Avoiding overfitting of models: an application to research data on the Internet videos

Radim Jiroušek [2,1], Iva Krejčová[2]

Abstract. A *model overfitting* is a well-known phenomenon both in statistics and machine learning. In this paper, we study this problem from the perspective of information theory. In this context, data-based model learning can be viewed as a transformation process; a process transforming the information contained in data into the information represented by a model. The overfitting of a model often occurs when one considers an unnecessarily complex model, which usually means that the considered model contains more information than the original data. Thus, using one of the basic laws of information theory saying that any transformation cannot increase the amount of information, we get the basic restriction laid on models constructed from data: A model is acceptable if it does not contain more information than the input data file. This idea is also in agreement with the *minimum description length* principle that, roughly speaking, advices to prefer models described with a small number of parameters to more complex models.

Keywords: Data-based learning, probabilistic models, composition, information theory, MDL principle, overfitting, lossless encoding.

JEL classification: C52
AMS classification: 90B60

1 Introduction

It is perhaps unnecessary to explain in detail what is understood by overfitted models in AI [1] and/or in statistics [14]. Just recall that the notion is connected with models constructed from data, in particular, with the models reflecting noninformative properties of the source data files (like noise and other random properties that each randomly generated data file possesses). This phenomenon is often illustrated on two stochastically dependent variables, the dependence of which is linear. Because the dependence is stochastic, if randomly generated data are plotted in a graph, the respective dots are concentrated along a straight line describing the dependence. Naturally, only a part of them lies on the line. If one tries to find a curve that connects all the dots in the plot (see Fig. 1), the model is much less informative and cannot be used for prediction (neither for interpolation nor for extrapolation). It is important to realize that such a complex curve must be described (defined) by a much larger number of parameters than the straight line, which can be determined just by two points.

Going back to the ideas of von Mises [12] and Kolmogorov [9], who both explored relations interconnecting randomness, complexity, and information, we can learn that they were interested (among others) about "the quantity of information conveyed by an individual object 'x' about another individual object 'y' " [9]. Having two sequences $\mathcal{S}_1$, $\mathcal{S}_2$ of 0's and 1's, which are both lossless encoding of a considered model $\mathbf{M}$, we can thus deduce that both these sequences $\mathcal{S}_1, \mathcal{S}_2$ convey the same amount of information about model $\mathbf{M}$. The same holds true also for an optimum lossless encoding $\mathcal{S}^*$ of model $\mathbf{M}$. Since the mutual information between two objects is always less or equal the information contained in any of these objects, the length of encoding $\mathcal{S}^*$ is the best lower estimate of the amount of information contained in model $\mathbf{M}$ we have. Assuming also that there is no (relative) redundancy in sequence $\mathcal{S}^*$, we can take its length as an estimate of the amount of information (measured in bits) contained in model $\mathbf{M}$ (in what follows we will omit the word estimate, and will speak about the information, or amount of information contained in $\mathbf{M}$).

The above-presented ideas are independent of the type of considered models. The best model containing all the information contained in data is the respective data file itself. Therefore, using the above ideas, the amount of information in data equals the number of bits necessary to store an optimum lossless encoding of the respective data file. This enables us to compare the amount of information contained in data and that contained in a datalearned model. In case we get a model with a greater amount of information than that in data, we are sure, that some undesirable information has been added into the model. In addition to this, we know that regardless the way the data were collected, they always contain some specific part of the information, employment of which results in the overfitting of the model. It should not be included in the model. Therefore, all the considered models should contain less information than the input data. Thus we enforce a principle under which models with the amount

[1] Inst. of Information Theory and Automation, Acad. of Sciences, Prague, Czech Republic, radim@utia.cas.cz
[2] Faculty of Management, University of Economics, Jindřichův Hradec, Czech Republic, iva.krejcova@gmail.com

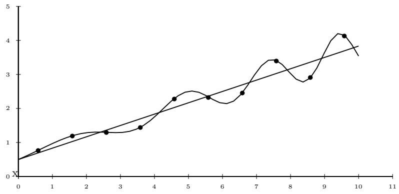

Figure 1 Overfitted linear dependence

of information greater or equal than that of the input data file are *unacceptable*. In fact, we accept only models containing *substantially* less information than the input data file. The meaning of the word *substantially* is usually left to the user's discretion.

Notice, that the above-mentioned principle is also fully compatible with the famous *Minimum Description Length* (MDL) principle that is often used in the process of model learning. For example, it was proposed for Bayesian network learning by Lam and Bacchus [11], (for general sources of this principle see e.g. [3], and [4]). For more discussion on this relationship see also [8].

The goal of this paper is to illustrate the above-described principle with the data-based learning of compositional models. Thus, in the next section, we introduce a minimum quota of concepts from compositional model theory necessary to describe these models and their lossless encodings. Section 3 is devoted to an example in which the approach is applied to data from a commercial research project *Video on Internet* performed by Nielsen Admosphere.

2 Compositional Models and their Encodings

By a compositional model, we understand a multidimensional probability distribution composed of a system of low-dimensional distributions (usually, marginals of the considered multidimensional one). From the viewpoint of this paper it is important to say that we consider discrete (finite-valued) variables $N = \{X_1, X_2, \ldots, X_n\}$; $\mathbb{X}_i$ denote the nonempty finite set of values of variable X_i. For a probability distribution $\pi(N)$, its marginal distributions for variables $M \subseteq N$ will be denoted either $\pi(M)$, or $\pi^{\downarrow M}$, and $\mathbb{X}_M$ denote the state-space of all the combinations of values of the included variables.

The most important concept from the theory of compositional models is an operator of composition $\triangleright$ that from two distributions, say $\kappa(K)$ and $\lambda(L)$ ($K, L \subseteq N$), constructs a more-dimensional distribution (denote it μ) of variables $K \cup L$

$$\mu(K \cup L) = \kappa(K) \triangleright \lambda(L) = \frac{\kappa(K) \cdot \lambda(L)}{\lambda(K \cap L)}. \tag{1}$$

In this paper, we do not need to care about the fact that the composition is not always defined, let alone what are its theoretical properties. For this, the reader is referred to survey papers [6, 7]. Here, we only have to know what are the parameters uniquely defining a compositional model.

Definition 1. Distribution $\pi(N)$ is a *compositional model* if there exists a cover $K_1, K_2, \ldots, K_m$ of N (i.e., $K_1 \cup \ldots \cup K_m = N$), such that[1]

$$\pi(N) = \pi^{\downarrow K_1} \triangleright \pi^{\downarrow K_2} \triangleright \ldots \triangleright \pi^{\downarrow K_m}. \tag{2}$$

It means that to define a compositional model it is enough to determine an ordered sequence of (marginal) distributions. Looking for its optimum lossless encoding we have to encode for each marginal distribution $\pi^{\downarrow K_i}$ the set of variables K_i and the respective probabilities. Realize that, while probability distribution $\pi(N)$ is defined with the help of $\prod_{i=1}^{n} |\mathbb{X}_i| - 1$ probabilities, to properly define compositional model (2) we need to specify only its marginals $\pi^{\downarrow K_i}$, i.e. we have to specify $\sum_{k=1}^{m} \left(\prod_{i \in K_k} |\mathbb{X}_i| - 1 \right)$ probabilities.

It is perhaps unnecessary to say that speaking about an optimum encoding of models is an unattainable idealization. In practical situations, solving such an optimization problem would be intractable and therefore we have

[1] The operator of composition is not associative, the expression (2) is evaluated from left to right, i.e.,

$$\pi^{\downarrow K_1} \triangleright \pi^{\downarrow K_2} \triangleright \ldots \triangleright \pi^{\downarrow K_m} = \left(\ldots \left((\pi^{\downarrow K_1} \triangleright \pi^{\downarrow K_2}) \triangleright \pi^{\downarrow K_3} \right) \triangleright \ldots \triangleright \pi^{\downarrow K_{m-1}} \right) \triangleright \pi^{\downarrow K_m}.$$

to accept a suboptimal solution. This problem was studied in [8], where five different encoding approaches were compared. It was shown that none of the proposed approaches dominated others. So, for practical applications, we suggest the following solution: Consider a battery of encoding procedures and define the amount of information contained in data/model equal to the minimal length from all the binary encodings achieved by the considered approaches. For the sake of simplicity, we use in this paper only two encoding approaches: Direct Data Encoding which appears to be felicitous for the considered input data file encoding, and Huffman Lexicographic Encoding apt for the encoding of models (more precisely for the encoding of its building stones - marginal distributions). Before describing these encoding approaches, let us stress that it is of great importance to properly select the precision with which the respective probabilities are specified. This is because, as we will see below, the encoding of probabilities takes the substantial part of the whole code (in particular for Huffman Lexicographic Encoding).

Taking into account the precision, with which the probabilities are specified, is fully sensible also from the statistical point of view. Due to the accepted principle, the less data we have, the less amount of bits we may use to encode the model. It means, among others, that for small data files we cannot consider probability values specified with a high precision. This fully corresponds with the fact that having a small number of data, the confidence intervals for the estimates of probability parameters are rather wide. Therefore it does not have the sense to specify these estimates with a high precision, with a great number of digits.

To make the description of the encoding procedures as simple as possible, assume that all the values of all considered variables are nonnegative integers: $\mathbb{X}_i = \{0, 1, \ldots, |\mathbb{X}_i| - 1\}$.

Direct Data Encoding. We will use this type of encoding to encode a source data file. Consider a record $(x_1, x_2, \ldots, x_n)$ from a data file. It means that $x_i \in \mathbb{X}_i = \{0, 1, \ldots, |\mathbb{X}_i| - 1\}$. Therefore, each record can be unambiguously represented by the integer[2]

$$\left(\sum_{k=1}^{n-1} x_k \prod_{j=k+1}^{n} |\mathbb{X}_i| \right) + x_n,$$

which is an nonnegative integer less than $|\mathbb{X}_N| = \prod_{j=1}^{n} |\mathbb{X}_i|$, and therefore it can be encoded into[3] $\log_2 \lceil |\mathbb{X}_N| \rceil$ bits.

Huffman Lexicographic Encoding. To encode a low-dimensional probability distribution we have to specify the ordered sequence of its arguments (i.e., the respective variables), and the respective probabilities in a predefined ordering. To control the precision, with which the probabilities are specified, we determine probabilities of a low-dimensional distribution as a ratio $\frac{\ell}{base}$, where $base$ is a properly determined positive integer. The highest reasonable precision is to set $base$ equal to the size of a source data file. Decreasing this integer we decrease the precision of the specified probabilities. Setting $base = 1000$ means that we take all the probability estimates with three digits of precision. Rounding these estimates to two decimal digit means to consider $base = 100$. Nevertheless, it is important to realize that we can consider any number less or equal size of the data file at our disposal. Specifying the integer $base$ for the respective low-dimensional distribution, the probabilities are determined by integers, numerators of the respective ratios. To encode these integers we will apply the famous Huffman's encoding algorithm [5] (the reader not familiar with this encoding technique will see its application in Section 3).

Before proceeding to the illustrative example of the application of the proposed principle, let us remark that for the considered compositional models there are principally two ways of model simplification, which can be applied when getting unacceptably complex models:

- *Structure simplification* means to consider smaller sets of variables K_i.
- *Considering lower precision* of probabilities means to decrease the constant base (which should always be considered especially for distributions with higher dimensions).

[2] Such an integer is in fact an order number of the combination $(x_1, x_2, \ldots, x_n)$ in the ordering

$$(0, \ldots, 0, 0), (0, \ldots, 0, 1), \ldots, (0, \ldots, 0, \mathbb{X}_n - 1), (0, \ldots, 1, 0), \ldots, (\mathbb{X}_1 - 1, \ldots, \mathbb{X}_{n-1} - 1, \mathbb{X}_n - 1).$$

[3] $\lceil r \rceil$ denotes the smallest integer, which is not less than r.

3 Example

Because of the lack of space, we cannot describe the model learning principles in more details. Perhaps the best way how to clarify not-well-declared steps is to describe the process by an example. Therefore, in this section we consider a compositional model learning from data acquired in the framework of a commercial research *Video on the Internet* realized in 2016 by Nielsen Admosphere. This research project was oriented to answer questions concerning what and how often the respondents watch different types of videos on the Internet. The data were collected from 1207 respondents from among the Czech Internet population in the age of 15-83. From a large number of questions, for the purpose of this paper, we selected only 24 variables concerning demographic characteristics of respondents (5 variables: *age, sex, education, region,* and *size-of-municipality*), what type of electronic device they use (variables: $Q1$ - $Q7$), frequency and type of watched videos (variables $V1$ - $V6$, $T1$ - $T6$). Among them[4]

9 variables are dichotomic (2-valued) variables: $sex, Q1, \ldots, Q7, T6$,
9 variables are trichotomic (3-valued) variables: $edu, region, size, V1, \ldots, V6$,
6 variables have 4 values: $age, T1, \ldots, T5$.

Having these 24 variables, each combination of their values, i.e., each record from the input data file, can be uniquely encoded as a nonnegative integer less than

$$2 \times 4 \times 3 \times 3 \times 3 \times \underbrace{2 \times \ldots \times 2}_{7} \times \underbrace{3 \times \ldots \times 3}_{6} \times \underbrace{4 \times \ldots \times 4}_{5} \times 2 = 41\ 278\ 242\ 816.$$

Among the 1207 records of the input data file only 8 records occur twice; it means that there is no way to find a substantially more efficient encoding of the data file than that encoding each record as a 36-digit binary number[5]. Therefore, we start the model learning process with an initial limitation given by the fact that the data file "contains the information" of $1207 \times 36 = 43\ 452$ bits.

In the data-based process of model learning, we look for groups of variables defining the model (more precisely, the respective model is defined by an ordering of distributions defined for these groups of variables). Naturally, the goal is to group together variables that are highly interconnected. The strength of the interconnection is measured by the *measure of interdependence* (sometimes also called *information content*), which simplifies to the well-known mutual information for two-dimensional distributions [13]

$$MI(\pi(N)) = \sum_{(x_1,\ldots,x_n)\in\mathbb{X}_N} \pi(X_1 = x_1, \ldots, X_n = x_n) \log_2 \left(\frac{\pi(X_1 = x_1, \ldots, X_n = x_n)}{\pi(X_1 = x_1) \cdot \ldots \cdot \pi(X_n = x_n)} \right).$$

There are several strategies how to select an appropriate system of variable groups. It is beyond the scope of this paper to discuss them. What is important that we get groups forming a cover of the considered set of variables (each variable is included at least in one group), and that the measure of interdependence for variables in each group is high. The length of an efficient encoding of all the respective marginal distributions defines the "amount of information" contained in the model, which should be kept, in agreement with the introduced principle, sufficiently below the limit given by the "amount of information" contained in the data file.

Let us explain on the example of five-dimensional distribution how we compute the length of its binary encoding using the Huffman Lexicographic Encoding. Consider five-dimensional distribution of the demographic variables $\pi(age, sex, edu, region, size)$. The size of the respective five-way frequency table is $4 \times 2 \times 3 \times 3 \times 3 = 216$. When going to encode such a marginal distribution, first what we have to do is to choose the precision with which the probabilities will be specified, i.e., to specify the above mentioned constant *base*. Let us consider the highest reasonable precision for the given data file by setting $base = 1207$. Then we get the five-way frequency table (it is too big to present it here), in which we can see that the most frequent entry is 0; it occurs 36-times. The highest entry is 42, which, not surprisingly, appears in the table only once. The whole summary of values (frequencies) appearing in the considered five-way frequency table is in Table 1. Using a block code (fixed length code) for numbers from 0 to 42 we would need six bits to encode each entry. However, since some entries are much more frequent than others, famous Huffman variable length code [5] is more advantageous. It is known that this code is in a way optimal, and it assigns shorter code-words to more frequent entries. Moreover, the reader can see from Table 1 that the resulting code is a *prefix-free* code, which means that no code-word is a prefix of another code-word. To compute the length of the five-way frequency table encoding we need the lengths of individual code words. Thus to encode the frequency table corresponding to $\pi(age, sex, edu, region, size)$ we need

$$3\times(36+28+24+23)+4\times(17+15+13+8)+5\times(8+6+5+5+5)+6\times(4+3)+7\times(2+2)+8\times(\underbrace{1 \times \ldots \times 1}_{12}) = 856$$

[4] For the sake of simplicity, we clustered the values of variables to decrease the total number of values. For example, T and V variables have originally 6 and 7 values, respectively.

[5] $36 = \lceil \log_2(41\ 278\ 242\ 816) \rceil = \lceil 35, 3 \rceil$

frequency	number of instances	Huffman code	length of code
0	36	100	3
2	28	110	3
1	24	000	3
3	23	010	3
5	17	1010	4
4	15	1110	4
7	13	0010	4
6	8	0110	4
8	8	10110	5
11	6	00111	5
10	5	00110	5
12	5	01110	5
13	5	01111	5
9	4	101110	6
14	3	111100	6
15	2	1011110	7
19	2	1011111	7
16	1	11110100	8
17	1	11110101	8
18	1	11110110	8
20	1	11110111	8
24	1	11111000	8
25	1	11111001	8
27	1	11111010	8
30	1	11111011	8
32	1	11111100	8
33	1	11111101	8
39	1	11111110	8
42	1	11111111	8

Table 1 Frequencies and their Huffman code

bits. Therefore, using Huffman code we need (in average) just $856/216 = 3.96$ bits to encode one entry of the considered five-way frequency table, i.e., one probability of the respective marginal probability distribution. Notice that we do not need to encode the respective denominator *base* because it can be got as a sum of all numbers from the frequency table. But we must not forget that we have to encode the respective Huffman code (otherwise nobody could decode it!). It means we have to encode three columns of Table 1: frequency, length of code, and Huffman code. For this we need:

11 bits to encode the maximum frequency;
29×6 bits to encode the frequencies;
6 bits to encode the maximum length of code word;
29×6 bits to encode the lengths of code words;
175 bits to encode the codewords.

Adding another 30 bits necessary to encode variables, for which the marginal distribution is defined, the described encoding needs 1426 bits to fully represent probability distribution $\pi(age, sex, edu, region, size)$. If we chose lesser precision of probabilities, say $base = 500$ (or $base = 200$), which means that we consider probabilities with the precision of 0.002 (0.005) we would need only 965 (550) bits to represent this distribution. Notice however that, if we considered five-dimensional marginal for four-valued variables $\pi(T1, \ldots, T5)$ with the highest reasonable precision $base = 1207$ we would need approximately four kilobits. Taking into account the fact that all reasonable models for the considered 24-dimensional distribution consisted of 18 – 21 marginal distributions, one can see that incorporating such space demanding marginals into the model would lead to models from our point of view unacceptable. Thus, for example, all five-dimensional distributions contained in the constructed *acceptable* models had always at least one binary variable among their arguments.

Recall that in the previous paragraph, we saw two ways how to keep learned models reasonably small: either we have to keep a simple structure, which means for compositional models that we keep the dimension of the considered marginals limited, or that we decrease the precision of probabilities. The trade-off between these two possibilities is fully under the control of the user. At this stage of research, we do not have any heuristics that could help them.

4 Conclusions

To avoid the well-known phenomenon of overfitting, different techniques for model verification and testing are usually used; starting with splitting the data into two parts: one part used for model learning and the other one for its testing (which decreases the amount of teaching data, though), or popular cross-validation technique. The common drawback of all these approaches is that they hardly ever make possible to distinguish the situations of overfitting from insufficient learning due to the lack of data (also called underfitting). Naturally, we do not claim the described principle is a general technique solving all such problems, but it can be used as one of stopping rules in the process of model learning. Its simplicity guarantees it can be used practically in all machine learning tasks. And what is important, its careful application makes the machine learning specialists realize which parts of the models are space demanding. In connection with probabilistic models, it is also important to realize that simplification can be achieved (at least) in two different ways: by simplification of the structure of constructed models, or by the roughening of the estimates of probabilistic parameters.

Acknowledgements

The authors want to express their gratitude to Nielsen Admosphere, a.s. for providing the research data free of charge. The research was partially supported by GAČR under Grant No. 15-00215S.

References

[1] Berka, P.: *Dobývání znalostí z databází. (Knowledge Discovery in Databases),* Academia, Praha 2003.

[2] Good, P. I., Hardin, J. W.: *Common Errors in Statistics (And How to Avoid Them),* Wiley, 2006.

[3] Grünwald, P.: A Tutorial Introduction to the Minimum Description Length Principle. [online]. 2004, p. 80 [cit. 2014-07-15]. Available from: http://eprints.pascal-network.org/archive/00000164/01/mdlintro.pdf

[4] Hansen, M. H., Yu, B.: Minimum Description Length Model Selection Criteria for Generalized Linear Models. [online]. p. 20. Available from: http://www.stat.ucla.edu/ cocteau/papers/pdf/glmdl.pdf

[5] Huffman, D. A.: A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the I.R.E.* (September 1952), 10981102.

[6] Jiroušek, R.: Foundations of compositional model theory. *Int. J. General Systems* **40** (2011), 623–678.

[7] Jiroušek, R., Kratochvíl, V.: Foundations of compositional model theory: structural properties. *Int. J. General Systems* **44** (2015), 2–25.

[8] Jiroušek, R., Krejčová I.: Minimum Description Length Principle for Compositional Model Learning. In: *Integrated Uncertainty in Knowledge Modelling and Decision Making. Proceedings of IUKM 2015* (Van-Nam Huynh, Masahiro Inuiguchi, and Thierry Denoeux, eds.), LNAI, Vol.: 9376 (2015), 254–266.

[9] Kolmogorov, A.N.: Tri podchoda k opredeleniju ponjatija kolichestvo informacii. *Problemy Peredachi Informatsii*, **1** (1965), 3–11 [Engl. transl. accessible at http://alexander.shen.free.fr/library/Kolmogorov65_Three-Approaches-to-Information.pdf]

[10] Kullback, S., Leibler, R. A.: On information and sufficiency. *Annals of Mathematical Statistics* **22** (1951), 76–86.

[11] Lam, W., Bacchus, F. Learning Bayesian Belief Networks: An approach based on the MDL Principle. *Computational intelligence* **10** (1994), 269–293. [Available from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.127.5504&rep=rep1&type=pdf]

[12] Von Mises, R.: *Probability, statistics, and truth.* Courier Corporation, Mineola, New York, 1957 [Originaly published in German by Springer, 1928]

[13] Rényi, A.: In: *On measures of information and entropy. Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability*, (1960), 547–561.

[14] Ryan, T.: *Modern Regression Methods,* Wiley, 2009.