Mathematical model of bank scoring in conditions of insufficient data

Vladimir Mosin[1], *Anton* Abashkin[1*], and *Olga* Yusupova[1]

[1]Samara State Technical University, Molodogvardeyskaya street, 244, 443100, Samara, Russia

Abstract. Recently, different methods of object classification using training datasets is actually. One of these methods is naive Bayesian classifier. Class of objects can consist of low number of elements. Such class is called poor class. In this paper we consider classification problem in poor class. Logical classifier doesn't work in this case. Metric classifier can give good results if and only if there are quite dense set of metrically nearby classified objects in neighborhood of the considering object. Bayesian classifier reevaluates all hypotheses about belonging of the object to certain class. Therefore, Bayesian classifier can solve this classification problem. For example, we considered classic problem of bank scoring. This scoring is based on two criteria. Classified object has two belonging hypotheses. We can apply such reasoning for more difficult cases.

1 Bank scoring

Algorithms of data classification by certain characteristics are actively used in finance. Probabilistic [1-8], logical [9,10] and metric [11-15] classifiers are used. One of them is naive Bayesian classifier. Classifiers are widely used in bank scoring.

Bank scoring is a procedure of evaluating a borrower's credit rating. As a result of this, the bank decides whether to issue a loan to the borrower. In the process of operation, the bank accumulates information about loans previously issued to this or that borrower thus making an extensive data frame that contains values characterizing the borrower. The target function here can show two possible values: 1) the borrower returned the loan, 2) the borrower did not return the loan.

We can image this data frame as table 1.

Table 1. Data frame.

	x_1	x_2	x_3	x_4	…	x_{n-1}	x_n	y
1	a_{11}	a_{12}	a_{13}	a_{14}	…	a_{1n-1}	a_{1n}	b_1
2	a_{21}	a_{22}	a_{23}	a_{24}	…	a_{2n-1}	a_{2n}	b_2
…	…	…	…	…	…	…	…	…
m	a_{m1}	a_{m2}	a_{m3}	a_{m4}	…	a_{mn-1}	a_{mn}	b_m
$m+1$	a_1	a_2	a_3	a_4	…	a_{n-1}	a_n	

* Corresponding author: samcocaa@rambler.ru

Based on the training data sample (the first m rows of the data frame), the probability that the$(m+1)$ borrower will repay the loan is calculated. If this probability is too low, the loan will be refused, if it is high enough, the loan will be issued. The loyalty threshold is set by the bank depending on its credit policy.

In terms of data analysis, bank scoring is a typical classification task. Classification procedures are well studied and described by both Russian (see [1-5]) and foreign researchers (see [6, 7]). The authors of this paper consider two versions of the probability classifier.

2 Classical Bayesian classifier

Bayes formulas are well known as a part of the initial course of the theory of probability. They demonstrate the relationship between a priori and a posteriori probability of a set of hypotheses. More specifically, if H_1, ..., H_r is a set of hypotheses, then after the condition$A$ is satisfied, the new probabilities of the hypotheses are calculated while using the following formulas:

$$P(H_i|A) = \frac{P(H_i)\cdot P(A|H_i)}{P(A)}. \quad (1)$$

If the two conditions A and $B$are satisfied, then the posteriori probabilities of the hypotheses are evaluated similarly:

$$P(H_i|AB) = \frac{P(H_i)\cdot P(AB|H_i)}{P(AB)}. \quad (2)$$

It is clear that there can be many conditions of this kind. In data analysis tasks, meeting a set of conditions means that the features represented in the data frame take on specific values:

$$\{x_1 = a_1\}\cdot\{x_2 = a_2\}\cdot\ldots\cdot\{x_n = a_n\} \quad (3)$$

In this case, the class of objects that meets these conditions may be very poor or even empty. Therefore, the classical Bayesian classifier is rarely used in classification problems.

3 Naive Bayesian classifier

Let us consider the naive Bayesian classifier algorithm for the case of two conditions. First, the classical Bayes formula is applied:

$$P(H_i|AB) = \frac{P(H_i)\cdot P(AB|H_i)}{P(AB)}.$$

The right part contains a fraction, so the left part is proportional to the numerator of the right part. Using the $\propto$symbol to denote proportionality, we get:

$$P(H_i|AB) \propto P(H_i)\cdot P(AB|H_i).$$

Now let us apply the so-called "naive assumption": we assume that the events A and B are independent. Then:

$$P(H_i|AB) \propto P(H_i)\cdot P(A|H_i)\cdot P(B|H_i).$$

Let us perform this operation for all hypotheses and get the proportionality of the sum:

$$P(H_1|AB) + \cdots + P(H_r|AB) \propto P(H_1) \cdot P(A|H_1) \cdot P(B|H_1) + \cdots + P(H_r) \cdot P(A|H_r) \cdot P(B|H_r).$$

The latter relation means an exact equality with some unknown coefficient α:

$$P(H_1|AB) + \cdots + P(H_r|AB) =$$

$$= \alpha\big(P(H_1) \cdot P(A|H_1) \cdot P(B|H_1) + \cdots + P(H_r) \cdot P(A|H_r) \cdot P(B|H_r)\big).$$

The coefficient is determined from the condition

$$P(H_1|AB) + \cdots + P(H_r|AB) = 1,$$

after that, the posteriori probability is calculated:

$$P(H_i|AB) = \alpha \cdot P(H_i) \cdot P(A|H_i) \cdot P(B|H_i).$$

The classical Bayesian classifier makes it possible to re-valuate one single hypothesis without appealing to the posteriori probabilities of other hypotheses, while the naive Bayesian classifier can only re-valuate all hypotheses at once.

In practice, the "naive assumption" is usually not fulfilled: it means that the value of one attribute accepted by an object somehow affects the values of other attributes. Nevertheless, when solving practical problems, the naive Bayesian classifier often shows better results than the classical one.

4 Examples

Example 1. Let us consider a data frame that contains information if the borrower has returned the loan (the Score attribute), as well as information about his gender identity (the Sex attribute) and whether he has a criminal record (the Crime attribute).

The basic information is presented in Table 2, Tables 3 and 4 differ from Table 1 only by permutations of rows, the data there are grouped for perception convenience.

Table 2. Random.

	Sex	Crime	Score	
1	Male	Yes	–	6
2	Male	No	–	8
3	Male	No	+	9
4	Female	Yes	–	11
5	Female	Yes	+	5
6	Female	No	+	14
7	Male	Yes	+	4
8	Female	No	+	13
9	Female	No	+	3
10	Male	Yes	–	15
11	Female	No	–	2
12	Male	No	–	12
13	Female	Yes	–	7
14	Female	Yes	+	1
15	Male	No	+	10
16	Male	Yes	–	16

Table 3. Group by Sex/Crime.

Sex	Crime	Score	
Female	No	+	6
Female	No	+	8
Female	No	+	9
Female	No	–	5
Female	Yes	+	14
Female	Yes	+	3
Female	Yes	–	15
Female	Yes	–	7
Male	No	+	11
Male	No	+	4
Male	No	–	13
Male	No	–	2
Male	Yes	+	12
Male	Yes	–	1
Male	Yes	–	10
Male	Yes	–	16

Table 4. Group by Score.

Sex	Crime	Score
Female	No	+
Female	No	+
Female	No	+
Female	Yes	+
Female	Yes	+
Male	No	+
Male	No	+
Male	Yes	+
Female	No	–
Female	Yes	–
Female	Yes	–
Male	No	–
Male	No	–
Male	Yes	–
Male	Yes	–
Male	Yes	–

Let us calculate the probability of credit repayment by the payer belonging to the class of "convicted male".

A. Classical Bayesian classifier

Let us denote:

$$P_B = P(\mathbf{Score} = " + " \,|\mathbf{Sex} = \text{"Male"}, \mathbf{Crime} = \text{"Yes"}).$$

By the Bayes formula:

$$P_B = \frac{P(\mathbf{Score} = \text{"+"})\text{P (Sex = "Male", Crime = "Yes"| Score = "+")}}{P(\ \mathbf{Sex} = \text{"Male"}, \mathbf{Crime} = \text{"Yes"})}$$

From Table 4 we get:

$$P(\mathbf{Score} = \text{"+"}) = 8/16, \quad \text{and} \quad P\ (\text{Sex = "Male", Crime = "Yes" | Score = "+"}),$$

from Table 3 we get:

$$P(\ \mathbf{Sex} = \text{"Male"}, \mathbf{Crime} = \text{"Yes"}) = 4/16.$$

Therefore,

$$P_B = \frac{8/16 \cdot 1/8}{4/16} = \frac{1}{4} = 0.25.$$

The same result can be obtained directly from Table 3 by separately analysing the last block of four rows.

B. Naive Bayesian classifier

Denote:

$$P_{NB} = P(\mathbf{Score} = \text{"+"} \mid \mathbf{Sex} = \text{"Male"}, \mathbf{Crime} = \text{"Yes"}).$$

Since the algorithm of the naive Bayesian classifier allows us to re-evaluate only all hypotheses at once, we introduce additional notations:

$$P_+ = P_{NB}, \qquad P_- = P(\mathbf{Score} = \text{"–"} \mid \mathbf{Sex} = \text{"Male"}, \mathbf{Crime} = \text{"Yes"}),$$

For the hypothesis P_+, we use the naive Bayesian assumption:

$$P_+ \propto P(\mathbf{Score} = \text{+"}) \cdot \text{P (Sex=Male| Score=+")} \cdot \\ \cdot P\ (\ \mathbf{Crime} = \text{"Yes"}|\ \mathbf{Score} = \text{"+"}).$$

From Table 4 we get the probability of the hypothesis:

$$P(\mathbf{Score} = \text{" + "}) = 8/16,$$

and two conditional probabilities:

$$(P\ (\text{Sex = "Male"| Score = "+"}) = 3/8, \qquad P\ (\ \mathbf{Crime} = \text{"Yes"}|\ \mathbf{Score} = \text{"+"}) = 3/8.$$

Therefore,

$$P_+ \propto \frac{8}{16} \cdot \frac{3}{8} \cdot \frac{3}{8} = \frac{9}{128}.$$

For the hypothesis P_-, we also use the naive Bayesian assumption:

$$P_- \propto P(\mathbf{Score} = \text{"–"}) \cdot \text{P (Sex="Male"| Score="–")} \cdot \\ \cdot P\ (\ \mathbf{Crime} = \text{"Yes"}|\ \mathbf{Score} = \text{"–"}).$$

From Table 4 we get the probability of the hypothesis:

$$P(\textbf{Score} = \text{"–"}) = 8/16,$$

and two conditional probabilities:

$$(P\ (\text{Sex} = \text{"Male"}|\ \text{Score} = \text{"–"}) = 5/8, \qquad P\ (\textbf{Crime} = \text{"Yes"}|\ \textbf{Score} = \text{"–"}) = 5/8.$$

Therefore,

$$P_{-} \propto \frac{8}{16} \cdot \frac{5}{8} \cdot \frac{5}{8} = \frac{25}{128}.$$

We get the proportionality for the sum of the probabilities:

$$P_{+} + P_{-} \propto \frac{9}{128} + \frac{25}{128}.$$

Therefore,

$$P_{+} + P_{-} = \alpha \cdot \frac{34}{128}.$$

The coefficient α is found based on the condition $P_{+} + P_{-} = 1$. We get $\alpha = 128/34$, and

$$P_{+} = \alpha \cdot \frac{9}{128} = \frac{128}{34} \cdot \frac{9}{128} = \frac{9}{34} \approx 0.2647.$$

Let us compare the results:

$$P_B = 0.25, \qquad P_{NB} = 0.2647.$$

Thus, in this case, the classical and naive Bayesian classifiers give approximately the same result.

Example 2. Let us again look at the data on borrowers, but this time, let the class of "convicted male" be very poor and contain only one representative (see Tables 5, 6 and 7).

Table 5. Random.

	Sex	Crime	Score
1	Male	No	–
2	Male	No	+
3	Female	Yes	–
4	Female	Yes	+
5	Female	No	+
6	Female	No	+
7	Male	No	–
8	Female	No	+
9	Female	No	+
10	Male	No	–
11	Female	No	–
12	Male	No	+
13	Female	Yes	+
14	Female	Yes	+
15	Female	Yes	–
16	Male	Yes	+

Table 6. Group by Sex/Crime.

	Sex	Crime	Score
6	Female	No	+
8	Female	No	+
9	Female	No	+
5	Female	No	+
11	Female	No	–
4	Female	Yes	+
13	Female	Yes	+
14	Female	Yes	+
3	Female	Yes	–
15	Female	Yes	–
2	Male	No	+
12	Male	No	+
7	Male	No	–
1	Male	No	–
10	Male	No	–
16	Male	Yes	+

Table 7. Group by Score.

	Sex	Crime	Score
6	Female	No	+
8	Female	No	+
9	Female	No	+
5	Female	No	+
4	Female	Yes	+
13	Female	Yes	+
14	Female	Yes	+
2	Male	No	+
12	Male	No	+
16	Male	Yes	+
11	Female	No	–
3	Female	Yes	–
15	Female	Yes	–
7	Male	No	–
1	Male	No	–
10	Male	No	–

Let us calculate the probability of credit repayment by the payer belonging to the class of "convicted male".

A. Classical Bayesian classifier

Denote

$$P_B = P(\mathbf{Score} = \text{"+"} \mid \mathbf{Sex} = \text{"Male"}, \mathbf{Crime} = \text{"Yes"}).$$

By the Bayes formula:

$$P_B = \frac{P(\mathbf{Score} = \text{"+"})P\,(\mathbf{Sex} = \text{"Male"}, \mathbf{Crime} = \text{"Yes"} \mid \mathbf{Score} = \text{"+"})}{P(\,\mathbf{Sex} = \text{"Male"}, \mathbf{Crime} = \text{"Yes"})}.$$

From Table 7 we get:

$$P(\mathbf{Score} = \text{"+"}) = 10/16, \quad \text{and} \quad P\,(\text{Sex} = \text{"Male"}, \text{Crime} = \text{"Yes"} \mid \text{Score} = \text{"+"}) = 1/10,$$

from Table 6 we get:

$$P(\,\mathbf{Sex} = \text{"Male"}, \mathbf{Crime} = \text{"Yes"}) = 1/16.$$

Therefore,

$$P_B = \frac{10/16 \cdot 1/10}{1/16} = \frac{1}{1} = 1.$$

The same result can be obtained directly from Table 6.

However, the result obtained contradicts common sense, since it is well known that a man is a less responsible social counterparty than a woman, and a convicted person is a less responsible social counterparty than a non-convicted person. At the same time, the formal calculations given above show that if the borrower is a man, and if he has a criminal record, then he will return the loan for sure. This is absurd.

B. Naive Bayesian classifier

Let us denote:

$$P_{NB} = P(\mathbf{Score} = \text{"+"} \mid \mathbf{Sex} = \text{"Male"}, \mathbf{Crime} = \text{"Yes"}).$$

Since the algorithm of the naive Bayesian classifier allows us to re-evaluate only all hypotheses at once, we introduce additional notations:

$$P_+ = P_{NB}, \qquad P_- = P(\mathbf{Score} = \text{"–"} \mid \mathbf{Sex} = \text{"Male"}, \mathbf{Crime} = \text{"Yes"}),$$

For the hypothesis P_+, we use the naive Bayesian assumption:

$$P_+ \propto P(\mathbf{Score} = \text{+"}) \cdot \text{P (Sex=Male| Score=+")} \cdot \\ \cdot P\,(\,\mathbf{Crime} = \text{"Yes"} \mid \mathbf{Score} = \text{"+"}).$$

From Table 3.3 we get the probability of the hypothesis:

$$P(\mathbf{Score} = \text{" + "}) = 10/16,$$

and two conditional probabilities:

$P($ Sex = "Male"| Score = "+") $= 3/10,$
P (**Crime** = "Yes"| **Score** = "+") $= 4/10.$

Therefore,

$$P_+ \propto \frac{10}{16} \cdot \frac{3}{10} \cdot \frac{4}{10} = \frac{3}{40}.$$

For the hypothesis P_-, we also use the naive Bayesian assumption:

$$P_- \propto P(\textbf{Score} = \text{–}") \cdot \text{P (Sex=Male| Score=–")} \cdot$$
$$\cdot P\ (\ \textbf{Crime} = \text{"Yes"}|\ \textbf{Score} = \text{"–"}).$$

From Table 3.3 we get the probability of the hypothesis:

$$P(\textbf{Score} = \text{"–"}) = 6/16,$$

and two conditional probabilities:

$$P(\text{Sex = "Male"| Score = "–"}) = 3/6,$$
$$P\ (\ \textbf{Crime} = \text{"Yes"}|\ \textbf{Score} = \text{"–"}) = 2/6.$$

Therefore,

$$P_- \propto \frac{6}{16} \cdot \frac{3}{6} \cdot \frac{2}{6} = \frac{1}{16}.$$

We get the proportionality for the sum of the probabilities:

$$P_+ + P_- \propto \frac{3}{40} + \frac{1}{16}$$

Therefore

$$P_+ + P_- = \alpha \cdot \frac{11}{80}.$$

The coefficient α is found based on the condition $P_+ + P_- = 1$. We get $\alpha = 80/11$ and

$$P_+ = \alpha \cdot \frac{3}{40} = \frac{80}{11} \cdot \frac{3}{40} = \frac{6}{11} \approx 0.5455.$$

Let us compare the results:

$$P_B = 1, \qquad P_{NB} = 0.5455.$$

The results vary significantly. Moreover, the naive Bayesian classifier gives a much more realistic probability of credit repayment by the borrower belonging to the class of "convicted male" than the classical Bayesian classifier.

References

1. P. Domingos, M. Pazzani, Machine Learning **29**, 103–137 (1997)
2. I. Rish, An empirical study of the naive Bayes classifier. IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence (2001)
3. D. Hand, K. Yu, International Statistical Review **69(3),** 385–399 (2001)

4. M. Mozina, J. Demsar, M. Kattan, B. Zupan, *Nomograms for Visualization of Naive Bayesian Classifier.* In Proc. of PKDD-2004, pp. 337—348 (2010)
5. M. Maron, Journal of the ACM (JACM) **8(3),** 404–417 (1961)
6. M. Minsky, Steps toward Artificial Intelligence. Proceedings of the IRE **49(1),** 8–30 (1961)
7. A. McCallum, K. Nigam, A Comparison of Event Models for Naive Bayes Text Classification. In AAAI/ICML-98 Workshop on Learning for Text Categorization pp. 41–48 (AAAI Press, 1998)
8. A. Ng, M. Jordan, Neural Information Processing Systems **14,** 841–848 (2001)
9. R. Duda, P. Hart, D. Stork, Pattern Classification (Wiley, New-York, 2001)
10. T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning (Springer, New-York, 2001)
11. P. Mills, International Journal of Remote Sensing **32(21),** 6109–6132 (2011)
12. G. Toussaint, International Journal of Computational Geometry and Applications **15(2),** 101–150 (2005)
13. T. Hastie, The elements of statistical learning : data mining, inference, and prediction : with 200 full-color illustrations (Springer, New-York, 2001)
14. F. Nigsch, A. Bender, B. van Buuren, J. Tissen, E. Nigsch, J. Mitchell, Journal of Chemical Information and Modeling **46(6),** 2412–2422 (2006)
15. N. Altman, The American Statistician **46(3),** 175–185 (1992)