This is an Accepted Manuscript of an article published by Taylor & Francis in The American Statistician on March 25, 2013 available online: http://www.tandfonline.com/10.1080/00031305.2013.778788

Majority Voting by Independent Classifiers Can Increase Error Rates

Stephen B. Vardeman

Max D. Morris

IMSE and Statistics Departments

Iowa State University

3004 Black Engineering Building

Ames, Iowa 50011-2164

Abstract

The technique of "majority voting" of classifiers is used in machine learning with the aim of constructing a new combined classification rule that has better characteristics than any of a given set of rules. The "Condorcet Jury Theorem" is often cited, incorrectly, as support for a claim that this practice leads to an improved classifier (i.e. one with smaller error probabilities) when the given classifiers are sufficiently good and are uncorrelated. We specifically address the case of 2-category classification, and argue that a correct claim can be made for independent (not just uncorrelated) classification errors (not the classifiers themselves), and offer an example demonstrating that the common claim is false.

Introduction

The body of written material on machine learning (e.g. books, online course notes, and even some research journal articles) is replete with assertions that "majority voting" by uncorrelated classifiers (that are sufficiently good individually) will improve on any one of the classifiers. The technical content of the argument usually given for those assertions concerns the two-class problem and amounts to the observation that for odd n if $p < .5$, the probability that a Binomial (n, p) variable

exceeds $.5n$ is less than p, a fact related to the "Condorcet Jury Theorem." This argument seems to be relevant only if applied to jointly independent (not just uncorrelated, or equivalently, pair-wise independent) *errors* (and *not to classifiers*). We formulate this issue carefully and provide a concrete numerical example that shows that not even complete independence (let alone uncorrelatedness) of good *classifiers* suffices to imply the commonly accepted conclusion. The issue serves as a simple example of the importance of precision of language when saying what mathematical results mean in practice, and how whole folklores and even literatures can develop around imprecision.

Probability Modeling for Three 2-Class Classifiers

For exposition purposes, consider a 2-group classification problem with observable $\mathbf{x}$ and group $y \in \{0,1\}$. Consider three classifiers $f_1(\mathbf{x}), f_2(\mathbf{x}), f_3(\mathbf{x})$ taking values in $\{0,1\}$ and a joint distribution P for $(\mathbf{x}, y)$. Since our interest is in error properties of classifiers, there is no real need to detail the input space or P beyond the probabilities assigned to the 16 basic events

$$\{(\mathbf{x}, y) \mid f_1(\mathbf{x}) = o_1, f_2(\mathbf{x}) = o_2, f_3(\mathbf{x}) = o_3, \text{ and } y = o_4\} \text{ for } \mathbf{o} = (o_1, o_2, o_3, o_4) \in \{0,1\}^4$$

and in fact we will suppress dependence upon $\mathbf{x}$; simply consider outcomes

$$(f_1, f_2, f_3, y) \in \{0,1\}^4$$

and treat P as a distribution on $\{0,1\}^4$.

The objective here is to consider the claim that under appropriate assumptions the "committee majority vote" classifier

$$g = I[f_1 + f_2 + f_3 > 1.5]$$

is necessarily better than each of f_1, f_2, f_3.

Associated with any classifier h (that for present purposes we assume is a function of f_1, f_2, f_3) is an error variable

$$e(h, y) = I[h \neq y]$$

and the error rate

$$\mathrm{E}e(h, y) = P[h \neq y]$$

Notice that by the very definition of g, the majority vote classifier makes an error if and only if at least two of f_1, f_2, f_3 make errors, that is

$$e(g, y) = I\left[e(f_1, y) + e(f_2, y) + e(f_3, y) \geq 2\right]$$

so that

$$\mathrm{E}e(g, y) = P\left[e(f_1, y) + e(f_2, y) + e(f_3, y) \geq 2\right]$$

If $e(f_1, y), e(f_2, y),$ and $e(f_3, y)$ are i.i.d. Bernoulli(p) for $p < .5$, then $e(f_1, y) + e(f_2, y) + e(f_3, y)$ is Binomial$(3, p)$ and it's fairly easy to show that the majority vote classifier error rate $\mathrm{E}e(g, y)$ (the Binomial probability of 2 or more successes) is less than p. This is a version of Condorcet's Theorem (see Boland (1998)) and is the argument usually put forward in technical support of the claim that g improves on each of f_1, f_2, f_3. (See, for example, page 274 of Webb (2002), pages 112-114 of Kuncheva (2004), Ruta and Gabrys (2002), Kuncheva *et al.* (2003), and Narasimhamurthy (2005); such errors can also be found in the statistics literature.) For example, the introduction of Narasimhamurthy (2005) says "A simple analytical justification for majority voting may be given by the well-known Condorcet's theorem. Under the assumption of independent classifiers, if the individual classifier error rate $e < .5$ (assume for simplicity that all classifiers have the same error rate), for odd number of classifiers (voters) N, the correct decision

rate increases with increasing N.") But the mathematical argument does not match the language used in making the claim. The claim is nearly always phrased in terms of assumptions on classifiers (not errors) and often is phrased in terms of uncorrelatedness (rather than independence).

Regarding these issues, notice that if random variables Z_1 and Z_2 both take values in $\{0,1\}$, they are uncorrelated if and only if they are independent. So three random variables Z_1, Z_2, Z_3 each taking values in $\{0,1\}$ are uncorrelated if and only if they are pair-wise independent. Joint independence of the three variables is a stronger condition. So uncorrelatedness of the classifiers does not imply complete independence of more than 2 classifiers (only pair-wise independence). Even an assumption of complete independence of the classifiers does not imply the complete independence of the errors $e(f_1, y), e(f_2, y), e(f_3, y)$; the first is a property only of the joint distribution of (f_1, f_2, f_3), while the second is a property of the full joint distribution of (f_1, f_2, f_3, y). Indeed, it seems to us that in the abstract the plausibility of an assumption of independence of the errors is much harder to contemplate than the plausibility of one regarding the classifiers. A sample of triples (f_1, f_2, f_3) would allow statistical investigation of the plausibility of the independence of the classifiers, while one would additionally need access to the corresponding y's in order to study the plausibility of independence of the errors. With a large training sample, contingency table methods could, in principle, be used to test assumptions of independent classifiers and errors.

A Counter-Example

It is simply not true that three (even completely) independent and identically distributed classifiers with small error rate will necessarily produce a committee classifier even as good as any one of the individuals when combined through majority voting. Consider the numerical example in Table 1. Rows of the table correspond to the 16 possible values of (f_1, f_2, f_3, y), show the

corresponding values of the majority vote classifier, g, and the error functions, and finally give probabilities for the outcomes in the last column.

Table 1: A Numerical Example

y	f_1	$e(f_1, y)$	f_2	$e(f_2, y)$	f_3	$e(f_3, y)$	g	$e(g, y)$	$P(f_1, f_2, f_3, y)$
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	0	0	.008
0	0	0	1	1	0	0	0	0	.008
0	1	1	0	0	0	0	0	0	.008
0	0	0	1	1	1	1	1	1	.08
0	1	1	0	0	1	1	1	1	.08
0	1	1	1	1	0	0	1	1	.08
0	1	1	1	1	1	1	1	1	0
1	0	1	0	1	0	1	0	1	$(.1)^3 = .001$
1	0	1	0	1	1	0	0	1	$(.9)(.1)^2 - .008 = .001$
1	0	1	1	0	0	1	0	1	$(.9)(.1)^2 - .008 = .001$
1	1	0	0	1	0	1	0	1	$(.9)(.1)^2 - .008 = .001$
1	0	1	1	0	1	0	1	0	$(.9)^2(.1) - .08 = .001$
1	1	0	0	1	1	0	1	0	$(.9)^2(.1) - .08 = .001$
1	1	0	1	0	0	1	1	0	$(.9)^2(.1) - .08 = .001$
1	1	0	1	0	1	0	1	0	$(.9)^3 = .729$

Under the distribution for (f_1, f_2, f_3, y) specified in Table 1, the three classifiers f_1, f_2, f_3 are independent Bernoulli$(.1)$. Each of them has the same error rate, namely

$$\begin{aligned} P\left[e(f_1, y) = 1\right] &= .008 + 2(.08) + (.1)^3 + 2\left((.9)(.1)^2 - .008\right) + (.9)^2(.1) - .08 \\ &= .008 + .16 + .001 + .002 + .001 \\ &= .172 \end{aligned}$$

On the other hand, the majority vote error rate is

$$\begin{aligned} P\left[e(g,y)=1\right] &= 3(.08)+(.1)^3+3\left((.9)(.1)^2-.008\right) \\ &= .24+.001+.003 \\ &= .244 \end{aligned}$$

The majority vote classifier is substantially worse than the individual classifiers in terms of error rate! In fact, the optimal classifier for this case DOES only depend on the sum of the three "votes" (see the next paragraph), but cannot be specified as a monotoic function of this quantity. (The sum of the classifiers is sufficient for the 2-distribution family of models for (f_1, f_2, f_3), but the likelihood ratio is not monotone in the sum.)

More intuitively, the example is "driven" by the largest probabilities in lines 5, 6, 7, and 16 of the table. In the last case, all individual classifiers are correct. But in each of rows 5-7, one classifier is correct while the other two are wrong; i.e. conditional on being in one of these states, the individual classisfiers have error probabilities of 2/3 while the majority-vote classifier is always wrong. The Editor has kindly observed that this structure can be generalized to whole family of examples with the same basic properties, and his analysis is outlined in the Appendix.

Note for comparison that, as an immediate consequence of the Neyman-Pearson Fundamental Lemma, the *optimal* (minimum error rate) classifier, based on the information given in the Table 1, simply selects the more probable value of *y* for each of the 8 values of (f_1, f_2, f_3). In this case, the optimal rule

$$h^* = I\left[f_1 = f_2 = f_3 = 0\right] + I\left[f_1 = f_2 = f_3 = 1\right]$$

is decidedly "undemocratic" in its functional form, with error rate $P\left[e\left(h^*, y\right)=1\right]=.006$ (the sum of probabilities from rows 1, 8, and 10-15 of the table), substantially less than the error rates of the individual classifiers or of g.

This optimal structure generalizes immediately for any finite number of classifiers that each produces any finite number of distinct predictions. Conditionally on any outcome of the ensemble of classifiers, the optimal combined classifier simply predicts the more probable y value (or most probable value where there are more than 2 distinct classes), with an overall error rate that is the sum of probabilities of y values not predicted in each case. To the extent that work on "classifier fusion" often begins with the premise that such probabilities are known/available, the large literature on the subject seems unjustified. The question of how to combine classifiers is only meaningful when one must approximate or estimate these joint probabilities.

A second counter-example based on continuous observable variables is described in the online supplementary material.

Conclusion

It should go without saying that there are many statistics and engineering publications that carefully and correctly consider the combination of classifiers, e.g. Breiman (2001) and Hu and Damper (2008). Even with regard to the error we address here, one might argue that the discussion is simply quibbling with small inconsequential instances of poor word choice. But precision of language is important. Classifiers are not errors, zero correlation is not independence, and seriously flawed claims follow when these facts are not understood. In order to judge whether an insight provided by a piece of mathematics is relevant and helpful, one must be careful to know exactly what that piece of mathematics says, and how it is to be interpreted in an application. In the present context, it is not so clear to us how to judge the practical reasonableness of an assumption that classification errors are i.i.d. across the members of a voting committee. More broadly, large folklores and even large published literatures can grow up around failed attempts to translate mathematics into practice.

References

Breiman, L. (2001). "Random Forests," *Machine Learning* Vol. 45, pp. 5-32.

Hu, R., and Damper, R.I. (2008). "A 'No Panacea Theorem' for Classifier Combination," *Pattern Recognition* Vol. 41, pp. 2665-2673.

Kuncheva, Ludmila I. (2004). *Combining Pattern Classifiers*, Wiley, New York.

Kuncheva, L.I., Whitaker, C.J., Shipp, C.A., and Duin, R.P.W. (2003). "Limits on the Majority Vote Accuracy in Classifier Fusion," *Pattern Analysis and Applications* Vol. 6, pp. 22-31.

Narasimhamurthy, Anand (2005). "Theoretical Bounds of Majority Voting Performance for a Binary Classification Problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, no. 12, pp. 1988-1995.

Ruta, Dymitr and Gabrys, Bogdan (2002). "A Theoretical Analysis of the Limits of Majority Voting Errors for Multiple Classifier Systems," *Pattern Analysis and Applications*, Vol. 5, pp. 333-350.

Webb, Andrew R. (2002). *Statistical Pattern Recognition*, 2nd Edition, Wiley, New York.

Appendix

A general structure including the counter-example is that where the three classifiers f_1, f_2, f_3 are i.i.d. Bernoulli(r) (so that $f_1 + f_2 + f_3$ is Binomial$(3,r)$) and conditional probabilities for y given the sum are (for $r > .1$)

$$P\left[y = 1 \mid f_1 + f_2 + f_3 = j\right] = \begin{cases} 0 & \text{if } j = 0 \text{ or } j = 3 \\ (.1)r^{-1} & \text{if } j = 1 \\ (.01)r^{-2} & \text{if } j = 2 \end{cases}$$

The optimal classifier based on this structure depends only upon this conditional distribution (choosing $y = 1$ when the conditional probability above is larger than $.5$) and is h^* as above provided $r > .2$.

Direct calculation using this structure shows that the majority vote classifier has error rate

$$P\left[e(g,y) = 1\right] = (1-r)\left(4r^2 - 2.3r + 1.27\right)$$

whereas the error rate of any single classifier is

$$P[e_1 = 1] = (1-r)\left(2r^2 - 1.1r + 1.09\right)$$

and the former exceeds the latter except when $r = .3$ and they are the same.

Online Supplementary Material

A second counter-example based on continuous observable variables is as follows. Let X_1, X_2, X_3, Y be random variables with $P[Y=0] = .264$ and $P[Y=1] = .736$ and conditional distributions for (X_1, X_2, X_3) given Y as follows. Let $G_{\boldsymbol{\mu},\sigma}$ be the spherical trivariate normal distribution with mean vector $\boldsymbol{\mu}$ and standard deviation σ. Suppose that conditional on $Y=0$, we let (X_1, X_2, X_3) have the mixture distribution

$$\frac{1}{33}\left(G_{(1,0,0),\sigma} + G_{(0,1,0),\sigma} + G_{(0,0,1),\sigma}\right) + \frac{10}{33}\left(G_{(1,1,0),\sigma} + G_{(0,1,1),\sigma} + G_{(1,0,1),\sigma}\right)$$

and conditional on $Y=1$, we let (X_1, X_2, X_3) have the mixture distribution

$$\frac{1}{736}\left(G_{(0,0,0),\sigma} + G_{(1,0,0),\sigma} + G_{(0,1,0),\sigma} + G_{(0,0,1),\sigma} + G_{(1,1,0),\sigma} + G_{(0,1,1),\sigma} + G_{(1,0,1),\sigma}\right) + \frac{729}{736} G_{(1,1,1),\sigma}$$

Then the variables $X_1, X_2,$ and X_3 are independent and marginally

$$.1\mathrm{N}\left(0,\sigma^2\right) + .9\mathrm{N}\left(1,\sigma^2\right)$$

Conditioned on $Y=0$ each X_i is

$$\frac{12}{33}\mathrm{N}\left(0,\sigma^2\right) + \frac{21}{33}\mathrm{N}\left(1,\sigma^2\right)$$

while conditioned on $Y=1$ each X_i is

$$\frac{4}{736}\mathrm{N}\left(0,\sigma^2\right)+\frac{732}{736}\mathrm{N}\left(1,\sigma^2\right)$$

For a specific example, take $\sigma=.25$ and suppose that

$$f_i = I\left[X_i \geq .25\right]$$

The marginal error rate is

$$\left[\frac{12}{33}\left(1-\Phi(1)\right)+\frac{21}{33}\left(1-\Phi(-3)\right)\right](.264)+\left[\frac{4}{736}\Phi(1)+\frac{732}{736}\Phi(-3)\right](.736)\approx .1873$$

The error rate for the majority vote classifier is easily shown to be bigger than $.1873$, as follows. Note that since the X_i are iid, so also are the f_i, and that the latter are each Bernoulli with success probability

$$P\left[X_1 \geq .25\right] = .1\left(1-\Phi(1)\right)+.9\left(1-\Phi(-3)\right)\approx .9147$$

Then,

$$\begin{aligned}&P\left[Y=0, X_1<.25, X_2<.25, X_3<.25\right]\\ &\quad=.264\left\{\frac{3}{33}(.8413)^2(.0013)+\frac{30}{33}(.8413)(.0013)^2\right\}\approx .000022424\end{aligned}$$

and

$$\begin{aligned}&P\left[Y=0, X_1<.25, X_2<.25, X_3>.25\right]\\ &\quad=\frac{.264}{33}\left\{\begin{array}{l}(.0013)(.8413)(.1587)+(.8413)(.0013)(.1587)\\ +(.8413)^2(.9987)+10(.0013)^2(.1587)\\ +10(.8413)(.0013)(.9987)+10(.0013)(.8413)(.9987)\end{array}\right\}\approx .00583249\end{aligned}$$

and

$$\begin{aligned}&P\left[Y=0, X_1<.25, X_2>.25, X_3>.25\right]\\ &\quad=\frac{.264}{33}\left\{\begin{array}{l}(.0013)(.1587)^2+2(.8413)(.9987)(.1587)\\ +20(.0013)(.9987)(.1587)+10(.8413)(.9987)^2\end{array}\right\}\approx .069296\end{aligned}$$

and

$$P\left[Y=0, X_1>.25, X_2>.25, X_3>.25\right]$$
$$=.264\left\{\frac{3}{33}(.9987)(.1587)^2+\frac{30}{33}(.9987)^2(.1587)\right\}\approx .03859271$$

The majority vote error rate is:

$$3(.069296)+.03859271+.000598+3(.0008225)\approx .2495$$

The complete joint distribution of (f_1, f_2, f_3, y) is given in Table A1.

Table A1: A Second Numerical Example

y	f_1	$e(f_1, y)$	f_2	$e(f_2, y)$	f_3	$e(f_3, y)$	g	$e(g, y)$	$P(f_1, f_2, f_3, y)$
0	0	0	0	0	0	0	0	0	.000022424
0	0	0	0	0	1	1	0	0	.00583249
0	0	0	1	1	0	0	0	0	.00583249
0	1	1	0	0	0	0	0	0	.00583249
0	0	0	1	1	1	1	1	1	.069296
0	1	1	0	0	1	1	1	1	.069296
0	1	1	1	1	0	0	1	1	.069296
0	1	1	1	1	1	1	1	1	.03859271
1	0	1	0	1	0	1	0	1	$(.0853)^3 - .000022424 = .000598$
1	0	1	0	1	1	0	0	1	$(.9147)(.0853)^2 - .00583249 = .0008225$
1	0	1	1	0	0	1	0	1	$(.9147)(.0853)^2 - .00583249 = .0008225$
1	1	0	0	1	0	1	0	1	$(.9147)(.0853)^2 - .00583249 = .0008225$
1	0	1	1	0	1	0	1	0	$(.9147)^2(.0853) - .069296 = .002072$
1	1	0	0	1	1	0	1	0	$(.9147)^2(.0853) - .069296 = .002072$
1	1	0	1	0	0	1	1	0	$(.9147)^2(.0853) - .069296 = .002072$
1	1	0	1	0	1	0	1	0	$(.9147)^3 - .03859271 = .7267149$