.26 \end{aligned}$$

Independence

An important concept in probability is that of **independence** $\rightarrow$ two events are independent if knowing one does not affect the probability of the other

Events A and B are independent if

- $P(A) = P(A|B)$
- i.e., $P(A \cap B) = P(A)P(B)$

In other words, the probability of seeing A and B together is the product of seeing each one individually because the one does not affect the other.

- $P(roll_i = 3|roll_{i-1} = 3) = \frac{1}{6} = P(roll_i = 3)$
- $P(roll_i = 3 \cap roll_{i-1} = 3) = P(roll_i = 3)P(roll_{i-1} = 3) = \frac{1}{36}$

Independence

Die rolling

Let A be outcome of rolling a number divisible by two and B be outcome of rolling a number divisible by three

- $P(A) = P(\{Two, Four, Six\}) = \frac{1}{2}$
- $P(B) = P(\{Three, Six\}) = \frac{1}{3}$
- $P(A \cap B) = P(\{Six\}) = \frac{1}{6}$

Since $\frac{1}{2} * \frac{1}{3} = \frac{1}{6}$, these two events are independent

- If C is outcome of rolling a number divisible by four, show that A and C are not independent

Bayes' Theorem

The following situation is fairly common in NLP:

- We want to find out $P(B|A)$ (e.g., $P(tag|word)$)
- We know $P(A|B)$ (e.g., $P(word|tag)$)

The good news is that with Bayes' Theorem, we can calculate $P(B|A)$ in terms of $P(A|B)$

(7) $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$

Bayes: Getting the most likely event

Bayes' Theorem takes into account the normalizing constant $P(A)$

(8) $P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A|B)P(B)}{P(A)}$

Often, $P(A)$ is the same for every event we are interested in, and we want to find the value of B which maximizes the function:

(9) arg max$_B$ $\frac{P(A|B)P(B)}{P(A)}$ = arg max$_B$ $P(A|B)P(B)$

... so, in these cases, we can ignore the denominator

Bayes Theorem

Die rolling

With A, B, and C as before ...

- $P(C|A) = \frac{P(A|C)P(C)}{P(A)}$
- $P(A|C) = 1$ since if a number is divisible by 4, it is always divisible by 2
- $P(A) = \frac{1}{2}$; $P(C) = \frac{1}{6}$

Plugging this in, we get $P(C|A) = \frac{1 * \frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$

Partitioning

Let's say we have i different, disjoint sets B_i, and these sets partition A (i.e., $A \subseteq \bigcup_i B_i$)

Then, the following is true:

(10) $P(A) = \sum_i P(A \cap B_i) = \sum_i P(A|B_i)P(B_i)$

This gives us the more complicated form of Bayes' Theorem:

(11) $P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_i P(A|B_i)P(B_i)}$

Partitioning

Die rolling

Let B be the outcome of rolling a number divisible by 3

- We can partition the sample space into $A_1 = \{$One$\}$, $A_2 = \{$Two$\}$, ...
- Thus, $P(A_i) = \frac{1}{6}$ for all i

Let $P(B|A_i) = 1$ when i is divisible by 3 and 0 otherwise

(12) $P(B) = 0 * \frac{1}{6} + 0 * \frac{1}{6} + 1 * \frac{1}{6} + 0 * \frac{1}{6} + 0 * \frac{1}{6} + 1 * \frac{1}{6} = \frac{1}{3}$

Example of Bayes' Theorem

Assume the following:

- Bowl B_1 ($P(B_1) = \frac{1}{3}$) has 2 red and 4 white chips
- Bowl B_2 ($P(B_2) = \frac{1}{6}$) has 1 red and 2 white chips
- Bowl B_3 ($P(B_3) = \frac{1}{2}$) has 5 red and 4 white chips

Given that we have pulled a red chip, what is the probability that it came from bowl B_1? In other words, what is $P(B_1|R)$?

- $P(B_1) = \frac{1}{3}$
- $P(R|B_1) = \frac{2}{2+4} = \frac{1}{3}$
- $$\begin{aligned} P(R) &= P(B_1 \cap R) + P(B_2 \cap R) + P(B_3 \cap R) \\ &= P(R|B_1)P(B_1) + P(R|B_2)P(B_2) + P(R|B_3)P(B_3) = \frac{4}{9} \end{aligned}$$

So, we have: $P(B_1|R) = \frac{P(R|B_1)P(B_1)}{P(R)} = \frac{(1/3)(1/3)}{4/9} = \frac{1}{4}$

Random variables

Randomly select a rat from a cage and determine its sex

- Sample space $\Omega = \{M,F\}$
- Let X be the function such that $X(M) = 1$ and $X(F) = 0$
- The range is thus $\{0, 1\}$, i.e., a set of real numbers

We call X a **random variable**, defined as a function from Ω to a subset of $\mathbb{R}^n$ (where n often equals 1)

- In other words, the sample space is mapped to a numeric value associated with each outcome
- In this case $X : \Omega \rightarrow \{0, 1\}$

Discrete random variables

A discrete random variable maps Ω to S, where S is a countable subset of $\mathbb{R}$.

- if $S = \{0,1\}$, as in the previous example, it is called a *Bernoulli trial*.

Example: roll two dice and look at their sum

- $\Omega = \{1;1, 1;2, 1;3, ... 3:6, 4:1, ... 5;6, 6;6\}$
- $S = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$
- The random variable X maps every item from Ω to S, and what we are interested in are the items from S (the sums)

Probability mass function

The **probability mass function (pmf)** for a random variable X allows us to say something about the probability of the values of S, i.e., the items related to the event space.

- $p(X = 2) = \frac{1}{36}$ (the case where 1 is rolled on both dice)
- $p(X = 3) = \frac{2}{36}$ (rolling 1;2 or 2;1)

In general, we can talk about $p(X = x)$, where x is a member of S, and we can rewrite this more simply as $p(x)$

Note that: $\sum_i p(x_i) = P(\Omega) = 1$

Example

Suppose that our experiment consists of tossing 3 fair coins. If we let Y denote the number of heads appearing, then Y is a random variable taking on one of the values 0, 1, 2, 3 with respective probabilities.

$$\begin{aligned} P\{\ Y = 0\} &= P(\{(T,T,T)\} = \frac{1}{8} \\ P\{\ Y = 1\} &= P(\{(T,T,H),(T,H,T),(H,T,T)\} = \frac{3}{8} \\ P\{\ Y = 2\} &= P(\{(H,H,T),(H,T,H),(T,H,H)\} = \frac{3}{8} \\ P\{\ Y = 3\} &= P(\{(H,H,H)\} = \frac{1}{8} \end{aligned}$$

Example (2)

$$\begin{aligned} P\{\ Y = 0\} &= P(\{(T,T,T)\} = \frac{1}{8} \\ P\{\ Y = 1\} &= P(\{(T,T,H),(T,H,T),(H,T,T)\} = \frac{3}{8} \\ P\{\ Y = 2\} &= P(\{(H,H,T),(H,T,H),(T,H,H)\} = \frac{3}{8} \\ P\{\ Y = 3\} &= P(\{(H,H,H)\} = \frac{1}{8} \end{aligned}$$

Since Y must take on one of the values 0 through 3,

$$1 = P(\bigcup_{i=1}^{3}\{Y = i\}) = \sum_{i=1}^{3} P\{Y = i\}$$

Expectation

The **expectation** of a random variable is simply the mean, or average

(13) $E(X) = \sum_x x \cdot p(x)$

In other words, the expectation of X, or $E(X)$, is the sum of each value times its probability

- Example: we roll one die and Y is the value that turns up. The expectation (expected value) is:

  (14) $E(Y) = \sum_{y=1}^{6} y \cdot p(y) = \sum_{y=1}^{6} y \cdot \frac{1}{6} = \frac{1+2+3+4+5+6}{6} = 3\frac{1}{2}$

- Because it's a uniform probability, the expectation in this case is the same as a mean/average

Weighted die

Let's assume we have the following weighted die:

- $p(X = 1) = p(X = 2) = p(X = 5) = \frac{1}{6}$
- $p(X = 3) = p(X = 4) = \frac{1}{12}$
- $p(X = 6) = \frac{1}{3}$

What is the expectation here?

Expectation of functions

Assume we have $Y = g(X)$

- Then, $E(Y) = E(g(X)) = \sum_x g(x) \cdot p_X(x)$

So, if $Y = X^2$, then:

- Then, $E(Y) = E(X^2) = \sum_x x^2 \cdot p_X(x) = \sum_{x=1}^{6} x^2 \cdot \frac{1}{6} = \frac{91}{6}$

Variance

Expectation alone ignores the question:

- Do the values of a random variable tend to be consistent over many trials or do they tend to vary a lot?

The measure of **variance** answers this question by calculating how much on average the values deviate from the mean (expectation).

(15)
$$\begin{aligned} Var(X) = \sigma^2 &= E((X - E(X))^2) \\ &= E(X^2) - E^2(X) \end{aligned}$$

The **standard deviation** is the square root of the variance

(16) $\sigma = \sqrt{\sigma^2}$

Variance

Die rolling

With one die rolled, we know the expectation ($E(X)$) to be $\frac{7}{2}$

- $Var(X) = E((X - E(X))^2) = \sum (x - \frac{7}{2})^2 \cdot \frac{1}{6} = \frac{35}{12}$

Alternatively, we could calculate this as:

- $Var(X) = E(X^2) - E(X)^2 = \frac{91}{6} - (\frac{7}{2})^2 = \frac{35}{12}$

Example for expectation and variance

When we roll two dice, what is the expectation and the variance for the sum of the numbers on the two dice?

(17)
$$\begin{aligned} E(X) &= E(Y + Y) \\ &= E(Y) + E(Y) \\ &= 3.5 + 3.5 = 7 \end{aligned}$$

(18)
$$\begin{aligned} Var(X) &= E((X - E(X))^2) \\ &= \textstyle\sum_x p(x)(x - E(X))^2 \\ &= \textstyle\sum_x p(x)(x - 7)^2 = 5\frac{5}{6} \end{aligned}$$

Joint distributions

With (discrete) random variables, we can define:

- **Joint pmf**: The probability of both x and y happening

  (19) $p(x, y) = P(X = x, Y = y)$

- **Marginal pmfs**: The probability of x happening is the sum of the occurrences of x with all the different y's

  (20) $p_X(x) = \sum_y p(x, y)$
  (21) $p_Y(y) = \sum_x p(x, y)$

If X and Y are independent, then $p(x, y) = p_X(x)p_Y(y)$, so, e.g., the probability of rolling two sixes is:

- $p(X = 6, Y = 6) = p(X = 6)p(Y = 6) = (\frac{1}{6})(\frac{1}{6}) = \frac{1}{36}$

Where do we get the probability distributions?

For natural language, we have to obtain our probabilities from corpora, typically by calculating the **relative frequency** of language phenomena

The relative frequency is the number of times an outcome (u) occurs, out of all (N) trials:

(22) $f(u) = \frac{C(u)}{N}$

From this, we can get probability estimates and fit them into a distribution

- We'll talk more about estimating frequencies when we talk about smoothing

Standard distributions

Certain probability distributions recur in lots of different kinds of data.

- These **distributions** are certain functions, but with different constants (**parameters**), which explain how different kinds of data fit into them

We will look at two common distributions:

- The binomial distribution—applicable to discrete data (i.e., countable)
- The normal distribution—applicable to continuous data (i.e., uncountably infinite)

Binomial distribution (discrete data)

- A binomial distribution results from a series of trials with two outcomes (Bernoulli trial)
- It is assumed that the trials are independent
  - This is not true in general for language tasks, such as the probability of one word being independent of the next one, but it is a useful approximation

Formula (r = number of successes, n = number of trials, p = probability of success):

(23) a. $b(r; n, p) = \binom{n}{r} p^r (1 - p)^{n-r}$
  b. with $\binom{n}{r} = \frac{n!}{(n-r)!r!}$ and $0 \leq r \leq n$

NB: $\binom{n}{r}$ is the number of ways to choose r objects from n objects

Binomial example

If we toss a coin n times (n trials), we want to know the number of heads (r successes) that come up, assuming that the coin has probability p of turning up a head

(24) This is the binomial distribution $p(r) = b(r; n, p)$

Assume the coin is fair ($p = .5$) and that we throw it ten times:

(25) $b(r; 10, 0.5) = \frac{10!}{(10-r)!r!}(0.5)^r(0.5)^{(10-r)}$

We define this for all values of r ($0 \leq r \leq 10$)

The probabilities of the distribution

(26) a. $b(0; 10, 0.5) = 0.0009765625$
  b. $b(1; 10, 0.5) = 0.009765625$
  c. $b(2; 10, 0.5) = 0.0439453125$
  d. $b(3; 10, 0.5) = 0.1171875$
  e. $b(4; 10, 0.5) = 0.205078125$
  f. $b(5; 10, 0.5) = 0.24609375$
  g. $b(6; 10, 0.5) = 0.205078125$
  h. $b(7; 10, 0.5) = 0.1171875$
  i. $b(8; 10, 0.5) = 0.0439453125$
  j. $b(9; 10, 0.5) = 0.009765625$
  k. $b(10; 10, 0.5) = 0.0009765625$

Normal distribution (continuous data)

If we graphed the previous distribution and then made the distribution continuous (connect the dots), it would look like the normal distribution.

When the data is continuous, we can use distributions like the **normal distribution**, or Gaussian distribution

(27) $n(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma)^2}$

We will not deal with continuous distributions as much this semester

Bayesian statistics

Sometimes we have an idea of what a probability is (prior probability), but further data will show that we need to alter this probability (posterior probability)

For example, we assume a coin is fair ($p = 0.5$), but then we flip it 10 times and get 8 heads.

- μ = our model, so $P(\mu) = 0.5$
- s = our observed sequence, so $P(s) = 0.8$

Glossing over the mathematics (involving Bayes' Law), the end result is that $P(\mu|s) = 0.75$ ... not quite 0.8, but further removed from 0.5

Bayesian decision theory

We can also use Bayesian statistics to compare two different models.

- Assume models μ and v, which explain some set of data
- We compare them by taking the ratio of $\frac{P(\mu|s)}{P(\nu|s)}$
- And, by Bayes' Theorem, this is the same as $\frac{P(s|\mu)P(\mu)}{P(s|\nu)P(\nu)}$
- And if, prior to this, both theories are equally likely ($P(\mu) = P(\nu)$), then this reduces to: $\frac{P(s|\mu)}{P(s|\nu)}$

This is called a **likelihood ratio**; if it's above 1, go with μ, and below 1, go with ν