On Estimating Model Accuracy with Repeated Cross-Validation.

Gitte Vanwinckelen GITTE.VANWINCKELEN@CS.KULEUVEN.BE
Hendrik Blockeel HENDRIK.BLOCKEEL@CS.KULEUVEN.BE
Department of Computer Science, KU Leuven, Celestijnenlaan 200A, B-3001 Heverlee, Belgium

Keywords: repeated cross-validation, predictive model evaluation, conditional prediction error

Abstract

Evaluation of predictive models is a ubiquitous task in machine learning and data mining. Cross-validation is often used as a means for evaluating models. There appears to be some confusion among researchers, however, about best practices for cross-validation, and about the interpretation of cross-validation results. In particular, repeated cross-validation is often advocated, and so is the reporting of standard deviations, confidence intervals, or an indication of "significance". In this paper, we argue that, under many practical circumstances, when the goal of the experiments is to see how well the model returned by a learner will perform in practice in a particular domain, repeated cross-validation is not useful, and the reporting of confidence intervals or significance is misleading. Our arguments are supported by experimental results.

1. Introduction

Evaluation of predictive models is a ubiquitous task in machine learning and data mining. The task is not as trivial as it may seem. It is generally known that, to get an unbiased estimate of the accuracy of a model learned via machine learning, one should test the model on unseen data, not on the training set. In some cases, the population accuracy can be estimated from the training error using mathematical formulas. For instance, in linear regression, no separate test set is needed to estimate the error of the model: This error can be estimated accurately from the training data, using the concept of "degrees of freedom" to transform a training error into an unbiased estimate for the population error. For many advanced data analysis methods, however, one can not mathematically derive an unbiased estimate of population accuracy from training set accuracy, and more empirical methods are needed.

A basic approach is to use hold-out sampling which splits the available data set into a training set to learn a model, and a test set to estimate the accuracy of this model on. This requires that the training and test sets are disjoint, and that the training set is no more representative for this test set than for the population as a whole.

Often, when a limited set of data is available, one wants to learn a model from the whole data set, in order to maximally exploit the available information. Unfortunately, that leaves no unseen data to evaluate the accuracy of the model. In this case, an often used procedure is to learn a model $\hat{f}$ from the whole data set, and estimate the population accuracy of $\hat{f}$ by using a resampling technique, such as cross-validation. Like any estimator, cross-validation has some bias and variance. The non-zero bias has been pointed out before by, for example, Hastie et al. (2011). Because its variance is known to be relatively high, it is often advocated to repeat the cross-validation a number of times and average out the results, or to add confidence intervals that indicate how accurate the estimates are.

There are obvious statistical problems with estimates based on repeated subsampling of one data set, and for this reason, one may doubt whether repeated cross-validation is all that useful. In this paper, we investigate this question. We start with clearly defining some concepts and terminology, showing that several types of experimental questions need to be distinguished, and results of cross-validation need to be interpreted carefully. Next, we show experimentally that for the questions that are most important in practice, it is not useful to conduct repeated cross-validation.

Appearing in *Proceedings of BeneLearn and PMLS 2012*.

2. Cross-validation based estimates

Consider the following problem. We have a data set S from some domain D. S is typically assumed to be a random sample drawn from a population P. The data consists of a set of predictor variables $\mathbf{X}$ which are in relation to a target variable Y as $Y = f(\mathbf{X})$. We also have a learner L, which, given a data set S, returns a model $\hat{f}(\mathbf{X})$. The loss function $l(Y, \hat{f}(\mathbf{X}))$ measures how well $\hat{f}$ approximates f and is a measure of the accuracy of L. We use the one-zero loss function which equals 0 if for a given $\mathbf{x}$, $\hat{f}(\mathbf{x})$ equals the real value $f(\mathbf{x})$, and which equals 1 otherwise. We can now consider several questions about L or $\hat{f}(\mathbf{X})$. Focusing on accuracy as the most important performance measure, we may be interested in estimating the following population parameters:

- $\alpha_1 = E[l(Y, \hat{f}(\mathbf{X}))]$: the mean accuracy of $\hat{f}(\mathbf{X})$ on P, taken over all data sets S' of the same size as S
- $\alpha_2 = E[l(Y, \hat{f}(\mathbf{X})|S]$: the accuracy of $\hat{f}(\mathbf{X})$ on P for a fixed sample S

α_1 is computed by computing the mean accuracy over all models $\hat{f}(\mathbf{X})$ that can be learned from data sets S' of the same size as S. For α_2 on the other hand, $\hat{f}(\mathbf{X})$ is a fixed model determined by the chosen S. α_1 is known as the unconditional prediction error, and indicates to some extent how well learner L is suited for this problem domain. α_2 is known as the conditional prediction error, and it indicates how well the specific model obtained by running L on the available data can be expected to perform. When it is the intention to deploy the model learned from S in practice, α_2 is the most relevant parameter. Therefore, we focus on estimating α_2.

As said, cross-validation is often used to estimate the performance of learners or models. k-fold cross-validation works as follows. The available data S is divided into k equally sized subsections S_i, also called folds. For each fold, a training set T_i is defined as $S \backslash S_i$, from which a model M_i is learned. Next, the accuracy of this model is computed on S_i, and finally the mean of all these accuracies is returned as an estimate $\widehat{A}$.

$\widehat{A}$ is usually interpreted as an estimate of the predictive accuracy of the model $\hat{f}(X)$ learned from the whole data set S. This estimate is pessimistically biased, because it really estimates the average accuracy of models learned from a subset of $(k-1)/k \cdot 100\%$ of the data, which is likely to be slightly less good than the accuracy of the more informed model that is learned from the whole data set. This type of bias can be minimized by performing *leave-one-out cross-validation*, which sets k to the number of instances in the data set.

In addition to bias, the results of a k-fold cross-validation also have high variance. If we run two different tenfold cross-validations for the same learner on the same data set S, but with a different random partitioning of S into subsets S_i, these two cross-validations can give quite different results. An estimate with smaller variance can be obtained by repeating the cross-validation several times, with different partitionings, and taking the average of the results obtained during each cross-validation.

Repeated cross-validation is often advocated, using as an argument the high variance of the result of a single cross-validation. However, while this procedure indeed reduces the variance of the estimates, it does not remove the bias. We now try to make this more precise.

We introduce the following notation:

- $\hat{A}$: the result returned by a single k-fold cross-validation
- $\mathcal{C}_k$: the population of all possible k-fold cross-validations over this particular data set S.
- μ_k: the mean of $\hat{A}$ taken over all possible k-fold cross-validations over S (i.e., taken over $\mathcal{C}_k$)
- α_3: the mean accuracy of $L(S')$ on P, taken over all S' of size $(k-1)/k|S|$

Repeated cross-validation boils down to repeatedly drawing an element from $\mathcal{C}_k$, say n times, and computing the average of all these results. It is clear that this average, $\bar{A} = \sum_{i=1}^{n} \hat{A}_i/n$, approximates μ_k as n goes to infinity: $E(\bar{A}) = \mu_k$ and $Var(\bar{A}) = \sigma_k^2/n$ with σ_k^2 the variance of $\hat{A}$ taken over $\mathcal{C}_k$.

Consequently, repeated cross-validation allows us to accurately estimate μ_k, the mean of all possible k-fold cross-validations over the given data set S. However, the parameter we are really interested in is α_2. It is unclear whether μ_k is a good estimator for α_2 and if it is not, whether this is because of bias or large variance.

One could argue that the estimator is biased due to the fact that $\bar{A}$ reflects the accuracy of models learned from only a proportion $(k-1)/k$ of the data. It estimates the accuracy of models learned using slightly less data than available in S. But this is perhaps only part of the truth: α_3 is the mean accuracy of models learned from equally few data, and it would be interesting to investigate whether, for a particular S, μ_k is

equal to α_3 (though there is no prior reason to believe it is higher or lower).

Thus, given a data set S, $\bar{A}$ asymptotically approximates μ_k, but not necessarily α_1, α_2 or α_3. It is uncertain whether it approximates any of the parameters we may be interested in, even the α_3 parameter that explicitly takes into account the differences in size of the training sets for the individual models.

Confidence intervals around $\bar{A}$ are sometimes constructed. When using the standard formula for confidence intervals, they are constructed so that they contain μ_k with a certain confidence, not any other parameter. It would be erroneous to interpret confidence intervals, based on repeated cross-validation, as "almost certainly containing α_2" (or any of the other parameters one might be interested in).

The conclusion is that statistical inference can easily be done for μ_k, but μ_k itself is of little interest; on the other hand, for the parameters we are interested in, the α_i, there is no guarantee that they will be estimated with higher precision as the number of cross-validation repetitions increases.

3. Related work

Several authors have discussed the experimental evaluation of learners, comparative or otherwise. We focus on those contributions that are most relevant for this work.

The fact that cross-validation based estimators have high variance and non-zero bias has been pointed out several times. Kohavi (1995) considers the goal of selecting the best learner among a set of possible learners, and, in this context, experimentally compares bootstrapping and cross-validation for varying numbers of folds. He studies the bias and variance of these methods, shows that k-fold cross-validation has smaller bias but higher variance as k increases, and concludes that, from the point of view of selecting the most suitable learner, stratified tenfold cross-validation is overall the best method, even when it is computationally possible to use more folds. He suggests that repeated stratified tenfold cross-validation may work even better, as it is likely to reduce variance, but he does not experiment with this.

Braga-Neto and Dougherty (2004) investigate cross-validation for estimating α_2 in the context of small-sample microarray classification. They provide a formal definition of the bias and variance of cross-validation estimates by looking at the *deviation distribution* of $\alpha_2 - \hat{A}$ for a certain data distribution. A k-fold cross-validation estimator is unbiased if $\mu_k, = E[\alpha_2 - \hat{A}] = 0$. A large spread of the variance $Var[\alpha_2 - \hat{A}]$ of the deviation distribution indicates a large variance of the estimator. Bias and variance are combined in the root-mean-square error $\sqrt{E[\alpha_2 - \hat{A}]}$ of the distribution. The focus of the paper lies on an investigation of the variance of cross-validation. The conclusion is that cross-validation estimators typically have high variance for small samples, which makes their use problematic for analysis on small microarray samples.

Hastie et al. (2011) also discuss the bias and variance of cross-validation, and the fallacies when using it for estimating model accuracy. They draw attention to the fact that dependencies are often unknowingly introduced between the training and the test set by first using test points to design the learner, and performing cross-validation afterwards. This leads to an overly optimistic accuracy estimate. The authors also empirically investigate that cross-validation typically results in a good estimate for α_1, but it does not for α_2.

Schaffer (1993) specifically views cross-validation as a meta-learning technique that allows us to choose which among a given set of learners is likely to give the best predictive model. He concludes that cross-validation selects the best learner in most cases. However, similar to any other learning technique, performance depends on the setting the learner is used in.

Dietterich (1998), in a very influential paper, showed that comparing learners on the basis of repeated resampling of the same data set can lead to very high Type-I errors. This result is quite generally known, and the paired t-test methods discussed in that paper are generally considered discredited. Still, based on current practice in cross-validation, it would seem that the underlying reasons for this result are less generally understood, since the construction of confidence intervals on the basis of repeated cross-validation can be expected to suffer from similar problems.

Repeated cross-validation is used quite frequently in the literature. Also, in the experience of the authors, reviewers sometimes insist that cross-validation experiments be based on repeated cross-validation, that confidence intervals are shown, or that it is indicated which of the cross–validation results are "significantly better" than previously published results. In the light of the above work, it would seem obvious that such information, at best, is not very informative, needs careful interpretation by the reader, and is prone to misinterpretation.

Since Dietterich's paper, there has been a series of re-

Table 1. Overview of the data sets and their properties.

data set	instances	attributes	classes
adult	48842	15	2
kropt	28056	7	18
letter	20000	17	26
krvskp	3196	37	2
mushroom	8124	23	2
nursery	12960	9	5
optdigits	5620	65	10
pageblocks	5473	11	5
pendigits	9737	17	10

sults where people propose more advanced methods for comparing learners (Alpaydin, 1999; Bouckaert, 2003; Demsar, 2006). While these methods are carefully designed, and are shown to improve upon previous methods in a number of ways, they suffer from the same risk as previous methods, namely that, the more complex a method is, the higher the risk that researchers will use it incorrectly, or interpret the result incorrectly.

4. Experiments

In order to test how accurate cross-validation based estimates are, we compare these estimates with the 'real' population accuracy α_2 of the model learned from the whole data set. To this purpose, we set up experiments as follows. We take a large data set D from an existing data repository; D will serve as our population P. Next, we create a data set S by randomly sampling n elements from D. We now act as if the learner has only the data set S available. Nevertheless, because we know the population P, we can evaluate any model learned from S on the population.

Two learners, C4.5, and Naive Bayes, are applied on the nine data sets shown in Table 1, which were selected from the UCI repository (Frank & Asuncion, 2010). We perform two experiments, with the size of S equal to 200 instances, and 1000 instances. An accuracy estimate $\widehat{A}$ is computed by performing tenfold cross-validation, $10\times$ repeated tenfold cross-validation, and $30\times$ repeated tenfold cross-validation. We also construct a 95% confidence interval CI around $\widehat{A}$ and investigate whether CI is a good interval estimate of α_2 by examining whether $\alpha_2 \in CI$.

Tables 2 and 3 show the results of these experiments. The symbol + indicates that α_2 is larger than the upper bound of CI, while $-$ indicates it is smaller than the lower bound of CI.

As can be seen from both tables, the length of the confidence interval decreases with the number of repetitions of cross-validation. However, this does not imply $\widehat{A}$ converges to α_2. On the contrary, while most of the confidence intervals for $\widehat{A}$ contain α_2 when using a single cross-validation, most of them do not when using repeated cross-validation. In fact, the number of intervals containing α_2 decreases with the number of repetitions. As mentioned before, repeated cross-validation improves the estimate of μ_k and this result demonstrates that μ_k is not necessarily close to α_2.

Another observation is that in most cases where $\alpha_2 \notin CI$, α_2 lies to the right of CI. This shows that there is a pessimistic bias, which is consistent with our expectations (as cross-validation models are learned from a subset of the data, they tend to be less accurate). It also shows that, for repeated cross-validation, this bias is often larger than half the width of the confidence interval; for a single cross-validation this is typically not the case.

One might argue that this problem can be avoided by giving the confidence intervals the same width as those constructed from a single cross-validation. In this case, one would expect the confidence intervals to contain α_2 about as frequently as when a single cross-validation is used, or even slightly more frequently, if the point estimates obtained are closer to α_2. However, when inspecting the tables, we see that this is certainly not always, and often only marginally, the case.

Lastly, we look at the influence of the sample size on the estimates. The slope of a learning curve is typically high around small training set sizes, and decreases with an increasing training set size. As a result, the pessimistic bias caused by not using all the available data should be large for small sample sizes and decreases with increasing sample size. Table 4 confirms this by showing that the difference between $\widehat{A}$ and α_2 is in most cases smaller for a sample size of 1000. However, a comparison of Table 2 and Table 3 shows that increasing the size of S from 200 instances to 1000 instances does not substantially increase the number of correct confidence intervals.

5. Conclusions

Repeated cross-validation is often advocated for the evaluation of models in machine learning, the argument being that cross-validation estimates have high variance, which can be reduced by using the mean of multiple cross-validations as an estimate. In this paper, we have argued that, due to the fact that the same data set is continuously resampled in cross-validation, this mean converges to another value than any of the values one might really be interested in estimating.

Table 2. The accuracy results for C4.5 and Naive Bayes (N.B.) with the sample size of S equal to 200 instances, computed on different data sets by tenfold cross-validation, 10× repeated tenfold cross-validation, and 30× repeated tenfold cross-validation. The last column shows the population accuracy α_2 computed on $D \setminus S$.

C4.5	cross-validation			10×cross-validation			30×cross-validation			Pop.
data set	$\widehat{A}$ (%)	95% CI		$\widehat{A}$ (%)	95% CI		$\widehat{A}$ (%)	95% CI		α_2 (%)
adult	72.0	(65.78, 78.22)		71.85	(69.88, 73.82)	+	72.13	(71.0, 73.27)	+	76.08
kropt	18.0	(12.68, 23.32)		17.4	(15.74, 19.06)		17.57	(16.6, 18.53)		16.68
letter	38.5	(31.76, 45.24)		39.7	(37.56, 41.84)	+	39.12	(37.88, 40.35)	+	44.81
krvskp	97.0	(94.64, 99.36)	-	96.9	(96.14, 97.66)	-	96.68	(96.23, 97.14)	-	93.93
mushroom	97.0	(94.64, 99.36)		97.0	(96.25, 97.75)	+	96.78	(96.34, 97.23)	+	98.56
nursery	85.0	(80.05, 89.95)		84.8	(83.23, 86.37)	-	84.75	(83.84, 85.66)	-	83.06
optdigits	63.5	(56.83, 70.17)	+	68.3	(66.26, 70.34)	+	68.47	(67.29, 69.64)	+	75.59
pageblocks	92.0	(88.24, 95.76)		92.25	(91.08, 93.42)		92.18	(91.5, 92.86)	+	92.89
pendigits	73.5	(67.38, 79.62)		74.75	(72.85, 76.65)	+	74.58	(73.48, 75.69)	+	76.87
N.B.	cross-validation			10×cross-validation			30×cross-validation			Pop.
data set	$\widehat{A}$ (%)	95% CI		$\widehat{A}$ (%)	95% CI		$\widehat{A}$ (%)	95% CI		α_2 (%)
adult	78.0	(72.26, 83.74)	+	79.0	(77.21, 80.79)	+	79.1	(78.07, 80.13)	+	84.06
kropt	24.0	(18.08, 29.92)		23.7	(21.84, 25.56)		23.37	(22.3, 24.44)		23.83
letter	42.5	(35.65, 49.35)		44.0	(41.82, 46.18)	+	44.58	(43.33, 45.84)	+	47.28
krvskp	92.5	(88.85, 96.15)	-	90.9	(89.64, 92.16)	-	90.53	(89.79, 91.27)	-	85.58
mushroom	90.5	(86.44, 94.56)		90.15	(88.84, 91.46)		90.12	(89.36, 90.87)	+	91.12
nursery	83.5	(78.36, 88.64)		83.35	(81.72, 84.98)	+	82.97	(82.02, 83.92)	+	87.08
optdigits	84.0	(78.92, 89.08)		85.5	(83.96, 87.04)		85.13	(84.23, 86.03)	+	86.53
pageblocks	90.5	(86.44, 94.56)		89.1	(87.73, 90.47)	+	89.62	(88.84, 90.39)	+	92.98
pendigits	81.5	(76.12, 86.88)		81.25	(79.54, 82.96)	+	81.32	(80.33, 82.3)	+	83.8

Table 3. The accuracy results with the sample size of S equal to 1000 instances.

C4.5	cross-validation			10×cross-validation			30×cross-validation			Pop.
data set	$\widehat{A}$ (%)	95% CI		$\widehat{A}$ (%)	95% CI		$\widehat{A}$ (%)	95% CI		α_2 (%)
adult	80.7	(78.25, 83.15)		80.52	(79.74, 81.3)	+	80.41	(79.96, 80.86)	+	82.3
kropt	27.5	(24.73, 30.27)	+	26.68	(25.81, 27.55)	+	26.46	(25.96, 26.96)	+	30.53
letter	63.3	(60.31, 66.29)		61.85	(60.9, 62.8)		61.7	(61.15, 62.25)		61.44
krvskp	97.6	(96.65, 98.55)		97.82	(97.53, 98.1)	-	97.83	(97.66, 97.99)	-	97.41
mushroom	98.7	(98.0, 99.4)		99.01	(98.82, 99.2)		98.96	(98.85, 99.07)		99.07
nursery	88.09	(86.08, 90.09)	+	87.54	(86.89, 88.19)	+	87.51	(87.13, 87.88)	+	91.21
optdigits	83.6	(81.31, 85.89)		83.3	(82.57, 84.03)	-	83.53	(83.11, 83.95)	-	82.42
pageblocks	95.9	(94.67, 97.13)		96.08	(95.7, 96.46)		95.9	(95.68, 96.12)		95.98
pendigits	86.6	(84.49, 88.71)		87.6	(86.95, 88.25)		87.65	(87.28, 88.03)		87.41
N.B.	cross-validation			10×cross-validation			30×cross-validation			Pop.
data set	$\widehat{A}$ (%)	95% CI		$\widehat{A}$ (%)	95% CI		$\widehat{A}$ (%)	95% CI		α_2 (%)
adult	82.4	(80.04, 84.76)		82.48	(81.73, 83.23)	+	82.48	(82.05, 82.91)	+	83.55
kropt	28.1	(25.31, 30.89)		27.67	(26.79, 28.55)	+	27.77	(27.26, 28.28)	+	29.3
letter	59.8	(56.76, 62.84)		58.98	(58.02, 59.94)		59.1	(58.55, 59.66)	-	58.5
krvskp	86.58	(84.47, 88.69)		86.69	(86.02, 87.35)	+	86.9	(86.52, 87.28)	+	87.8
mushroom	94.1	(92.64, 95.56)		94.02	(93.56, 94.48)		94.02	(93.75, 94.29)		94.02
nursery	88.48	(86.5, 90.46)	+	88.65	(88.03, 89.27)	+	88.75	(88.39, 89.11)	+	90.67
optdigits	90.9	(89.12, 92.68)		90.46	(89.88, 91.04)	-	90.59	(90.26, 90.92)	-	89.48
pageblocks	93.29	(91.74, 94.85)		93.35	(92.87, 93.84)	+	93.5	(93.22, 93.78)	+	94.1
pendigits	87.3	(85.24, 89.36)		86.99	(86.33, 87.65)	-	87.13	(86.75, 87.51)	-	85.97

Table 4. A comparison of the differences between $\widehat{A}$ and α_2 for sample sizes 200 and 1000. The symbol '*' indicates a case where the difference is smallest for sample size 200.

C4.5									
	cross-validation $\lvert\widehat{A}-\alpha_2\rvert$ (%)			10×cross-validation $\lvert\widehat{A}-\alpha_2\rvert$ (%)			30×cross-validation $\lvert\widehat{A}-\alpha_2\rvert$ (%)		
$\lvert S\rvert$	200	1000		200	1000		200	1000	
adult	4.08	1.08		4.23	0.41		3.95	1.91	
kropt	1.32	3.02	*	0.72	3.81	*	0.89	1.66	*
letter	6.31	2.40		5.11	1.82		5.69	2.5	
krvskp	3.07	1.18		2.97	0.9		2.75	1.35	
mushroom	1.56	0.16		1.56	0.15		1.78	0.48	
nursery	1.94	0.92		1.74	0.43		1.69	2.31	*
optdigits	12.09	0.42		7.29	0.55		7.12	0.45	
pageblocks	0.89	0.58		0.64	0.85	*	0.71	0.28	
pendigits	3.37	1.29		2.12	0.76		2.29	0.16	
Naive Bayes									
adult	6.06	0.84		5.06	0.99		4.96	0.96	
kropt	0.17	1.25	*	5.06	1.21		0.46	2.98	*
letter	4.78	0.99		0.13	0.61	*	2.7	1.89	
krvskp	6.92	0.80		3.28	0.42		4.95	0.3	
mushroom	0.62	1.06	*	5.32	1.02		1	0.24	
nursery	3.58	0.18		0.97	0.32		4.11	0.47	
optdigits	2.53	0.67		3.73	0.06		1.4	0.76	
pageblocks	2.48	2.27		1.03	2.02	*	3.36	12.03	*
pendigits	2.30	0.23		3.88	0.42		2.48	3.7	*

Repeated cross-validation should not be assumed to give much more precise estimates of a model's predictive accuracy. The pessimistic bias due to the fact that cross-validation models are learned from smaller data sets (in the paper's notation, $\alpha_3 - \alpha_2$), together with the bias introduced by using a single data set ($\mu_k - \alpha_3$), can easily dominate the estimation error, which means reducing the variance is, in many cases, not very useful, and essentially a waste of computational resources.

Acknowledgments

This research was supported by the Research Foundation-Flanders (FWO Vlaanderen); Project G.0179.10 Multi-objective compiler optimization space exploration).

References

Alpaydin, E. (1999). Combined 5×2 cv f test for comparing supervised classification learning algorithms. *Neural Computation*, *11*, 1885–1892.

Bouckaert, R. R. (2003). Choosing between two learning algorithms based on calibrated tests. *International Conference on Machine Learning* (pp. 51–58).

Braga-Neto, U., & Dougherty, E. R. (2004). Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, *20*, 374–380.

Demsar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1–30.

Dietterich, T. G. (1998). Approximate statistical test for comparing supervised classification learning algorithms. *Neural Computation*, *10*, 1895–1923.

Frank, A., & Asuncion, A. (2010). UCI machine learning repository.

Hastie, T., Tibshirani, R., & Friedman, J. (2011). *The elements of statistical learning*. Springer Series in Statistics. Springer New York Inc. Second edition.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial intelligence, 2* (pp. 1137–1143). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Schaffer, C. (1993). Selecting a classification method by cross-validation. *Machine Learning, 13* (pp. 135–143).