What is entropy, how is it connected to business data and why should it be maximised? Entropy is often understood as a measure of the degree of disorder or uncertainty in a system. So, why shall we want to maximise uncertainty or disorder? The present paper tries to answer these questions. We show that the maximum entropy principle ensures that the outcome of a statistical model is unbiased, robust, and actionable. The maximum entropy method is a general-purpose technique in machine learning. It has a simple and precise mathematical foundation. A number of aspects make it well suited for the modelling of distributions of business data.

1 Introduction

The concept of entropy has been introduced to physics in the beginning of the 19th century in connection with the second law of thermodynamics. Later in the 19th century, Boltzmann, Gibbs, and others put entropy on a statistical basis (for a short historical overview see [1]). Whereas other concepts of thermodynamics such as energy, temperature, and pressure directly impact our daily life, entropy remains mysterious and its relevance to daily life is unclear for most people. In popular literature, it is associated with disorder or with the heat death of the universe (to the interested reader I recommend [2]). Given these historical facts, it is unclear where there is a connection of entropy to statistics or even business.

To understand what entropy has to do with statistics, we have to explain the statistical concept behind it. In fact, entropy is connected to information. We will therefore first discuss information. We will then explain its importance in statistics. To do that we use an example of business data. We will argue that the maximum entropy method is useful because it allows to carefully avoid anything that is not known. It yields results which are essentially free of prejudice or bias. This makes it important for data analysis.

2 Information and Entropy

Let us begin with a simple and well known example, DNA. It is widely accepted that the information that nature needs to build a living organism is stored in its DNA. How large is that information? We know that DNA is build out of nucleotides, abbreviated by the four letters A, G, C, and T. A string of DNA is therefore a very long sentence formed of these four letters. Just like the information coded in a usual English text, the sequence of the letters in DNA codes information. Let us take a short DNA sequence of three letters. There are \( N = 4^3 = 64 \) different possibilities to form such three letter sequences. Suppose, you want to store the information of a three letter sequence in a bit sequence in a computer. Since we have four letters, it is sufficient to code each letter with two bits. We could for instance represent A by 00, G by 01, C by 10, and T by 11. For a three letter word we therefore need a storage of 6 bits. We can therefore say that the information stored in a DNA sequence of length three is 6 bits.

This concept can be generalised. Suppose that we have an arbitrary code and an arbitrary sequence of letters.
or words formed using that code. Then, the information contained in that sequence is \( \log_2 N \), where \( N \) is the number of possibilities to form such a sequence. We need \( \log_2 N \) bits to store the sequence in a computer.

In physics, information theory, and statistics one uses the natural logarithm \( \ln \) instead of the binary logarithm \( \log_2 \), but that does not alter the general concept. We will therefore use the natural logarithm in what follows as well. If we have \( N \) possible sequences of letters, each occurring with the same probability, the information is \( I = \ln N \). The probability to pick one of these sentences is \( p = \frac{1}{N} \). We can therefore write the information as

\[
I = \ln N = -\ln \frac{1}{N} = -\ln p.
\]

In this way we express the information \( I \) stored in a sequence of letters by the probability \( p \) that this sequence occurs. Let us now suppose, that for some reason the probabilities for the sequences of letters are not the same. This is the case for letters, words, or sentences in natural languages. In English language, the letter e is much more often used than e.g. j [3]. The same is true for DNA, different combinations of letters occur more often than others. Therefore, for a given sentence \( s \) we denote the probability by \( p_s \). The information for the given sentence is \( I_s = -\ln p_s \). The average information for all possible sentences is then

\[
\bar{I} = -\sum_s p_s \ln p_s.
\]

Up to the minus sign in front of this expression, this is the expression for the entropy used in statistical physics. The entropy is

\[
S = -\bar{I} = \sum_s p_s \ln p_s.
\]

This expression for the entropy is therefore sometimes called information entropy. You can already find it in the work of Boltzmann, Gibbs, and others [1]. It is as well important in algorithmic information theory and complexity theory [4].

Knowing that entropy is negative information, we may interpret entropy as missing information. This makes the relation to disorder clear. Higher entropy means less information. Disorder destroys information and thus increases entropy.

### 3 Maximum Entropy and Statistics

At a first glance, the concept of maximum entropy may sound stupid: If entropy is negative information, increasing (or even maximising) entropy means decreasing information. In an information driven world of big data, shouldn't we try to increase information?

To illustrate the maximum entropy principle let us take the following example: Suppose a manager wants to know whether and how DSO in his company depends on delivery reliability. DSO is the time it takes until a customer pays an invoice. As described in a previous paper in this series [5], it may be important to look at the distribution of DSO. We show a simplified example of such a distribution in Fig. 1.

![DSO distribution with payment target 30 days.](image)

The question of how DSO depends on delivery reliability is therefore more complicated than one may expect. The distribution has a maximum, a tail, and several outliers. Improving delivery reliability may influence the maximum, the tail, and the outliers differently. More precisely, the manager should ask how the distribution of DSO depends on delivery reliability. To answer that question, one needs a statistical model for the distribution of DSO depending on one or more KPIs which are suitable in the given enterprise to describe delivery reliability.

It is possible to invent such a model with a huge number of parameters which fits exactly the distribution found in the data. In statistics, that would be called over-fitting. From a business point of view, the parameters would have no operational meaning. One does not know how they are connected to business processes and what a change in a parameter actually means or how it could be achieved in a concrete situation. This shows that over-fitting yields poor and useless results.

From an information point of view over-fitting means that more information was put into the model than is actually present in the data. The result is a biased model. This is a very important point: To extract meaningful information from data, which in the end leads to actions, we must avoid bias as much as possible. What we actually want is an unbiased result. Jaynes, who was the first to propose the maximum entropy principle [8],
formulated it as follows:
The fact that a certain probability distribution maximises entropy subject to certain constraints representing our incomplete information, is the fundamental property which justifies use of that distribution for inference; it agrees with everything that is known, but carefully avoids anything that is not known. [9]

To obtain an unbiased statistical model for the DSO distribution, we proceed as follows. We collect the meaningful information like the payment targets (in Fig. 1 only data for a single payment target are shown), the KPIs describing delivery reliability, and eventually some further potentially influencing factors. We take this information as constraints and calculate the distribution such that its information is minimal under these constraints. Doing that, we make sure that only the information really contained in the constraints enters. The resulting distribution has a maximum entropy under the given constraints. It is unbiased. And it depends on meaningful parameters. Changing a KPI in that model shows directly how it influences the distribution and thereby DSO. These changes are actionable. The maximum entropy principle yields an unbiased, robust, and actionable description of our data.

This answers as well the paradox mentioned above. The point is not to minimise information in an absolute sense. Instead, the relevant information is taken into account, it is put into certain constraints. What the maximum entropy principle does is to make sure that only this information is taken into account, not some additional spurious information.

The idea of maximum entropy as described so far has to be extended to situations where certain fixed conditions have to be taken into account. And it may be that latent variables are needed, variables which are not directly part of the data but which nevertheless influence the distribution. These are technical aspects. One then deals with conditional maximum entropy or latent maximum entropy. For business data and their distributions, a combination of both is often suitable. How this can be achieved has been discussed in [6, 7].

4 Few mathematical details

This short section provides some mathematical details of the maximum entropy principle. We try to be quite elementary here. The reader who is not interested in these details may skip this section.

The goal of the maximum entropy principle is to obtain an unbiased model distribution. Let us start with a very simple example. Suppose that there are \( n \) different events. For the event \( s \), we denote the probability as \( p_s \). As stated above, the entropy is \( S = \sum s p_s \ln p_s \). The different \( p_s \) are not completely independent of each other. Since they are probabilities, one has \( \sum s p_s = 1 \).

Suppose this is the only information we have. Then, in order to obtain an unbiased estimate for \( p_s \) we have to maximise \( S \) under the condition \( \sum s p_s = 1 \). The standard way to take the additional condition into account is the technique of Lagrangian multipliers. We calculate the maximum of the quantity

\[
S + \lambda (\sum s p_s - 1).
\]

The second term in this expression is the additional condition. \( \lambda \) is the Lagrangian multiplier. It is a number that will be determined later such that the condition is fulfilled. If the condition is fulfilled, the additional term vanishes. Maximising \( S \) or the above expression is thus equivalent. We calculate the maximum by taking the derivative of that quantity with respect to \( p_s \) and setting it to zero. This yields the condition

\[
\ln p_s + 1 - \lambda = 0.
\]

This equation is fulfilled if all \( p_s \) are the same. We determine \( \lambda \) such that the condition \( \sum s p_s = 1 \) is fulfilled. The result is \( p_s = \frac{1}{n} \). Indeed, if nothing is known about the events, this is the best unbiased estimate for the probabilities. All events occur with the same probability.

In a more realistic and interesting situation like the distribution of a quantity like DSO, additional conditions have to be taken into account. We have more data available and the estimate should take the information from the additional data into account. Solving this problem is a bit more complicated than the trivial example, but the principle of calculation is the same. For each additional condition, we introduce an additional Lagrangian multiplier and add it to the above expression. In the end, we determine the Lagrangian multipliers such that all conditions are fulfilled.

To really apply the maximum entropy method to business data we need two extensions.

The first is conditional entropy. Instead of bare probabilities \( p_s \) used above, in most cases we look at conditional probabilities. In our above example, we included the condition that the payment target is 30 days. In fact, conditional probabilities are the more fundamental concept than bare probabilities because in reality there are always additional conditions. Unfortunately, in many cases they are not made explicit, they are often not mentioned, or we are not even aware of them. And implicit condition in the above data is that they stem from a certain company. Stating that condition may be important because the results of a calculation based on those data are eventually not applicable to another company, even if it belongs to the same enterprise.

The second is latent variables. It may be that the distribution is influenced by some unknown quantity, a latent variable. One can extend the maximum entropy principle to include such a latent variable. A typical example would be some seasonal effect. Since the latent variable is unknown, we do not know the distribution
of it. As a consequence, the information contained in the latent variable has to be extracted from the same data as the information that yields the distribution of the quantity we are interested in. The resulting equations are implicit equations which are difficult to solve. The numerical effort is a lot higher than for the usual maximum entropy method.

In [6, 7] we showed that the conditional maximum entropy method with latent variables included can describe distributions of relevant business quantities quite well and yields indeed meaningful results.

5 Some history

Maximum entropy methods are a part of Bayesian statistics. Although named after Thomas Bayes, who lived in the 18th century, the term Bayesian statistics is used only since the 1950s. The concept of maximum entropy was formulated in that time as well [8], although it existed before in statistical physics. Most of the methods summarised under the term Bayesian statistics need much more computational effort than usual statistics. Therefore, Bayesian statistics became increasingly popular only after sufficient computational power has been available. This also holds for maximum entropy methods. Most of the research has been done in the past 30 years.

Today, most statistics packages provide well tested tools for Bayesian statistics, including maximum entropy methods. In Trufa, we use R [10], for which packages implementing maximum entropy methods existed for a while on the CRAN [11].

References

This communication contains general information only not suitable for addressing the particular circumstances of any individual case and is not intended to be used as a basis for commercial decisions or decisions of any other kind. None of Deloitte GmbH Wirtschaftsprüfungsgesellschaft or Deloitte Touche Tohmatsu Limited, its member firms, or their related entities (collectively, the “Deloitte network”) is, by means of this communication, rendering professional advice or services. No entity in the Deloitte network shall be responsible for any loss whatsoever sustained by any person who relies on this communication.

Deloitte refers to one or more of Deloitte Touche Tohmatsu Limited, a UK private company limited by guarantee (“DTTL”), its network of member firms, and their related entities. DTTL and each of its member firms are legally separate and independent entities. DTTL (also referred to as “Deloitte Global”) does not provide services to clients. Please see www.deloitte.com/about for a more detailed description of DTTL and its member firms.

Deloitte provides audit, risk advisory, tax, financial advisory and consulting services to public and private clients spanning multiple industries; legal advisory services in Germany are provided by Deloitte Legal. With a globally connected network of member firms in more than 150 countries, Deloitte brings world-class capabilities and high-quality service to clients, delivering the insights they need to address their most complex business challenges. Deloitte’s approximately 286,000 professionals are committed to making an impact that matters.