Hold-out Risk Bounds for Classifier Performance Evaluation

Mohak Shah MOHAK@CIM.MCGILL.CA
Centre for Intelligent Machines, McGill University, Montreal, H3A 2A7, Canada

Sara Shanian SARA.SHANIAN@IFT.ULAVAL.CA
IFT-GLO, Pav. Adrien Pouliot, Laval University, Quebec, G1K 7P4, Canada

Abstract

We present an empirical study of the generalization error bounds on the empirical risk of classifier on a test set. We show how this approach, by modeling the empirical risk as a binomial, can be used to obtain realistic confidence intervals that lie strictly in the $[0, 1]$ interval. This is in contrast to the traditional confidence interval approach that impose an asymptotic Gaussian assumption on the empirical risk which rarely holds for low risk-values resulting in unrealistic estimates on the limits of the intervals.

1. Introduction

One of the most common techniques of evaluating the performance of a machine learning algorithm is its empirical evaluation on a separate set of test examples (not used for training the algorithm). This is generally referred to as the hold-out testing. In this case, either the full dataset is divided into a training and a hold-out set or such a division is already provided with the data description in case an empirical evaluation estimate is desired specifically on the chosen hold-out set. In either case, a learning algorithm is trained on the training set using an apt model selection strategy and the classifier output by the learning algorithm after the training is then tested on the held-out dataset. Further one aims to provide a confidence interval around the performance estimate of the learned classifier on the test set. Naturally, to do so, we assume that the test set is representative of the underlying distribution of the test data. Providing such confidence interval around the empirical risk estimate of the chosen classifier on the test data is the issue that we focus on here. The main aim of such evaluation is to answer the following questions:

- Given the observed accuracy of a learning algorithm over a limited sample of data, what can we say about the behavior of the learning algorithm over future unseen examples?
- Given that one learning algorithm outperforms another over some sample data, how probable is it that this learning algorithm is more accurate, in general?

The estimates on the future performance of the empirical risk of the classifier, or more appropriately the degree of deviation of the empirical risk from the true risk is generally obtained using a confidence interval in which we believe the true risk of the classifier to lie.

The most common method of obtaining such confidence interval relies on the assumption that the empirical risk of the classifier on the test data can be modeled, in the limit, as a Gaussian. Based on this assumption, the necessary statistics are obtained from testing the classifier on the test data. That is, the mean classification error and its corresponding variance on the test examples are obtained. A confidence interval is then provided in terms of a Gaussian around the mean empirical risk with its tails removed at twice the standard deviation estimate on either sides. This provides both a lower and an upper bound on the true risk of the classifier (effectively a 95% confidence interval).

However, there is a strong caveat in this approach. The confidence interval strategy described above relies very significantly on the Gaussian assumption. But the basis of this Gaussian assumption generally comes from the central limit theorem in the statistics theory. This results implies that given a true estimate of the data statistic, the sampling distribution of this statistic approaches a Gaussian distribution as the number of samplings reaches infinity. That is, the Gaussian assumption holds on a *fixed underlying statistic and that too asymptotically.* However, this might not, and

Appearing in *the 4^{th} Workshop on Evaluation Methods for Machine Learning in conjunction with 26^{th} Intl Conf on Machine Learning*, Montreal, Canada, 2009.

indeed is not, generally the case.

The risk in the case of classification is modeled as a zero-one loss. This is equivalent then to having an indicator function which is true when the classifier errs on an example. This would lead to a binomial distribution over a number of trials (tests of classifier on a number of samples). Further, the aim of learning is to obtain as low an empirical risk as possible. That is, we are interested in modeling the empirical risk of the classifier for lower values (values closer to zero). However, for smaller values of empirical risk a binomial distribution cannot be approximated by a Gaussian. This observation was also made by Langford [2005]. As a result, applying a Gaussian assumption results in estimates that are overly pessimistic when obtaining an upper bound and overly optimistic when obtaining a lower bound around the empirical risk. Langford [2005] also showed a comparison between the behavior of the two distributions with an empirical example of upper bounds on the risk of a decision tree classifier on test datasets.

Shah [2008] gave a qualitative analysis of this approach and discussed some important extension possibilities. In this paper, we further the empirical validation of the test set bound approach [Langford, 2005] by looking at the behavior of both the upper and the lower bounds on the true risk of the classifiers. This is analogous to providing a confidence interval around a binomial distribution. We compare this against the traditional Gaussian confidence interval approach and show on a range of classifiers and datasets, how the test set bound approach yields more realistic estimates as opposed to the Gaussian confidence intervals. For the purpose, we compare six binary classifiers on a total of 16 datasets from the UCI machine learning repository.

The rest of the paper is organized as follows. In the next section we give the hold-out bound on the true risk of the classifier. Section 3 then gives the empirical results of applying the test set risk bound approach and compares it against the Gaussian confidence interval approach. We conclude in Section 4.

2. An Hold-out Risk bound

We consider binary classification problems where the input space $\mathcal{X}$ consists of an arbitrary subset of $\mathbb{R}^n$ and the output space $\mathcal{Y} = \{-1, +1\}$. An example $\mathbf{z} \stackrel{\text{def}}{=} (\mathbf{x}, y)$ is an input-output pair where $\mathbf{x} \in \mathcal{X}$ and $y \in \mathcal{Y}$. We adopt the PAC setting where each example $\mathbf{z}$ is drawn according to a fixed, but unknown, probability distribution $\mathcal{D}$ on $\mathcal{X} \times \mathcal{Y}$. The true risk $R(f)$ of any classifier f is defined as the probability that it misclassifies an example drawn according to $\mathcal{D}$:

$$R(f) \stackrel{\text{def}}{=} \Pr_{(\mathbf{x},y)\sim\mathcal{D}}(f(\mathbf{x}) \neq y) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}} I(f(\mathbf{x}) \neq y)$$

where $I(a) = 1$ if predicate a is true and 0 otherwise. Given a classifier f, and a test set $T = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$ of m examples, the *empirical risk* $R_T(f)$ on T, of any classifier f, is defined according to:

$$R_T(f) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} I(f(\mathbf{x}_i) \neq y_i) \stackrel{\text{def}}{=} \mathbf{E}_{(\mathbf{x},y)\sim T} I(f(\mathbf{x}) \neq y)$$

Now, we model $R_T(f)$ as binomial. The distribution is defined as the probability of λ errors on a set of m examples with true risk of the classifier f being $R(f)$.

$$\Pr_{T\sim\mathcal{D}^m}(mR_T(f) = \lambda | R(f)) = \binom{m}{\lambda}(R(f))^{\lambda}(1-R(f))^{m-\lambda}$$

We use the cumulative which is the probability of λ or fewer errors on m examples.

$$\begin{aligned}\text{Bin}(m, \lambda, R(f)) &= \Pr_{T\sim\mathcal{D}^m}((mR_T(f) \leq k | R(f)) \\ &= \sum_{i=0}^{m}\binom{m}{\lambda}(R(f))^i(1-R(f))^{m-i}\end{aligned}$$

We define binomial inversion tail [Langford, 2005] as:

$$\overline{\text{Bin}}(m, \lambda, \delta) = \max\{p : \text{Bin}(m, \lambda, p) \geq \delta)\}$$

which is the largest true error such that the probability of observing λ or fewer errors is at least δ.

Then, the risk bound on the true risk of the classifier is defined as [Langford, 2005]:

Theorem 1 *For all classifiers f, for all $\mathcal{D}$, for all $\delta \in (0, 1]$:*

$$\Pr_{T\sim\mathcal{D}^m}(R(f) \leq \overline{\text{Bin}}(m, \lambda, \delta)) \geq 1 - \delta$$

From this result, it follows that $\overline{\text{Bin}}(m, \lambda, \delta)$ is the smallest upper bound which holds with probability at least $1 - \delta$, on the true risk $R(f)$ of any classifier f with an observed empirical risk $R_T(f)$ on a set of m examples.

In an analogous manner, a lower bound on $R(f)$ can be shown to be [Langford, 2005]:

Theorem 2 *For all classifiers f, for all $\mathcal{D}$, for all $\delta \in (0, 1]$:*

$$\Pr_{T\sim\mathcal{D}^m}(R(f) \geq \min_p\{p : 1-\text{Bin}(m, \lambda, \delta) \geq \delta\}) \geq 1-\delta$$

3. Empirical Results

In this section, we examine empirically how the estimates of the risk bounds around the empirical risk fare compared to the traditionally utilized method of obtaining confidence intervals around the empirical risk based on the Gaussian assumption. We compare six learning algorithms on 16 different datasets. The learning algorithms compared are the Support Vector Machine equiped with an radial basis function kernel, the Set Covering Machine for learning conjunctions of data-dependent balls [Marchand and Shawe-Taylor, 2002], Adaboost with decision stumps, Decision Trees and the Naive Bayes algorithms. With the exception of the SCM for which an in-house implementation was used, the other algortihms were the ones implemented in the Weka machine learning toolkit [Witten and Frank, 2005].

Each data set was divided into two parts, a training set S and a test set T. The training set was used to train the learning algorithm and perform model selection to obtain the best parameters from a pre-defined set of parameter values. The learning parameters of all algorithms were determined from the training set only. To do the model selection a 10-fold Cross Validation (CV) was used on the training set and the parameters with the best average CV error were chosen for each of the learning algorithms on each dataset. The parameters included the C and the γ values in the case of the SVM, the penalty parameter p and the best number of features s for the SCM, the confidence parameter for pruning C and the minimum leaf nodes in the case of Decision trees, and the number of iterations in the case of Adaboost. The algorithms were then trained with the chosen parameter values on the training set. The final classifier output by each algorithm was then tested on the test set. The details of the datasets are provided in Table 1. The columns $|S|$ and $|T|$ refers to the number of examples in the training and the test sets respectively. The column n refers to the number of attributes in each dataset. The results of testing each of the classifier on these datasets are presented in Table 2. The column labeled R_T denotes the empirical risk of the classifier on the test set, the columns CI_l and CI_u denote the lower and upper limits of the confidence interval obtained using the asymptotic Gaussian assumption on the sampling distribution of the empirical risk. These limits are the two standard deviation limits around the empirical risk. The variance of the risk is obtained on the test set data samples with the empirical risk assumed as the mean of the distribution. Finally, the B_l and B_u columns denote, respectively, the lower and upper intervals generated from computing the lower and upper risk bounds of Theorems 2 and 1 of Section 2 with $\delta = 0.025$. This value of δ is chosen to obtain the intervals comparable to the two standard deviations intervals obtained with the Gaussian assumption approach.

Data-Set	**|T|**	**|S|**	**n**
Usvotes	200	235	16
bupa	175	170	6
Credit	300	353	15
Glass	107	107	9
Haberman	150	144	3
HeartS	147	150	13
sonar	103	105	60
SonarM	104	104	60
BreastCancer	343	340	9
Wdbc	284	285	30
Tic-tac-toe	479	479	9
Ionosphere	175	176	34
Letter_AB	1055	500	16
Letter_OQ	1036	500	16
Letter_DO	1055	500	16
Mushroom	4062	4062	22

Table 1. Data Set Description

4. Discussion and Conclusion

As mentioned above, the risk bound technique can be considered as an alternate approach to obtain confidence intervals around the empirical risk of the classifiers. It is different from the traditional confidence interval technique in the sense that the empirical risk is modeled as a binomial distribution. In contrast, the classical approach to obtain confidence intervals makes an implicit use of the central limit theorem in imposing an asymptotic Gaussian assumption on the distribution of the empirical risk considering the true risk to be fixed. However, for lower values of the empirical risk (closer to zero), this assumption rarely, if ever, holds. As a result the limits of the confidence intervals obtained using the classical technique are either overly pessimistic (the upper limits) or overly optimistic (the lower limits). Moreover, the limits of these intervals are also not restricted to the $[0, 1]$ intervals rendering them meaningless in most scenarios. For instance, upper limits of the confidence interval around the empirical risk exceeding unity can hardly be interpreted. Indeed, the empirical risk of the classifier, by definition, should always be constrained in the $[0, 1]$, and so should be its true risk. Hence, obtaining confidence intervals that spill over this known interval do not make much sense. On the other hand, the risk bound approach is guaranteed to lie in the $[0, 1]$ interval. Moreover, as we also saw in the results of Table 2, this technique allows us to obtain tight inter-

Data-Set	A	R_T	B_l	B_u	CI_l	CI_u
USvotes	SVM	0.05	0.027	0.096	-0.407	0.507
	Ada	0.04	0.017	0.077	-0.352	0.432
	DT	0.055	0.027	0.096	-0.402	0.512
	DL	0.045	0.020	0.083	-0.370	0.460
	NB	0.07	0.038	0.114	-0.441	0.581
	SCM	0.105	0.066	0.156	-0.509	0.719
Bupa	SVM	0.352	0.235	0.376	-0.574	1.278
	Ada	0.291	0.225	0.364	-0.620	1.202
	DT	0.325	0.256	0.400	-0.614	1.264
	DL	0.325	0.256	0.400	-0.614	1.264
	NB	0.4	0.326	0.476	-0.582	1.382
	SCM	0.377	0.305	0.453	-0.595	1.349
Credit	SVM	0.183	0.141	0.231	-0.592	0.958
	Ada	0.17	0.129	0.217	-0.582	0.922
	DT	0.13	0.094	0.173	-0.543	0.803
	DL	0.193	0.150	0.242	-0.598	0.984
	NB	0.2	0.156	0.249	-0.603	1.003
	SCM	0.19	0.147	0.239	-0.596	0.976
Glass	SVM	0.168	0.102	0.252	-0.583	0.919
	Ada	0	0	0.033	0.0	0.0
	DT	0.186	0.118	0.273	-0.597	0.969
	DL	0.065	0.026	0.130	-0.431	0.561
	NB	0.299	0.214	0.395	-0.621	1.219
	SCM	0.215	0.141	0.304	-0.610	1.040
Haberman	SVM	0.273	0.203	0.352	-0.621	1.167
	Ada	0.233	0.185	0.330	-0.640	1.106
	DT	0.273	0.203	0.352	-0.621	1.167
	DL	0.273	0.203	0.352	-0.621	1.167
	NB	0.246	0.180	0.323	-0.619	1.111
	SCM	0.253	0.185	0.330	-0.619	1.125
HeartS	SVM	0.204	0.142	0.278	-0.604	1.012
	Ada	0.272	0.202	0.351	-0.621	1.165
	DT	0.197	0.136	0.270	-0.601	0.995
	DL	0.156	0.101	0.225	-0.574	0.886
	NB	0.136	0.085	0.202	-0.552	0.824
	SCM	0.190	0.130	0.263	-0.598	0.978
Sonar	SVM	0.116	0.061	0.194	-0.528	0.760
	Ada	0.135	0.076	0.217	-0.553	0.823
	DT	0.365	0.099	0.251	-0.581	0.911
	DL	0.281	0.197	0.378	-0.622	1.184
	NB	0.262	0.180	0.358	-0.621	1.145
	SCM	0.310	0.223	0.409	-0.620	1.240
SonarM	SVM	0.182	0.113	0.270	-0.594	0.958
	Ada	0.153	0.090	0.237	-0.615	0.921
	DT	0.365	0.273	0.465	-0.602	1.332
	DL	0.221	0.145	0.313	-0.613	1.055
	NB	0.269	0.186	0.365	-0.622	1.160
	SCM	0.403	0.308	0.504	-0.583	1.389
BreastCancer	SVM	0.038	0.020	0.063	-0.344	0.420
	Ada	0.049	0.029	0.078	-0.385	0.483
	DT	0.061	0.038	0.092	-0.419	0.541
	DL	0.046	0.026	0.074	-0.376	0.468
	NB	0.046	0.026	0.074	-0.376	0.468
	SCM	0.037	0.020	0.063	-0.345	0.419
Wdbc	SVM	0.070	0.043	0.106	-0.442	0.582
	Ada	0.042	0.022	0.072	-0.361	0.445
	DT	0.052	0.029	0.085	-0.396	0.500
	DL	0.059	0.035	0.094	-0.416	0.534
	NB	0.049	0.027	0.081	-0.384	0.482
	SCM	0.056	0.032	0.089	-0.406	0.518
Tic-Tac-Toe	SVM	0.062	0.042	0.088	-0.423	0.547
	Ada	0.016	0.007	0.326	-0.240	0.272
	DT	0.135	0.106	0.169	-0.550	0.820
	DL	0.048	0.030	0.071	-0.372	0.468
	NB	0.340	0.297	0.384	-0.608	1.288
	SCM	0.106	0.080	0.137	-0.511	0.723
Ionosphere	SVM	0.045	0.019	0.088	-0.373	0.463
	Ada	0.091	0.053	0.144	-0.487	0.669
	DT	0.091	0.053	0.144	-0.487	0.669
	DL	0.142	0.094	0.203	-0.559	0.843
	NB	0.16	0.109	0.222	-0.574	0.894
	SCM	0.24	0.178	0.310	-0.617	1.097
Letter-AB	SVM	0.001	0	0.005	-0.060	0.062
	Ada	3.8e-3	0.001	0.009	-0.119	0.126
	DT	0.017	0.010	0.026	-0.242	0.276
	DL	0.016	0.009	0.025	-0.235	0.267
	NB	0.080	0.064	0.098	-0.464	0.624
	SCM	0.029	0.020	0.041	-0.308	0.366
Letter-OQ	SVM	0.010	0.005	0.018	-0.195	0.215
	Ada	0.043	0.031	0.057	-0.364	0.450
	DT	0.077	0.061	0.095	-0.457	0.611
	DL	0.055	0.041	0.070	-0.401	0.511
	NB	0.157	0.135	0.180	-0.571	0.885
	SCM	0.109	0.090	0.129	-0.514	0.732
Letter-DO	SVM	0.013	0.007	0.022	-0.215	0.241
	Ada	0.024	0.016	0.035	-0.286	0.334
	DT	0.061	0.047	0.077	-0.420	0.542
	DL	0.054	0.042	0.070	-0.402	0.510
	NB	0.080	0.064	0.098	-0.464	0.624
	SCM	0.061	0.047	0.077	-0.420	0.542
Mushroom	SVM	0	0	0.0009	0.0	0.0
	Ada	0	0	0.0009	0.0	0.0
	DT	0	0	0.0009	0.0	0.0
	DL	0	0	0.0009	0.0	0.0
	NB	0.091	0.083	0.101	-0.486	0.668
	SCM	0.025	0.020	0.304	-0.287	0.337

Table 2. Results of various classifiers on UCI Datasets

vals in practice. The upper bound never results in an overly pessimistic estimate greater than 1 while the lower bound never becomes too optimistic. Further, the confidence interval technique can't yield a confidence interval in the case when the observed empirical risk is zero. This can be seen directly since the resulting Gaussian in this case has both a zero mean and a zero variance. Hence, in the case of zero empirical risk, the confidence interval technique becomes overly optimistic. The risk bound on the other hand, still yields a finite upper bound (of course very small since $R_T(f) = 0$).

Hence, we show empirically how a risk bound based approach yields more realistic estimates on the limits of the confidence intervals and make a case for its wider use. Currently, this approach is confined to certain learning theoretic venues and formats. However, the approach is promising and robust. We believe, that it should be the metric of choice for reporting results based on the hold out test set. However, with regard to obtaining such guarantees based on other data resampling techniques, such robust results are yet not available. As also discussed in [Shah, 2008], however, there are some approaches, such as the sample compression bounds [Marchand and Shawe-Taylor, 2002, Shah, 2006], for obtaining the training set bounds that can result in practical realizable bounds on the true risk of the classifier. Such techniques would not only enable a comparison of the behavior of the classifier on the data but can also take into account other characteristics of the learning algorithm such as the complexity of the hypothesis class.

References

John Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 3:273–306, 2005.

Mario Marchand and John Shawe-Taylor. The Set Covering Machine. *Journal of Machine Learning Reasearch*, 3:723–746, 2002.

Mohak Shah. *Sample Compression, Margins and Generalization: Extensions to the Set Covering Machine*. PhD thesis, SITE, University of Ottawa, Ottawa, Canada, May 2006.

Mohak Shah. Risk Bounds for Classifier Evaluation: Possibilities and Challenges. In *Proceedings of the 3rd Workshop on Evaluation Methods for Machine Learning*. in conjunction with 25th International Conference on Machine Learning Helsinki, Finland, 2008.

Ian H. Witten and Eibe Frank. *Data Mining: Practical machine learning tools and techniques, 2nd Ed.* Morgan Kaufmann, San Francisco, 2005.