A Bayesian Information Criterion for Unsupervised Learning Based on an Objective Prior

Ildar Baimuratov[1], Yulia Shichkina[2(✉)], Elena Stankova[3], Nataly Zhukova[1,4], and Nguyen Than[1]

[1] ITMO University, St. Petersburg, Russia
baimuratov.i@gmail.com, nazhukova@mail.ru
[2] St. Petersburg State Electrotechnical University, St. Petersburg, Russia
strange.y@mail.ru
[3] St. Petersburg State University, St. Petersburg, Russia
[4] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, St. Petersburg, Russia

Abstract. Data processing techniques, such as mathematical formulas, statistical methods and machine learning algorithms, require a set of tools for evaluating knowledge extracted from data. In unsupervised learning it is impossible to use referential or predictive estimation. Therefore, the only reliable way to evaluate results of unsupervised learning is information estimation. Unfortunately, information estimation suffer from underfitting and overfitting. We propose a new method for evaluating unsupervised learning results, which is based on the Bayesian criterion for optimal decision and an objective prior probability distribution of partitions. We illustrate the proposed method application on Fisher's iris data set by comparing original label distribution with results of clustering with different numbers of clusters. We show the method prevents underfitting and overfitting and verify it by comparing the recommended value with posterior distribution.

Keywords: Unsupervised learning · Information estimation · Bayesian criterion · Objective prior

1 Introduction

Data processing techniques, such as mathematical formulas, statistical methods and machine learning algorithms, require a set of tools for evaluating knowledge extracted from data. There are several ways to perform such estimation. Based on the research presented in [1], we consider the following estimation methods: referential, predictive and informational.

In unsupervised learning it is impossible to use referential or predictive estimation. Referential methods require some reference model which is believed to

S. Misra et al. (Eds.): ICCSA 2019, LNCS 11619, pp. 707–716, 2019.
https://doi.org/10.1007/978-3-030-24289-3_52

represent our best guess about the knowledge hidden in data. Predictive methods also have limitation as they require some outer evaluation of predictions. Therefore, the only reliable way to estimate results of unsupervised learning is information estimation.

1.1 Information Measures Underfitting/Overfitting

Unfortunately, information estimation, like many other methods for solving optimization problems, suffer from overfitting. Overfitting is a situation where a model corresponds not only to the relations between variables, but also to random noise in data [2]. A model incompleteness can be caused either by a bias—an error that occurs since the model used is not able to describe the dependencies in the data or by a variance—an error that occurs due to increased sensitivity to noise. Bias and variance, in turn, are related to model complexity—a quantity derived from the type of the model, the amount of input data and the number of parameters. A model that is too simple has high bias and vice versa—a model that is too complex has high variance.

Let us show that informational measures suffer from overfitting. In order to do it consider the Fisher's iris data set [3] and estimate mutual information of two random variables A and B [4], where A is the real class of an instance and B is some unique identifier, with respect to the original probability distribution of instances. This data set contains $N = 150$ instances of $A = 3$ different classes, therefore, $P(n) = \frac{1}{150}$, $P(a) = \frac{1}{3}$ and $P(b) = \frac{1}{150}$. Mutual information $MI(Y, X)$ of two discrete random variables is determined by the formula

$$MI(Y, X) = \sum_x \sum_y P(x, y) \log \frac{P(x, y)}{P(x)P(y)}. \tag{1}$$

Therefore, $MI(A, N) \approx 1,58$ and $MI(B, N) \approx 7,23$.

One can use mutual information to compare a result of classification with a reference model or to estimate the amount of knowledge gained after data processing. We are interested in unsupervised learning, therefore, in the second case. For estimating the amount of extracted knowledge mutual information is used in some algorithms for constructing decision trees. In such algorithms, preference is given to attributes with the highest mutual information. Therefore, for a decision tree of the mentioned data set one should choose the attribute B as the first node, but the resulting model, consisting of 150 classes, would be too complicated. Therefore, it is an example of overfitting.

There are various ways of avoiding overfitting:

- Cross-validation,
- Regularization,
- Prior probabilities,
- Bayes factor et al. [1].

One of the typical methods for solving the problem described above is the regularization of mutual information $MI(X,Y)$ using the entropy $H(X)$. The ratio

$$IGR(X,Y) = \frac{MI(X,Y)}{H(Y)} \tag{2}$$

is called the information gain ratio [9].

However, there are cases, where IGR is not applicable. One case is where Y is completely determined by X, i.e. there is a function $f : X \to Y$. In this case $H(Y|X) = 0$. Then, as

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

it holds that,

$$H(Y|X) = 0 \Rightarrow MI(Y,X) = H(Y) \tag{4}$$

therefore,

$$H(X|Y) = 0 \Rightarrow IGR(X,Y) = 1 \tag{5}$$

Thus, if there is a function $f : X \to Y$, IGR is useless. Classification task is indeed a search of a function for assigning labels to data, therefore, IGR is useless for estimating classification in unsupervised learning. The example described above is an example of classification.

1.2 Objective Priors

In this paper we suggest to consider objective priors for preventing overfitting/underfitting. A prior probability distribution of a random variable is a distribution that characterizes its value before obtaining experimental data. There is informative and uninformative, or objective, prior distributions. An uninformative, or objective, distribution expresses vague or general information about a variable and has the following advantages:

- invariance with respect to parameters structure;
- inverse dependence on model complexity;
- independence from subjective assumptions.

Nowadays, applications of objective priors in regression and classification tasks are actively researched [5–8].

An example of objective prior usage is adjusted mutual information [10]. The idea of adjusted mutual information is to adjust mutual information of two random variables with joint probability distribution of their partitions. Suppose there are N points, two partitions U and V, and the number of points $a_i = |U_i|$ for $U_i \subseteq U, i = 1...R$ and $b_j = |V_j|$ for $V_j \in V, j = 1...C$, then the total number of ways to distribute the set N over the two partitions U and V is Ω

$$\Omega = \frac{(N!)^2}{\prod_i a_i! \prod_j b_j!} \tag{6}$$

Every two joint partitions U and V can be represented as a contingency table M

$$M = [n_{ij}]_{j=1\ldots C}^{i=1\ldots R} \tag{7}$$

Suppose there is some contingency table M, then there are w different ways to distribute points so that this M is obtained

$$w = \frac{N!}{\prod_i \prod_j n_{ij}!} \tag{8}$$

Thus, the probability $P(M|a,b)$ for some M with respect to the set $\mathcal{M}$ of all possible contingency tables is determined by the formula

$$P(M|a,b) = \frac{w}{W} \tag{9}$$

Mutual information $MI(M)$ of the contingency table M is determined by the formula

$$MI(M) = \sum_i \sum_j \frac{n_{ij}}{N} \log \frac{N\dot{n}_{ij}}{a_i b_j} \tag{10}$$

then the average mutual information of all possible joint partitions of random variables X and Y is defined as the expectation of mutual information $E(MI(M)|a,b)$

$$E(MI(X,Y)) = E(MI(M)|a,b) = \sum_{M\in\mathcal{M}} MI(M)P(M|a,b) \tag{11}$$

Finally, adjusted mutual information $AMI(X,Y)$ is defined as follows

$$AMI(X,Y) = \frac{MI(X,Y) - E(MI(X,Y))}{max(H(X),H(Y)) - E(MI(X,Y))} \tag{12}$$

Thus, hypergeometric distribution of two joint partitions is used in adjusted mutual information as an objective prior. The most natural way to use adjusted mutual information is to estimate clustering result. But again it estimates similarity of a resulting partition to a reference partition, while we are interested in a method for estimating knowledge gain.

2 A Bayesian Information Criterion Based on an Objective Prior

In this paper we propose a new method for information estimation, intended to estimate knowledge, gained in result of data processing, that is useful in unsupervised learning. The method is applicable to functionally dependent random variables and based on the Bayes criterion for optimal decision and an objective prior probability distribution of partitions.

2.1 Basic Definitions

Suppose there is a function $f : X \to Y$, then for every $y_i \in Y$ there is an inverse image $X_i \subseteq X$

$$X_i = \{x : f(x) = y_i\} \tag{13}$$

We consider partition $part(X) = X_1 \cup ... \cup X_k$ as a resulting model of X, knowledge gain of which we are going to evaluate. A structure of a partition is described by a number of subsets k and a partition $\bar{n}$ of a number $n = |X|$, such that $n_i = |X_i|$ and $n_1 + ... + n_k = n$.

We use relative entropy, or Kullback—Leibler divergence [11], for basic information gain estimation

$$D_{KL}(X||Y) = \sum_i P(x_i) \log \frac{P(x_i)}{Q(x_i)} \tag{14}$$

Considering $P(x_i) = \frac{|X_i|}{|X|}$ and $Q(x_i) = \frac{1}{|X|}$, we get

$$D_{KL}(part(X)||X) = \sum_i \frac{|X_i|}{|X|} \log |X_i| \tag{15}$$

But we propose to adjust relative entropy with an objective prior to prevent overfitting/underfitting.

2.2 The Bayesian Criterion for Optimal Decision

Preventing overfitting/underfitting means to find an optimal structure of a model. In decision theory various criteria are used to find the optimal choice. Given a set of all possible actions A, a set of states of nature Θ, its probability distribution $P(\theta)$ and a utility function $U(a, \theta)$ for an action $a \in A$ and a state of nature $\theta \in \Theta$, consider the Bayes criterion for choosing an optimal action a^* [12]

$$a_b^* = \underset{i}{\operatorname{argmax}} \sum_j P(\theta_j) U(a_i, \theta_j) \tag{16}$$

Given a set X, we consider the set of all possible functions X^X on X, which forms the set of all possible partitions $Part(X)$ of X. Every partition $part(X) \in Part(X)$ has a certain structure $k_i, \bar{n}_j$. Assuming $A = \{k\}$, $\Theta = \{\bar{n}\}$ and $U(a_i, \theta_j) = D_{KL}(part_{ij}(X)||X)$, we get expected relative entropy $E(D_{KL}(part_k(X)||X))$, ED_{KL} for short, of a partition $part_k(X)$ for given number of subsets k

$$E(D_{KL}(part_k(X)||X)) = \sum_j P(\bar{n}_j) D_{KL}(part_{kj}(X)||X) \tag{17}$$

and the Bayesian criterion for optimal number of subsets k_b^*

$$k_b^* = \underset{i}{\operatorname{argmax}} \sum_j P(\bar{n}_j) D_{KL}(part_{ij}(X)||X) \tag{18}$$

It only requires to define the probability distribution $P(\bar{n})$.

2.3 The Objective Prior Distribution of Number Partitions

Let us define the objective prior distribution $P(\bar{n})$ for Bayes criterion of optimal number of subsets k_b^*. Given a set X, we denote a set of all possible functions on X as $F = X^X$ and a set of functions that result in some number partition $\bar{n_i}$ as $F_i \subseteq F$. Let $|X| = n$ and m is a number of elements of $\bar{n}$ with a given value v, such that $v_1 + ... + v_m = k$, then

$$P(\bar{n}_i) = \frac{|F_i|}{|F|} \tag{19}$$

where $|F| = n^n$ and

$$|F_i| = \frac{n!}{n_1!...n_k!} \frac{n!}{k!(n-k)!} \frac{k!}{v_1!...v_m!} \tag{20}$$

Objective prior $P(\bar{n}_i)$ and, therefore, the criterion k_b^* is defined for every set X.

2.4 Other Criteria

There are other criteria in decision theory. If probability distribution $P(\theta)$ is unknown, one can assume states of nature Θ to be equiprobable

$$a_l^* = \underset{i}{\operatorname{argmax}} \sum_j \frac{1}{|\Theta|} U(a_i, \theta_j) \tag{21}$$

or minimize possible loss

$$a_w^* = \underset{i}{\operatorname{argmax}} \min_j U(a_i, \theta_j) \tag{22}$$

Laplace and Wald criteria respectively.

Performing analogous substitutions, we get Laplace criterion for optimal number of subsets k_l^*

$$k_l^* = \underset{i}{\operatorname{argmax}} \sum_j \frac{1}{|\{\bar{n}\}|} D_{KL}(part_{ij}(X)||X) \tag{23}$$

and Wald criterion for k_w^*

$$k_w^* = \underset{i}{\operatorname{argmax}} \min_j D_{KL}(part_{ij}(X)||X) \tag{24}$$

However, the last criterion does not help with underfitting, as for any n it holds that $k_w^* = 1$, because

$$\min D_{KL}(part_1(X)||X) = \max D_{KL}(part(X)||X) = 1 \tag{25}$$

while Laplace criterion k_l^* may be used, if probability $P(\bar{n})$ is too complex to calculate.

3 Evaluating Unsupervised Learning Results

Consider the Fisher's iris data set, mentioned above, to demonstrate the proposed method. We compare the original labels distribution with $k = 3$ with more or less numbers of classes, $k = 2$ and $k = 4$ respectively. We use k-means clustering algorithm for $k = 2, 4$ to compare expected relative entropy of particular partitions. The results are given in the Table 1. As we can see, partitioning a set of 150 objects yields more expected relative entropy for $k = 4$ than for $k = 3$ and $k = 2$.

Table 1. Results of comparing

k	$\bar{n}$	D_{KL}	P	PD_{KL}	ED_{KL}
2	53, 97	0.85	1.20e-281	1.02e-281	5.30e-278
3	50, 50, 50	0.78	4.32e-252	3.37e-252	6.13e-250
4	50, 40, 28, 32	0.72	5.38e-233	3.87e-233	1.15e-229

However, this function has a global maximum, i.e. there is k_b^* such that for $k' > k_b^*$ expected relative entropy decreases. To show that, let us consider another experiment. The idea is to simulate numerous results of various unsupervised classification and clustering instances to get posterior distribution and calculate expected relative entropy for every possible k. Thus, we will get a posterior number of subsets k_p^* with the maximum expected relative entropy and compare it with k_b^*.

Due to complexity issues we consider $n = 75$. The plot of expected relative entropy for $k = 1, ..., 75$ is given at the Fig. 1. As we see, $max(ED_{KL}) = 1.20e-02$ for $k_b^* = 47$.

To simulate posterior k_b^* distribution, we considered various clustering methods that do not require k as an input parameter, such as hierarchical clustering [13], affinity propagation [14], mean shift [15] and DBSCAN [16], but it turned out that their results have subjective bias, caused by hyperparameter values, set by a user, or by an algorithm logic itself, which may differ from objective dependencies beneath data. Therefore, we suggest to assign labels randomly, assuming that for large number of samples s the distribution would be similar to real results.

The result of the experiment is given at the Fig. 2. As we can see, k_p^* comes near to k_b^* and equals to it nearly after 1000 samples. Thus, we demonstrated that prior criterion k_b^* has global maximum and correlates with posterior distribution.

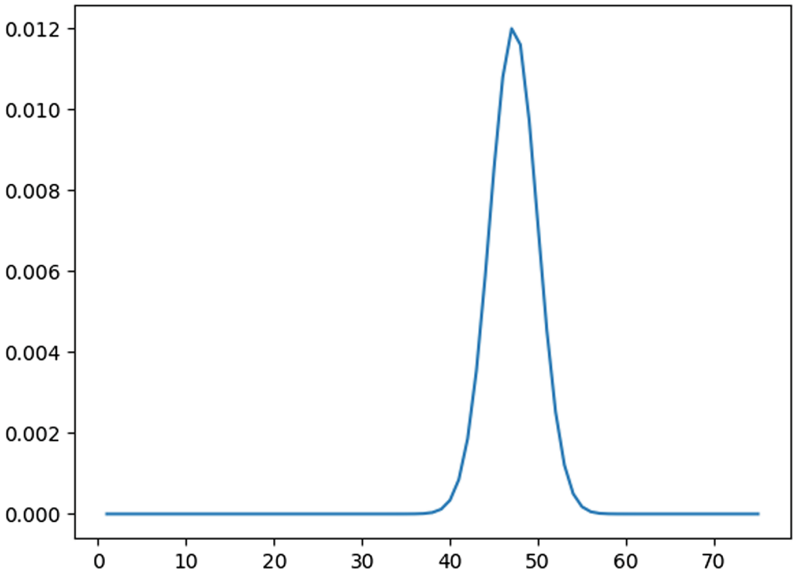

Fig. 1. Expected entropy for $n = 75$ and $k = 1, ..., 75$

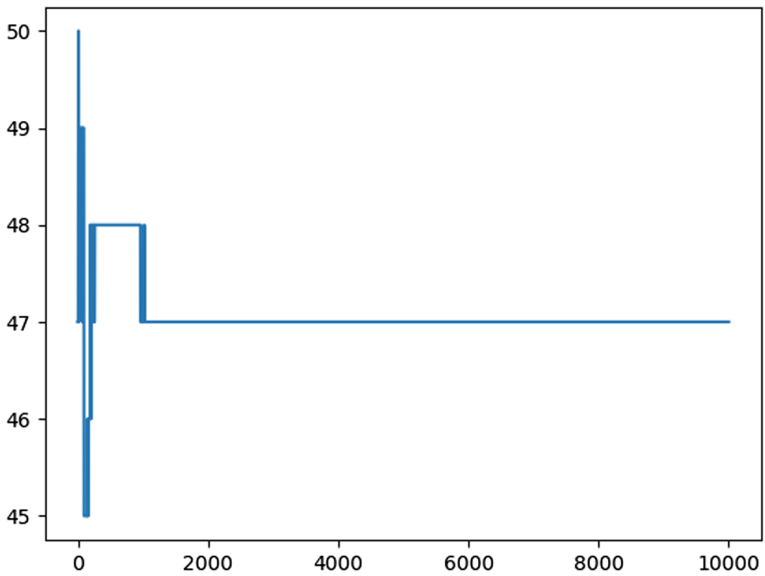

Fig. 2. k_p^* for $n = 75$ and $s = 1, ..., 10000$

4 Conclusion

Summing up, we suggested that information estimation is the only reliable way to evaluate results of unsupervised learning and demonstrated that information measures, like mutual information, suffer from underfitting and overfitting. We considered objective priors as a method for preventing underfitting/overfitting, since they have particular advantages, and proposed a new method for evaluating partition of a set, considered as unsupervised classification or clustering result. It includes, on the one hand, the Bayesian criterion for optimal decision, considering number of subsets as set of actions, corresponding number partitions as states of nature and relative entropy as utility function, and, on the other hand, the objective prior probability distribution of number partitions with respect to a number of all set partitions. We illustrated the resulting criterion application on Fisher's iris data set by comparing original label distribution with results of clustering with different numbers of clusters and demonstrated that the criterion has a global maximum and correlates with posterior distribution for large numbers.

Acknowledgments. The research was funded by RFBR and CITMA according to the research project №18-57-34001 and was funded by RFBR according to the research project №19-07-00784

References

1. Piironen, J., Vehtari, A.: Comparison of Bayesian predictive methods for model selection. Stat. Comput. **27**(3), 711–735 (2017). https://doi.org/10.1007/s11222-016-9649-y
2. Lever, J., Krzywinski, M., Altman, N.: Points of significance: model selection and overfitting. Nat. Methods **13**(9), 703–704 (2016)
3. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**(2), 179–188 (1936)
4. Cover, T.M., Thomas, J.A.: Elements of Information Theory. Wiley, New York (1991)
5. Simpson, D., Rue, H., Riebler, A., Martins, T.G., Sørbye, S.H.: Penalising model component complexity: a principled, practical approach to constructing priors. Stat. Sci. **32**(1), 1–28 (2017)
6. Mattingly, H.H., Transtrum, M.K., Abbott, M.C., Machta, B.B.: Maximizing the information learned from finite data selects a simple model. Proc. Nat. Acad. Sci. U.S.A. **115**(8), 1760–1765 (2018)
7. Palmieri, F.A.N., Ciuonzo, D.: Objective priors from maximum entropy in data classification. Inf. Fusion **14**(2), 186–198 (2013)
8. Sørbye, S.H., Rue, H.: Penalised complexity priors for stationary autoregressive processes. J. Time Ser. Anal. **38**(6), 923–935 (2017)
9. Quinlan, J.R.: Induction of decision trees. Mach. learn. **1**(1), 81–106 (1986)
10. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. J. Mach. Learn. Res. **11**, 2837–2854 (2010)
11. Kullback, S., Leibler, R.A.: On information and sufficiency. Ann. Math. Stat. **22**(1), 79–86 (1951)

12. Berger, J.O.: Statistical Decision Theory and Bayesian Analysis. Springer Series in Statistics, 2nd edn. Springer, New York (1985). https://doi.org/10.1007/978-1-4757-4286-2
13. Ward Jr., J.H.: Hierarchical grouping to optimize an objective function. J. Am. Stat. Assoc. **58**(301), 236–244 (1963)
14. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science **315**(5814), 972–976 (2007)
15. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 603–619 (2002)
16. Ester, M., Kriegel, H.P., Sander, J., Xu, X.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226–231, Portland (1996)