NIST Technical Note 2044

Unreliable evidence in binary classification problems

David Flater

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2044

National Institute of
Standards and Technology
U.S. Department of Commerce

NIST Technical Note 2044

Unreliable evidence in binary classification problems

David Flater
Software and Systems Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2044

May 2019

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

Certain commercial entities, equipment, or materials may be identified in this document in order to describe an experimental procedure or concept adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the entities, materials, or equipment are necessarily the best available for the purpose.

National Institute of Standards and Technology Technical Note 2044
Natl. Inst. Stand. Technol. Tech. Note 2044, 14 pages (May 2019)
CODEN: NTNOEF

This publication is available free of charge from:
https://doi.org/10.6028/NIST.TN.2044

Unreliable evidence in binary classification problems

David Flater

May 2019

This publication is available free of charge from: https://doi.org/10.6028/NIST.TN.2044

Abstract

Binary classification problems include such things as classifying email messages as spam or non-spam and screening for the presence of disease (which can be seen as classifying a subject as disease-positive or disease-negative). Both Bayesian and frequentist approaches have been applied to these problems. Both kinds of approaches provide poor estimates of the predictive value of tests for which the number of positive results in the sample is either very small or very large. A classifier that does not account for the uncertainty of these estimates is vulnerable to making inferences from unreliable evidence. This report explains the problem and explores options for accounting for the often-neglected uncertainty. A neat solution that does no harm to less uncertain cases remains elusive.

1 Introduction

Consider the following scenario:

- We have a sample of size N, meaning N individuals chosen randomly from a much larger population. The nature of the individuals does not matter; they could be people, objects, email messages, or something abstract.
- Each individual in the sample has been assigned to one of two classes. This assignment is "ground truth."
- There is a set of tests that we can apply to any individual and obtain a positive or negative result. Equivalently, we can think of a set of properties that each individual either has or does not have.
- We assume (for the purposes of this discussion) that each individual test or inspection has no cost and its result has no uncertainty. We have all of the test results "for free."
- We assume that the available tests, either singly or in combination, can possibly and plausibly be predictors or indicators of individuals' classes.
- Given the sample and the set of tests, we want to find a reliable method of classifying other individuals that are drawn from the population without ground truth.

This is a generic binary classification problem. The pattern arises in different contexts, and different models and methods are used to tackle it.

One example is classifying email messages as spam or non-spam. The properties that we test for are whether particular words appear in the emails or not; ground truth is whether the emails are deemed to be spam or non-spam by the recipient. A simple solution would identify particular words that are treated as reliable indicators that a message is either spam or non-spam. A more complex solution would assign to each word an estimate of the strength of evidence that its presence or absence provides that the message is spam or non-spam and then combine the evidence from all of the words to reach a classification for the message.

A different example is diagnosing a disease based on medical tests. Each test, for each patient, produces a positive or negative result; ground truth is whether the patient actually does or does not have the disease.

Subject: re : urgent buy recommendation
X-Bogosity: Ham, tests=bogofilter, spamicity=0.284 240, version=1.2.4

	n	pgood	pbad	fw	U
“highly”	3	0.033 333	0.000 000	0.003 067	+
“initial”	2	0.022 222	0.000 000	0.004 587	+
“medical”	2	0.022 222	0.000 000	0.004 587	+
“price”	2	0.022 222	0.000 000	0.004 587	+
“strong”	2	0.022 222	0.000 000	0.004 587	+
“quote”	1	0.011 111	0.000 000	0.009 094	+
“share”	1	0.011 111	0.000 000	0.009 094	+
“which”	32	0.344 444	0.100 000	0.225 164	−
“has”	31	0.333 333	0.100 000	0.230 935	−
“buy”	0	0.000 000	0.000 000	0.520 000	−
.....					
“yahoo”	0	0.000 000	0.000 000	0.520 000	−
“the”	92	0.911 111	1.000 000	0.523 255	−
“for”	71	0.688 889	0.900 000	0.566 422	−
“this”	61	0.588 889	0.800 000	0.575 984	−
“their”	27	0.255 556	0.400 000	0.610 110	−
“more”	33	0.311 111	0.500 000	0.616 386	−
“with”	57	0.533 333	0.900 000	0.627 873	−
“information”	24	0.200 000	0.600 000	0.749 830	−
“announced”	3	0.022 222	0.100 000	0.816 423	−
“report”	3	0.022 222	0.100 000	0.816 423	−
“almost”	6	0.044 444	0.200 000	0.817 300	−
“place”	8	0.055 556	0.300 000	0.843 031	−
“com”	9	0.044 444	0.500 000	0.917 581	+
“priority”	3	0.011 111	0.200 000	0.944 848	+
“http”	1	0.000 000	0.100 000	0.991 605	+
N_P_Q_S_s_x_md	10	0.431 520	0.000 000	0.284 240	
		0.017 800	0.520 000	0.375 000	

Figure 1: Demonstrative example of words with few occurrences dominating a Bogofilter classification.

For email classification, it is common to use Bayesian and Bayes-like approaches to estimate the evidence provided by the individual tests, and then combine the evidence from all of the tests (since in this case the real cost of testing is low). For disease diagnosis, it is more common to speak in terms of sensitivity, specificity, positive or negative predictive value, false positive and false negative rates, likelihood ratios, and related metrics, and then select one or two tests that are found to be the most useful (since in this case the real cost of testing is high).

The issue at hand is that both approaches provide poor estimates regarding tests for which the number of positive results in the sample is either very small or very large; for example, where only one or two individuals in the sample had positive results, or where only one or two individuals in the sample had negative results. In oversimplified applications of the statistics, predictive value is assessed based on the correlations between test results and ground truth for a given sample, neglecting the separate effect that a large imbalance between positive and negative results has on the uncertainty of that estimate. As a result, the weakest evidence appears to be the strongest. Tests with unproven predictive value float to the top and their combined force determines the outcome. Figure 1 shows an example using Bogofilter [1] where a spam message was classified as “ham” (non-spam) due to the presence of words having 3 or fewer occurrences in the training corpus.

The remainder of this report is organized as follows. Section 2 surveys related work. Section 3 gives an intuitive overview of the problem. Section 4 and Section 5 provide Bayesian and frequentist examples respectively. Section 6 addresses philosophical matters about multiple layers of uncertainty. Section 7 finds another side of the problem using a hypothesis testing approach. Section 8 describes an attempt to resolve the problem using likelihood ratios. Finally, Section 9 wraps up with conclusions.

This publication is available free of charge from: https://doi.org/10.6028/NIST.TN.2044

2 Related work

A screening test for membership in a given class can be evaluated in terms of sensitivity, specificity, positive or negative predictive value, and related statistics. With these statistics being in common use under various names across different fields, a canonical reference for them is difficult to identify, but their definitions are widely available [2]. Confidence intervals for these statistics can be approached in many ways using recipes that apply to binomial proportions [3, 4] or with Bayesian methods [5].

Bayesian classification is one branch of work that was built upon Bayes' theorem [6, 7, 8, 9]. A lot of practical implementation and refinement of Bayesian and Bayes-like classification methods was done for the purpose of detecting spam email [10, 11, 12, 13, 14, 15, 16, 17]. Bogofilter [1] was one application of that work, but there are many other email filters that implement the same or similar methods.

Bayesian classification also is used in the evaluation of forensic evidence, particularly through likelihood ratios. National Institute of Standards and Technology (NIST) statisticians recently looked at the question of uncertainty characterization for this context [18].

3 Intuition of the problem

Since the problem is symmetrical, I will discuss only the case of classifying individuals or diagnosing disease based on positive test results.

For many statistics, the size of the sample is a dominant factor in the statistical significance ascribed to results. This is the case when the scenario described in Section 1 is modelled as a Bernoulli process. Each test is applied N times, once to each of the N individuals in the sample. Should a test return a positive result only once, the $N-1$ times that it declined to do so are just as informative as the one time that it did.

Now consider the example of a rare word appearing in a sample of emails. Since it is a rare word, its prevalence in the corpus is going to be low. This is not consequential to the testing; it is an independent property of the word itself. If it has 1 occurrence in the sample, is it right to conclude that it is specific to spam or non-spam? What if it has 2 occurrences that happen to both be spam or non-spam?

Analogously, consider the example of a medical test that seldom returns a positive result. Positive results will have low prevalence in the sample, regardless of the sample. If it comes up positive only once, and it happens in a case that is positive for the disease, is it right to conclude that the test has predictive value?

The words "regardless of the sample" are at the crux of the matter. We do not know whether the distribution of results actually has anything to do with the sample it is applied to. As a surrogate for this knowledge, we analyze the correlations with statistics. It just so happens that these statistics are misleading for a test that randomly distributes positive results at a rate close to $1/N$.

4 Bayesian example

Let N be the size of the sample. Let r be the proportion of the sample that is class C (as ground truth). Suppose that a test t returns just one positive result, and it is for a member of C. Using naive Bayes estimates, we find t to be sufficient to determine class C with 100 % probability regardless of the values of N and r.

This publication is available free of charge from: https://doi.org/10.6028/NIST.TN.2044

$p(C\|t) = \frac{p(C) \cdot p(t\|C)}{p(t)}$	Bayes' rule
$\hat{p}(C) = r$	Proportion of sample that is class C
$\hat{p}(t\|C) = 1/rN$	1 individual in C has t
$\hat{p}(t) = 1/N$	Prevalence of t in the sample
$\hat{p}(C\|t) = \frac{r \cdot 1/rN}{1/N} = 1$	Q.E.D.

This conclusion is wrong because the uncertainty associated with the use of estimates has been ignored.

While the Bayesian pedigree of Bogofilter's model has been questioned [17, §1][19], it is close enough in any event to serve as a real world example of the issue just described. With all configuration parameters left at their default values, Bogofilter will assign a word that appears only once in a corpus, in a spam message, a "spamminess" (ostensible probability of indicating spam) of 0.991 605 on a scale from 0 to 1. The extremity of this value is then interpreted as an indication of the evidentiary significance of the presence of the word, and it is used in the combined assessment of the entire message while words with less extreme values are excluded. If the word appears in a non-spam message instead, it is assigned a spamminess of 0.009 094, and it is again taken as useful evidence, but for the opposite conclusion. These outliers are unaffected by the size of the training corpus or its balance of ham and spam, and it does not take many of them to overwhelm real evidence and destroy the performance of the system.

Instead of calculating a point estimate, one could use a more complete Bayesian prediction approach with probability distributions to model the uncertainty, resulting in a wide posterior predictive distribution for $\hat{p}(C|t)$ when the training corpus is small. However, I have not encountered an example of this being done, much less consensus on an acceptable recipe.

Discarding words with low counts is a "common text classification practice" [17, §2]. Even without misspellings, there is always a proliferation of words that appear only once in any given corpus, so ignoring the problem often leads to poor results. However, it is worrisome to discard low-frequency words just as a rule of thumb without quantifying the underlying issue.

Bogofilter provides two parameters, s and x, for the purpose of de-weighting words with few occurrences in the training corpus. x is the prior value that is used for the spamminess and s is the weighting factor for that prior value, relative to a single occurrence of the word. By tuning these parameters, one can moderate the effects of rare words, but it is far from obvious whether the action of these parameters in the spamminess formula is the right tool for the job.

5 Frequentist example

Let N be the size of the sample. Let r be the proportion of the sample that is positive for a disease (as ground truth). Suppose that a test t returns just one positive result, and it is for someone who has the disease. The estimated specificity and positive predictive value (PPV) of t are 1 and its estimated positive likelihood ratio is infinite, so we find a positive result from t to be sufficient to diagnose the disease with 100 % probability regardless of the values of N and r.

This conclusion is wrong because the uncertainty of the statistics has been ignored.

Various recipes are available to derive confidence intervals [3]. Practitioners seeking the most simple and familiar approach might use a Wald type formulation of the PPV confidence interval using standard estimates for sensitivity and specificity [4]. But unless continuity corrections are applied, this approach yields a degenerate, zero-width interval around 1. Users may therefore be misled into believing that the estimates are certain when the opposite is true.

As of December 11, 2018, a Google search for "confidence interval predictive value" returns an implementation of this misleading approach as the very first result [20].

This publication is available free of charge from: https://doi.org/10.6028/NIST.TN.2044

A better choice in this case is the Clopper-Pearson interval [21]. For the scenario described above, the Clopper-Pearson interval for the specificity narrows around 1 as N increases, but the same does not occur for the PPV. PPV depends only on the positive test results, of which we have a fixed number. The 95 % (minimum) Clopper-Pearson interval for PPV with just one true positive and no false positives is approximately (0.03 to 1.00) regardless of the values of N and r.

The Clopper-Pearson interval for PPV thus accurately conveys the fact that the strength of evidence provided by a test that has few positive results is limited by that small number of positive results and is not improved by having a larger sample. Other intervals surveyed in [3] are also suitable and are said to be better.

6 Layers of uncertainty

Both Bayesian and frequentist approaches to the scenario produce misleading estimates of the predictive value of a test. Predictive value can be interpreted as a probability estimate or level of confidence about whether a test result is indicative of a particular ground truth. The problem therefore could be filed under "uncertainty of the uncertainty," which is sometimes philosophically rejected or categorically disregarded.

Multiple layers of uncertainty can be collapsed through an argument of the following kind:

1. The estimated PPV expresses certainty that a positive test result reliably indicates the presence of the disease.
2. However, the PPV estimate is very uncertain.
3. Therefore, we have little confidence that a positive test result reliably indicates the presence of the disease.

The problem is then reduced to: one of the significant components of uncertainty for the reliability of the indication has been disregarded. We conclude that the values being used for $\hat{p}(C|t)$ are wrong.

7 Significance test

Ref. [3] begins, "Applied statisticians have long been aware of the serious limitations of hypothesis tests when used as the principal method of summarizing data. Following persuasion... [leading medical journals] have indicated that in general confidence intervals (CIs) are preferred to p-values in the presentation of results." While that point is well-taken, framing the problem at hand in terms of p-values provides a different perspective that reveals another side of the original problem.

In the context of statistics literature, the following may be recognized as an application of Fisher's exact test [22, §21·02][23]; more generically, it may simply be called the hypergeometric test [24].

We take it as given that the prevalence of positive test results is not informative about the likelihood that the test is valid. Test results that are manufactured by coin toss (a fake test) are no less fraudulent if the coin is heavily biased. The significance test, therefore, is by comparison with a random distribution having the same number of "hits." If the correlation of a test's results with ground truth has no statistically significant separation from what could be expected to occur if test results were fabricated using a random number source, then its predictive value has not been established and it should not be relied upon as evidence.

For a given sample, test, and level of confidence, the question of significance is answered using the hypergeometric distribution. As defined in R [25], the hypergeometric probability distribution function phyper is

$$\text{phyper}(x, m, n, k) = \frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}} \tag{1}$$

This publication is available free of charge from: https://doi.org/10.6028/NIST.TN.2044

for x in the range 0 to k, where x is "the number of white balls drawn without replacement from an urn which contains both black and white balls," m is "the number of white balls in the urn," n is "the number of black balls in the urn," and k is "the number of balls drawn from the urn."

The hypergeometric inverse cumulative distribution function qhyper then is

$$\text{qhyper}(p, m, n, k) = \text{minimum } q \text{ such that} \sum_{x=0}^{q} \text{phyper}(x, m, n, k) \geq p \tag{2}$$

The null hypothesis is that the screening test is not more sensitive than a fake test in which the same number of positive test results are distributed randomly through the sample without any reference to ground truth. (Since the other parameters are fixed, specificity is an equivalent criterion.)

$$H_0 \leftrightarrow x_{11} \leq \text{qhyper}\left(c, n_1, n_0, m\right) \tag{3}$$

where c is the level of confidence $(1 - \alpha)$ adjusted for multiple comparisons as explained below, n_1 is the number of disease-positives in the sample, n_0 is the number of disease-negatives in the sample, m is the number of positive test results, and x_{11} is the number of true positives.

It is impossible to reject the null hypothesis when the probability of a fake test having no false positives exceeds α. This occurs readily at small m. However, it is also impossible to reject the null hypothesis when the probability of a fake test having no false negatives exceeds α. This occurs readily at large m, not as a symmetry but as a separate mode of failure.

With a large number of tests, the probability of some test appearing to be significant when it is not becomes high. To filter out insignificant tests effectively, c must be adjusted to account for the fact that we are testing one independent hypothesis for each different test that is identified for potential use in classification. A sufficient correction is obtained using the Šidák inequality [26, 27], which amounts to raising c to the power of 1 over the number of simultaneous comparisons (i.e., the number of tests). When the number of simultaneous comparisons is large, the adjusted level of confidence that is then required to reject any of the individual null hypotheses rises dramatically. For example, if 1000 tests are considered, an adjusted confidence level of 0.9999 ("four nines") provides only 90 % confidence against a false discovery. In a sample with 10 disease-positives and 90 disease-negatives, the resulting criterion is unsatisfiable for $m < 4$ or $m > 42$.

The two distinct modes of failure for small m and large m reveal that a rule of thumb to discard tests having few positive results is justified, but not sufficient. In addition, the cutoff values that this criterion produces vary based on the parameters, so an arbitrarily chosen threshold cannot be best for all cases.

8 Classification using likelihood ratios

The preceding section provided a way to detect when a given test does not meet a standard of significance. However, a method to adjust the evidential weight that is given to individual tests is still needed. Discarding tests that do not meet the standard while assigning full weight to those that barely meet it is a crude approach.

One option is to reformulate the classifier in terms of likelihood ratios (LR) [18, §5]. An LR for the combination of all tests is the product of the prior odds (if any) and the LRs, positive or negative as applicable, for each individual test. The evidential significance of an individual test manifests in the confidence interval for its LR. To avoid giving weight to insignificant evidence, one can choose the value from the interval that is closest to 1 as the estimate to be used in subsequent calculations.

Unfortunately, confidence intervals for LRs are even more troublesome than those for PPV. Marill et al. [28] surveyed previous work and commented, "Calculating the 95 % CI for the LR, a ratio of binomial proportions that incorporates both sensitivity and specificity, can be challenging." They defined a bootstrap method of finding a confidence interval and provided an implementation in the R package bootLR [29]. While it avoids

This publication is available free of charge from: https://doi.org/10.6028/NIST.TN.2044

```
> BayesianLR.test(truePos=1, totalDzPos=10, trueNeg=90, totalDzNeg=90)

Likelihood ratio test of a 2x2 table

data:
   truePos totalDzPos    trueNeg totalDzNeg
         1         10         90         90
Positive LR: Inf (2.505 - Inf)
Negative LR: 0.900 (0.664 - 1.022)
95% confidence intervals computed via BCa bootstrapping.
```

Figure 2: Confidence intervals computed according to Marill et al. [28].

the degeneracy that arises with methods based on normal approximations, it does not solve the problem of a word appearing only once being considered evidentially significant. For the example that was discussed earlier (1 true positive in a sample of size 100), the 95 % interval for the positive LR remains stubbornly above 1 (see Figure 2).

To work around this, we can instead construct a conservative interval for the LR using the endpoints of Clopper-Pearson confidence intervals for the numerator and denominator. The full "LR method" of classification then is as follows:

1. Assign an initial message LR value of 1, indicating equal prior probability of the message being ham or spam.
2. A word that exists in the database is either present in (T+) or absent from (T−) the message. For whichever applies (Tx), compute Clopper-Pearson confidence intervals for the numerator $p(\text{Tx}|\text{spam})$ and denominator $p(\text{Tx}|\text{ham})$ of the LR using the relevant counts from the database.
3. Form a conservative LR interval by using (numerator minimum) / (denominator maximum) and (numerator maximum) / (denominator minimum) as endpoints.
4. Choose the value from that interval that is closest to 1 and multiply the message's LR by that value.
5. The final message LR r, which is the product of all of the LRs that were computed for each word in the database, is converted to a probability p that the message is spam (labelled "spamicity" in Bogofilter's output) via $p = r/(1 + r)$.

The LR method uses different approaches to filtering, weighting, and combining evidence than does Bogofilter. Also, it considers the absence of a word from a message as possible evidence, while Bogofilter only considers the presence of words as evidence.

The first step in evaluating this method is to see how it compares with Bogofilter as the baseline in a "normal" scenario. For this, we use the bare version of the Ling-Spam corpus [11, 30]. This corpus is evenly divided into 10 partitions of comparable size and content. The message counts for each of the 10 parts are shown in Table 1. The ranges of the counts for ham and spam are 241 to 242 and 48 to 49 respectively, for a total count of 289 to 291 and a spam prevalence of approximately 17 % in each part.

The purpose of the partitioning was to enable a 10-fold cross-validation where the classifier is trained on 9 parts before being tested on the 10th. This strategy was followed to collect data on the performance of the original Bogofilter method and the LR method described above. LR data were collected by replacing the scoring mechanism in Bogofilter, keeping the tokenization logic and everything else the same.

A kernel density plot of the resulting spamicities (Figure 3) shows that they are tightly clustered for both methods, but the LR method produced fewer inconclusive classifications.

Unfortunately, the receiver operating characteristic (ROC) curves reveal that classification accuracy suffered under the LR method. Figure 4 compares the Bogofilter method with the LR method, first using a 95 % level of confidence for the individual Clopper-Pearson intervals and then using a 99.9999 % level of confidence

This publication is available free of charge from: https://doi.org/10.6028/NIST.TN.2044

Table 1: Message counts in the 10 parts of the Ling-Spam corpus.

Part	Ham	Spam	Total	Part	Ham	Spam	Total
part1	241	48	289	part6	241	48	289
part2	241	48	289	part7	241	48	289
part3	241	48	289	part8	241	48	289
part4	241	48	289	part9	241	48	289
part5	242	48	290	part10	242	49	291

for the same. Since the curves for the LR method are almost entirely underneath the curve for the original Bogofilter method, it seems that the LR method is actually harmful.

The value 0.4 at the elbow of the curve for the Bogofilter method shows that Bogofilter's default thresholds of 0.45 and 0.99 for ham and spam respectively are suboptimal for this corpus. The filter's parameters would need to be tuned to realize the performance represented by the ROC curve. The much lower values at the elbow of the curve for the LR method indicate that a significant number of spam messages received spamicities that place them in the ham cluster.

9 Conclusion

This report called attention to a component of uncertainty that is sometimes overlooked in both Bayesian and frequentist approaches to binary classification problems. Discarding tests that show few positive results in the sample reduces the impact that this uncertainty can have, but this mitigation is incomplete and the threshold might have been set arbitrarily.

To handle the uncertainty, a frequentist confidence interval for predictive value can be used to adjust the evidential weight that is given to individual tests. Alternatively, a significance test, instead of an arbitrary rule of thumb, can be used as a quantitative basis on which to discard tests.

Questions remain of how to adjust inferences to account for the uncertainty of predictive value. An attempt to do this using confidence intervals for likelihood ratios actually degraded performance on the corpus tested, for the particular model and predictor used. It is possible that the likelihood ratio approach could perform better in different circumstances. For classifiers that combine evidence using Fisher's combined probability test [22, §21·1][31], perhaps it should be replaced with the weighted Z-method [32], with weights being set according to the uncertainty of the individual probabilities.

If the sample is insufficient for the intended purpose, suppressing evidence or propagating uncertainty does not solve the problem. Nevertheless, it is better to get an "unsure" result and know that the evidence is insufficient to support a conclusion than not to know that a conclusion is unreliable.

Acknowledgment

Thanks to Steven Lund, Hari Iyer, and William F. Guthrie for statistical advice and to Ya-Shian Li-Baboud for other helpful suggestions.

References

[1] Eric S. Raymond, David Relson, Matthias Andree, Greg Louis, et al. Bogofilter, 2002. http://bogofilter.sourceforge.net/.

This publication is available free of charge from: https://doi.org/10.6028/NIST.TN.2044

This publication is available free of charge from: https://doi.org/10.6028/NIST.TN.2044

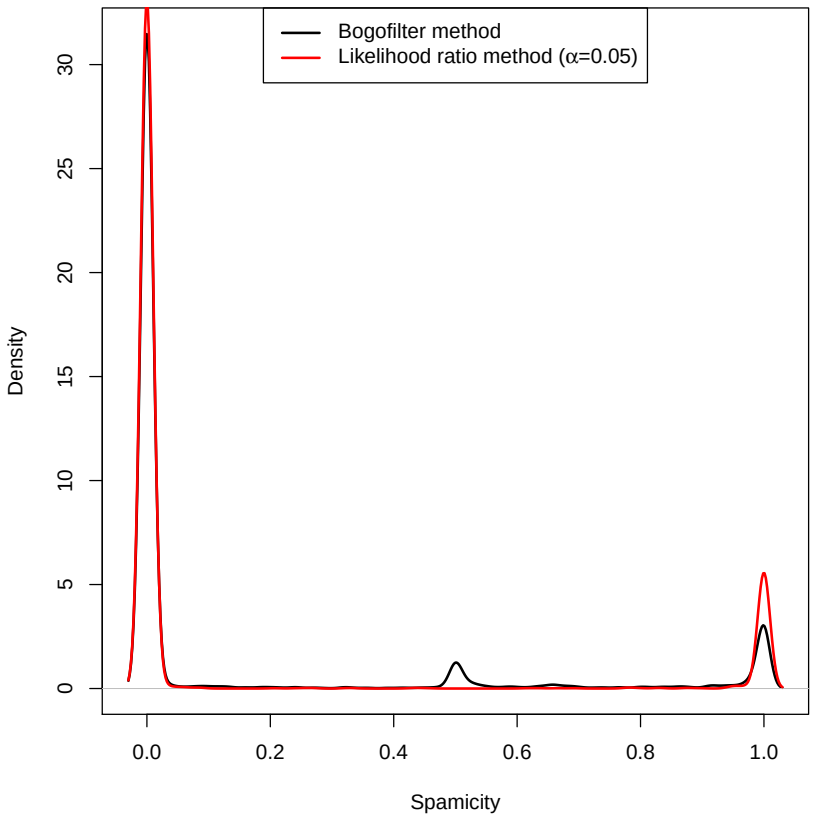

Figure 3: Gaussian kernel density (bandwidth = 0.01) of spamicities

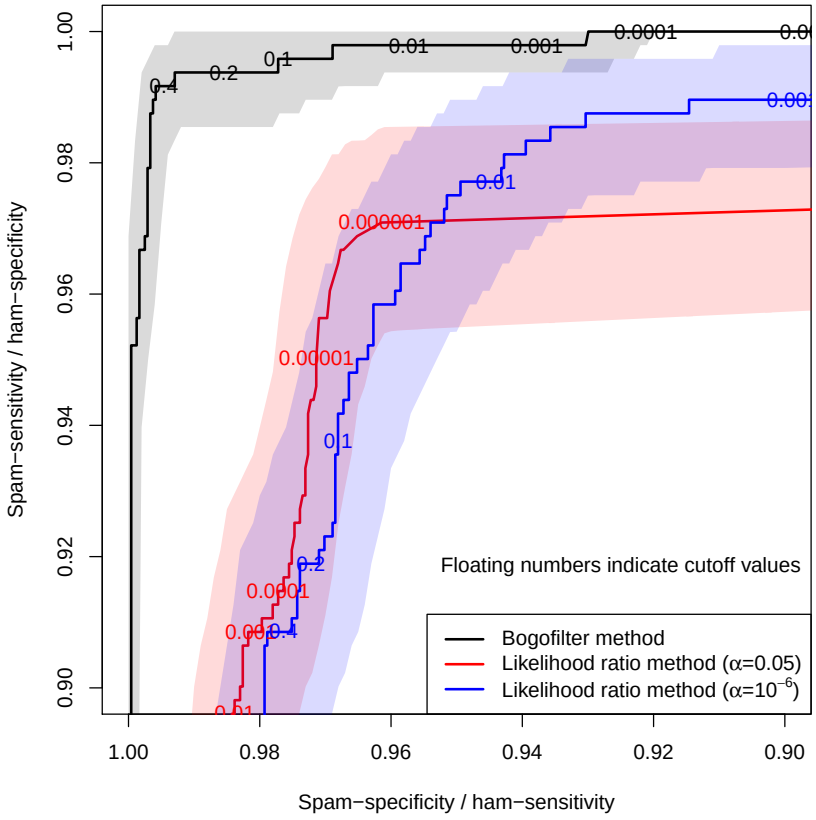

Figure 4: Approximate ROC curves (means with 95 % confidence intervals)

[2] Wikipedia. Sensitivity and specificity, June 2018. https://en.wikipedia.org/wiki/Sensitivity_and_specificity.

[3] Robert G. Newcombe. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statistics in Medicine*, 17:857–872, 1998.

[4] Nathaniel David Mercaldo, Xiao-Hua Zhou, and Kit F. Lau. Confidence intervals for predictive values using data from a case control study. Working Paper 271, UW Biostatistics Working Paper Series, 2005. http://biostats.bepress.com/uwbiostat/paper271.

[5] James D. Stamey and Melinda M. Holt. Bayesian interval estimation for predictive values from case-control studies. *Communications in Statistics—Simulation and Computation*, 39(1):101–110, 2009. https://doi.org/10.1080/03610910903312219.

[6] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, 53:370–418, 1763. https://doi.org/10.1098/rstl.1763.0053.

[7] Pierre-Simon Laplace. Mémoire sur la probabilité des causes par les événements. In *Œuvres complètes de Laplace: publiées sous les auspices de l'Académie des sciences*, volume 8, pages 27–65. 1891. Ebook available in Google Play.

[8] Pierre-Simon Laplace. Mémoire sur les probabilités. In *Œuvres complètes de Laplace: publiées sous les auspices de l'Académie des sciences*, volume 9, pages 383–485. 1893. Ebook available in Google Play.

[9] Pierre-Simon Laplace. Mémoire sur les approximations des formules qui sont fonctions de très grands nombres. In *Œuvres complètes de Laplace: publiées sous les auspices de l'Académie des sciences*, volume 10, pages 295–338. 1894. Ebook available in Google Play.

[10] Mehran Sahami, Susan Dumais, David Heckerman, and Eric Horvitz. A Bayesian approach to filtering junk e-mail. In *Proceedings of the AAAI Workshop on Learning for Text Categorization*, AAAI Technical Report WS-98-05, pages 55–62, 1998. https://www.microsoft.com/en-us/research/wp-content/uploads/1998/01/junkfilter.pdf or https://www.aaai.org/Papers/Workshops/1998/WS-98-05/WS98-05-009.pdf.

[11] Ion Androutsopoulos, John Koutsias, Konstantinos V. Chandrinos, George Paliouras, and Constantine D. Spyropoulos. An evaluation of naive Bayesian anti-spam filtering. In *Proceedings of the Workshop on Machine Learning in the New Information Age, 11th European Conference on Machine Learning (ECML 2000)*, pages 9–17, 2000. http://www.aueb.gr/users/ion/docs/mlnet_paper.pdf.

[12] Paul Graham. A plan for spam, August 2002. http://paulgraham.com/spam.html.

[13] Gary Robinson. Spam detection, September 2002. http://radio-weblogs.com/0101454/stories/2002/09/16/spamDetection.html.

[14] Paul Graham. Better Bayesian filtering, January 2003. http://paulgraham.com/better.html.

[15] Gary Robinson. A statistical approach to the spam problem. *Linux Journal*, March 1 2003. https://linuxjournal.com/article/6467.

[16] Gary Robinson. Handling redundancy in email token probabilities, May 6 2004. http://garyrob.blogs.com/handlingtokenredundancy94.pdf.

[17] Vangelis Metsis, Ion Androutsopoulos, and Georgios Paliouras. Spam filtering with naive Bayes—which naive Bayes? In *Proceedings of the 3rd Conference on Email and Anti-Spam (CEAS 2006)*, July 27–28 2006. With corrections: http://www2.aueb.gr/users/ion/docs/ceas2006_paper.pdf.

[18] Steven P. Lund and Hari Iyer. Likelihood ratio as weight of forensic evidence: A closer look. *Journal of Research of National Institute of Standards and Technology*, 122(27), 2017. https://doi.org/10.6028/jres.122.027.

[19] Tim Peters. Email on Python-Dev mailing list, August 22 2002. https://mail.python.org/pipermail/python-dev/2002-August/028216.html.

This publication is available free of charge from: https://doi.org/10.6028/NIST.TN.2044

[20] Unspecified author, Centre for Clinical Research and Biostatistics, Chinese University of Hong Kong. C.I. calculator: Diagnostic statistics. https://www2.ccrb.cuhk.edu.hk/stat/confidence%20interval/Diagnostic%20Statistic.htm. Accessed 2018-12-11.

[21] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

[22] Ronald A. Fisher. *Statistical Methods for Research Workers*. Number 5 in Biological Monographs and Manuals. Hafner, 12th edition, 1954.

[23] Wikipedia. Fisher's exact test, September 2018. https://en.wikipedia.org/wiki/Fisher%27s_exact_test.

[24] Wikipedia. Hypergeometric test, in hypergeometric distribution, September 2018. https://en.wikipedia.org/wiki/Hypergeometric_distribution#Hypergeometric_test.

[25] The R project for statistical computing, 2018. https://r-project.org/.

[26] Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, June 1967. Corollary 1. https://doi.org/10.1080/01621459.1967.10482935, https://jstor.org/stable/2283989.

[27] William F. Guthrie. Example of statistical tests for hypergeometric data with multiple comparison corrections using the Šidák method. Unpublished white paper, July 19 2018.

[28] Keith A. Marill, Yuchiao Chang, Kim F. Wong, and Ari B. Friedman. Estimating negative likelihood ratio confidence when test sensitivity is 100%: A bootstrapping approach. *Statistical Methods in Medical Research*, 26(4):1936–1948, 2017. https://doi.org/10.1177/0962280215592907.

[29] Keith A. Marill and Ari B. Friedman. bootLR: Bootstrapped confidence intervals for (negative) likelihood ratio tests. https://cran.r-project.org/package=bootLR.

[30] Ling-Spam corpus, July 17 2000. http://www.aueb.gr/users/ion/data/lingspam_public.tar.gz.

[31] Wikipedia. Fisher's method, September 2018. https://en.wikipedia.org/wiki/Fisher%27s_method.

[32] Michael C. Whitlock. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology*, 18(5), August 2005. https://doi.org/10.1111/j.1420-9101.2005.00917.x.

This publication is available free of charge from: https://doi.org/10.6028/NIST.TN.2044