Information Retrieval

Chapter 2: Mathematic Foundations and Text Mining Basics

Sanasam Ranbir Singh

Contents

In this chapter we review a number of basic concepts in probability and text analysis that are used throughout the book. Each of these topics is worthy of an entire chapter (or even a whole book) by itself, so our treatment is necessarily brief. Nonetheless, the goal of this chapter is to provide the introductory foundations for models and techniques that will be widely used in the remainder of the book. Readers who are already familiar with these concepts, or who want to avoid mathematical details during a first reading, can safely skip to the next chapter.

1 Probability

1.1 Discrete Random Variables

Let x be a discrete random variable[1] which can take any of the finite number of n different values from a sample space[2] $\mathcal{X} = \{v_1, v_2, ..., v_n\}$. Then p_i denotes the probability that the random variable x takes the value v_i.

$$p_i = Pr(x = v_i), \quad \text{for } i = 1, 2, ..., n$$

It is also denoted by the *probability mass function* $P(x)$. It satisfies the following two conditions.

- For any $x \in \mathcal{X}$, $P(x) \geq 0$.
- For the sample space $\mathcal{X}$, $Pr(\mathcal{X}) = 1$ (probability of the random variable x taking any value from $\mathcal{X}$ is 1) i.e. $\sum_{i=1}^{n} p_i = 1$

In probability, an event is a subset of the sample space $\mathcal{X}$ i.e., an event A can be associated with more than one outcome of an experiment ($A \subseteq \mathcal{X}$). Two events A and B are called *mutually disjoint* or *pairwise disjoint*, if $A \cap B = \phi$. According to Kolmogorovs axioms, each event A has a probability, which is a number. These numbers satisfy the following axioms.

- For any event A, $Pr(A) \geq 0$.
- For the sample space $\mathcal{X}$, $Pr(\mathcal{X}) = 1$
- For any k number of mutually disjoint events $\{A_1, A_2, ..., A_k\}$, $Pr(A_1 \cup A_2 \cup ... \cup A_k) = \sum_{i=1}^{k} Pr(A_i)$

If A and B are two events of a sample space $\mathcal{X}$, the $Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$. This property is known as inclusion-exclusion properties.

1.2 Expected Values

If x be a discrete random variable which takes values from the sample space $\mathcal{X} = \{v_1, v_2, ..., v_n\}$, the expected value or mean of x is the number $E(x)$ given by the formula

$$E(x) = \sum_{i=1}^{n} Pr(x = v_i).v_i$$

It is nothing but weighted average of all possible values. If we assume that the probability of x taking any value in $\mathcal{X}$ is equally likely, expectation is the sample average i.e., $\frac{1}{n} \sum_{i=1}^{n} v_i$. A random variable may also take infinitely many values where expectation will be defined as $E(x) = \sum_{i=1}^{\infty} Pr(x = v_i).v_i$. However, majority of the estimation in this book will be dealing with finite sample space.

[1] A random variable is neither random nor a variable, but a random variable is a function defined on a sample space.

[2] set of possible outcomes

Further variance of x is defined as $Var(x) = E(x^2) - E(x)^2$ where $E(x^2)$ is the expectation of the values x^2 i.e.,

$$E(x^2) = \sum_{i=1}^{n} Pr(x = v_i).v_i^2$$

If μ is the mean of the random variable i.e., $E(x)$, $Var(x)$ can also be defined as

$$Var(x) = E(x - \mu)^2 = \sum_{i=1}^{n} Pr(x = v_i).(v_i - \mu)^2$$

The *standard deviation* σ of the random variable x is defined as $\sqrt{Var(x)}$.
Formally, if $f(x)$ be any function of x, expectation of $f(x)$ is defined as

$$E(f(x)) = \sum_{x \in \mathcal{X}} P(x).f(x)$$

Expectation is linear in nature i.e., $E(\omega_1 f_1(x) + \omega_2 f_2(x)) = \omega_1 E(f_1(x)) + \omega_2 E(f_2(x))$.

1.3 Joint probability mass function of two random variables

Let x and y be two discrete random variables which take values from $\mathcal{X} = \{v_1, v_2, ..., v_n\}$ and $\mathcal{Y} = \{w_1, w_2, ..., w_m\}$ respectively. Then, the P_{ij} denotes the probability that the r.v. x takes the value v_i and the y takes w_j i.e., $Pr(x = v_i, y = w_j)$. It also satisfy the conditions

- $p_{ij} \geq 0$
- $\sum_{i=1}^{n} \sum_{j=1}^{m} P_{ij}$

Further, the *marginal distribution* of the random variables x and y can also be defined as

$$Pr(x = v_i) = \sum_{j=1}^{m} Pr(x = v_i, y = w_j)$$

and

$$Pr(y = w_j) = \sum_{i=1}^{n} Pr(x = v_i, y = w_j)$$

The two random variable x and y are statistically independent if and only if

$$Pr(x = v_i, y = w_j) = Pr(x = v_i).Pr(y = w_j)$$

For any function $f(x, y)$, the expectation of $f(.)$ can be defined as

$$E(f(x, y)) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y).f(x, y)$$

The mean and variance of the random variables x and y are

$$\mu_x = E(x) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y).x$$

$$\mu_y = E(y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y).y$$

$$\sigma_x^2 = Var(x) = E((x - \mu_x)^2) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P(x, y).(x - \mu_x)^2$$

and

$$\sigma_y^2 = Var(y) = E((y-\mu_y)^2) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y).(y-\mu_y)^2$$

The covariance of x and y is

$$\sigma_{xy} = Var(x,y) = E((x-\mu_x)(y-\mu_y)) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} P(x,y).(x-\mu)(y-\mu_y)$$

It is often used as one measure of the degree of statistical dependence between x and y. The normalized covariance is known as correlation co-efficient and defined it as

$$\rho = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$

1.4 Conditional Probability

The probability of the value of one random variable x, given the value of another random variable y is defined as

$$Pr(x = v_i | y = w_j) = \frac{Pr(x = v_i, y = w_j)}{Pr(y = w_j)}$$

which can also be written as

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

If x and y are statistically independent, then $P(x|y) = P(x)$.
Further, let $A_1, A_2, ..., A_k$ be mutually exclusive events satisfying the following conditions

- $A_i \cup A_j = \phi$ for any i and j such that $i \neq j$.
- $A_1 \cup A_2 \cup .... \cup A_k = \mathcal{S}$, where $\mathcal{S}$ is the sample space.

Then the probability of an event B can be defined as

$$P(B) = \sum_{i=1}^{k} P(B|A_i).P(A_i)$$

This is known as *law of total probability*. Now, we can write

$$P(y) = \sum_{x\in\mathcal{X}} P(y|x)P(x)$$

It is equivalent to the marginal probability defined above. Now, the conditional probability $P(x|y)$ can be written as

$$P(x|y) = \frac{P(x,y)}{P(y)} = \frac{P(y|x).P(x)}{\sum_{x\in\mathcal{X}} P(y|x)P(x)}$$

This expression is also known as *Bayes Rule*. It is also realized as

$$posterior = \frac{likelihood \times prior}{evidence}$$

2 Information Theory

2.1 Information Content

The *Shannon* information content of an outcome x is defined as

$$h(x) = \log_2 \frac{1}{p(x)}$$

It is measured in *bits*.

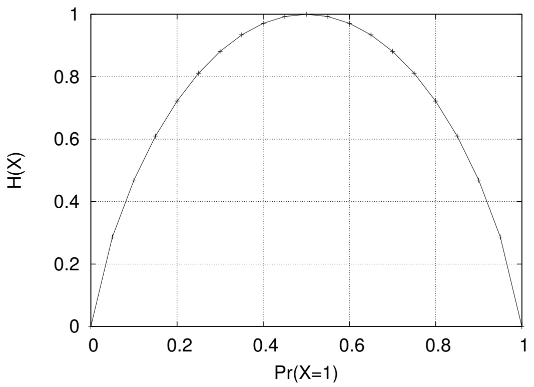

Figure 1: Entropy distribution of a binary variable.

2.2 Entropy

In information theory, entropy is a measure of randomness or uncertainty of a random variable being drawn. Assume we have a random variable X with a discrete set of symbols $\{x_1, x_2, , ..., x_m\}$ with associated probabilities $Pr(X = x_i)$, then the entropy of the discrete distribution is defined as follows:

$$H(X) = -\sum_{i=1}^{m} Pr(X = x_i) \log_2 Pr(X = x_i) \tag{1}$$

It is measured in *bits*, if the logarithm is used in base 2. The value of $H(X)$ is maximum when $Pr(X = x_i)$ is uniformly distributed over all variables. Considering a binary random variable $X = \{0, 1\}$, we plot the entropy distribution of the variable X in Figure 1. If $p = Pr(X = 0)$ be a probability of the symbol 0, $1 - p$ is the probability of the symbol 1. It can be clearly seen from the figure that entropy is maximum when $p = 0.5$ (i.e., uniform distribution) and it is minimum when $p = 0$ or $p = 1$ (i.e., most skewed distribution).

2.3 Joint Entropy

Given two random variables X and Y, the joint entropy of X and Y is defined as

$$H(X, Y) = -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P(x, y) \log P(x, y)$$

2.4 Conditional Entropy

Given two random variables X and Y, the conditional entropy of X given Y is defined as

$$H(X|Y) = -\sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} P(x, y) \log P(x|y)$$

2.5 Pointwise Mutual Information (PMI)

Suppose we have two probability distributions $Pr(X = x_i)$ and $Pr(Y = y_j)$ over the discrete set of symbols x_i and y_j of two random variables X and Y respectively. Then, pointwise mutual information between x_i and y_j is defined as follows:

$$PMI(x_i, y_j) = \log_2 \frac{Pr(X = x_i, Y = y_j)}{Pr(X = x_i) \cdot Pr(Y = y_j)} \tag{2}$$

where $Pr(X = x_i, Y = y_j)$ is the joint probability between x_i and y_j. Pointwise mutual information is also referred to as *association norm* in some literatures [**?**, **?**]. $PMI(x_i, y_j)$ represents the amount of information convey by the occurrence of one symbol about the another.

2.6 Mutual Information

Mutual information between two random variables X and Y is the reduction in uncertainty about one variable due to the knowledge of another variable and can be expressed as follows:

$$MI(X,Y) = H(X) - H(X|Y) \tag{3}$$

where $H(X|Y)$ is the conditional entropy of X given Y. This equation can be further simplified as follows:

$$MI(X,Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} Pr(X=x_i, Y=y_j)\log_2 \frac{Pr(X=x_i, Y=y_j)}{Pr(X=x_i)\cdot Pr(Y=y_j)} \tag{4}$$

Therefore, MI is the expectation of the $PMI(x_i, y_j)$, where $x_i \in X$ and $y_j \in Y$. If the two random variables X and Y are independent of each other, then $MI(X,Y) = 0$.

2.7 Kullback-Leibler divergence (KLD)

Suppose we have two probability distributions $p(X = x_i)$ and $q(X = x_j)$ over a random variable X, the distance between $p(X = x_i)$ and $q(X = x_j)$ can be defined by the Kullback-Leibler divergence as follows.

$$KLD\big(p(X=x_i)||q(X=x_j)\big) = p(X=x_i).\log\left(\frac{p(X=x_i)}{q(X=x_j)}\right) \tag{5}$$

KLD (also known as relative entropy) is a non-symmetric measure of the difference between two probability distributions of a random variable.

3 Distance Measure

3.1 Cosine Similarity

Given two vectors $\mathbf{d}_i = \{w_{i1}, w_{i2}, ..., w_{in}\}$ and $\mathbf{d}_j = \{w_{j1}, w_{j2}, ..., w_{jn}\}$, the cosine similarity between the two vectors $\mathbf{d}_i$ and $\mathbf{d}_j$ is defined as follows:

$$\text{cosine}(\mathbf{d}_i, \mathbf{d}_j) = \frac{\sum_{k=0}^{n} w_{ik}.w_{jk}}{\sqrt{\sum_{k=0}^{n} w_{ik}^2}.\sqrt{\sum_{k=0}^{n} w_{jk}^2}} \tag{6}$$

3.2 Pearson's correlation coefficient

Pearson's coefficient, also known as *correlation coefficient* or *Pearson's product-moment coefficient* represents the linear relationship between two variables that are measured on the same interval or ratio scale. If X and Y are variables having a series of n values written as x_i and y_i where $i = 1, 2, ..., n$, then the Pearson product-moment correlation coefficient can be used to estimate the correlation between X and Y. The Pearson correlation coefficient, which is often denoted as ρ, is defined as:

$$\rho = \frac{\sum_{i=1}^{n}\big(x_i - \bar{x}\big)\big(y_i - \bar{y}\big)}{(n-1).\sigma_x.\sigma_y} \tag{7}$$

where $\bar{x}$ and $\bar{y}$ are sample means of X and Y respectively, and, σ_x and σ_y are standard deviations of X and Y respectively. As with the population correlation, it can be written as

$$\rho = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}} \tag{8}$$

The value of ρ ranges from -1 to +1. A value of +1 is the result of a perfect positive relationship between the variables. Conversely, a value of -1 represents a perfect negative relationship. It has been shown that the Pearson coefficient can be deceptively small when it is used with a non-linear equation.

4 Linear Algebra

—— yet to edit

5 Text Mining Basics

Text (or documents) collected from various sources using programs such as *crawler* or *feeds* are heterogeneous in nature. They come in variety of formats such as HTML, XML, pdf, Microsoft word, power point and so on. Traditional IR systems (search engines) require that these documents are converted into a format (mostly plain text) accepted by the system. Crawlers are the programs that are responsible for identifying and collecting Web pages from Web (needed for building Web search engines). Whereas *feeds* is a mechanism for accessing or collecting a real time stream of documents such as news feed. RSS feed is a common standard for document feeding from pre specified sources. While crawlers generally attempt to discover new documents (it can also do updates) from previously unknown sources, feeds mainly focus on acquiring new documents or updates from known sources. Apart from identifying new sources, a Web crawler has several responsibilities (details of designing Web crawlers are discussed in Chapter **??**) such as identifying death pages, updating previously visited pages, maintaining the quality of the document repository etc.
One of the tasks at the core of any text IR system is document representation so that several IR techniques (weighing, indexing, searching etc) can be applied uniformly automatically across all the documents in the repository. For do so, it is necessary to pre-process the text documents and store the information in a data structure, which is more appropriate for further processing. In the following section, we briefly introduce some of the terminologies which would be used throughout the book.

5.1 Text Preprocessing

5.1.1 Tokenization

A process of getting all the words present in a document. A text document is split into a stream of words (tokens) by removing all punctuation marks and by replacing tabs and other non-text characters by single white spaces. This tokenized representation is then used for further processing. Depending on the language or nature of the text contents, several important decisions (that potentially affect retrieval performance) need to be taken at the tokenization step. For example, are `information` and `Information` same? or how to deal with capital letters, hyphens or apostrophes? or is `on-line` a single word or two words. All this issues make the tokenization process a non-trivial process.

5.1.2 Parser

It is a component that is responsible for scanning a tokenized document and identify structure elements such as titles, headings, meta data, figures, links etc. Depending on the format of the underlying documents such as HTML or XML, parser need to know the structural properties of the document format and remove unnecessary tags, and/or infer meaningful structural properties to the output texts.

5.1.3 Dictionary of a text collection

The set of different words obtained by merging all text documents of a collection is called the dictionary of a document collection.

5.1.4 Stop words removal

Commons words such as `a, the, an, of, at, in` etc. are known stop words. Stop words are considered to possess little information by their own (explained in chapter **??**) and thus removed. There is no well accepted list of stop words. However, traditional systems work with some pre-defined set of words. Effect of stop word removal on IR performance is discussed in section ??.

5.1.5 Stemming

A process of transforming a word to its *stem* (root) word. For example, transform the word `retrieved` to its root word `retriev`. The words `retrieve, retrieves, retrieved` and `retrieving` are different tokens, however all of them probably have same meaning. After stemming, all these terms refer to the same root word `retriev`. There are several algorithms to stem a word. Some of the commonly used stemming algorithms are *Porters stemming*[3], Lovins Stemmer[4], Dawson Stemmer[5] etc. Effect of word stemming on IR performance is discussed in section ??.

5.1.6 Document Representation

Most of the text mining techniques considers *bag-of-words* representation of a document i.e. a text document is described based on the set of words contained in it (it does not consider the sequence in which the words appear). *Vector space model*, *Probabilistic model* and *Language Model* are commonly used representation models using bag-of-word format.

In vector space model, it is assumed that each document d is represented by a *term vector* of the form $\mathbf{d} = \{w_1, w_2, ..., w_n\}$, where w_i is a weight associated with the term f_i. In boolean representation, w_i is either 0 or 1 to represent *absence* or *presence* of the term f_i in a document. Otherwise, w_i is defined by some scoring functions. Several term weighting methods are introduced in chapter ??.

5.2 Power law distribution

5.2.1 Zipf's and Heaps' Law

5.2.2 Scale-free properties

References

[3]

[4]

[5]