Assessing and Comparing Classification Algorithms

Ethem Alpaydın
Department of Computer Engineering
Boğaziçi University
TR-80815 Istanbul, Turkey
alpaydin@boun.edu.tr

Abstract

Machine learning algorithms induce classifiers that depend on the training set and hyperparameters and there is a need for statistical testing for (i) assessing the expected error rate of a classifier, and (ii) comparing the expected error rates of two classifiers. We review interval estimation and hypothesis testing and discuss three tests for error rate assessment and four tests for error rate comparison.

1 Introduction

In the machine learning literature, there are several classification algorithms and given a certain application, more than one is applicable. In this paper, we are concerned with two questions:

1. How can we assess the error rate of a classifier induced by an algorithm ?
2. Given two classification algorithms, how can we say one is better than the other one, for a given application ?

The classifiers compared can be trained with different learning algorithms, or they can use the same algorithm with different hyperparameter settings. For example, given a multilayer perceptron with four hidden units and another one with eight hidden units, we would like to be able say which one has less error.

It should be clear that we cannot look at the training set accuracies of classifiers and decide based on that, i.e., choose the less erroneous one. This is because over the training set, the more complex one having more parameters, will almost always give less error than a simple one. So we need a validation set different from the training set to assess the performance of a classifier, or compare two classifiers. Even over a validation set though, just one comparison is not enough. There are two reasons for this: First, the validation set may be small and may contain exceptional instances which may mislead us. The second reason is that the learning method may depend on other factors which may affect generalization. For example with a multilayer perceptron trained with backpropagation, the initial weights affect the final weights and given the exact same architecture and training set, there may be multiple possible final classifiers.

So we should have a number of training/validation set pairs and we should run the two methods over all these pairs and then compare their errors on the validation sets. Therefore, our first need is to get a number of training/validation set pairs from a dataset D. If D is large enough, we can randomly divide it into K parts then divide each part randomly into two and use one for training and the other for validation. Unfortunately, datasets are never large enough to do this. So we should do our best with small datasets. This is done by repeated use of the same data splitted differently though this makes the error percentages dependent as they share data. Different tests use different splits of the data for this purpose, as we will see shortly. Once we get a number of validation error percentages from the two classifiers, we can define a statistic from them which, according to some simplifying assumptions, obeys a certain distribution. Then, the problem reduces to *hypothesis testing*, and this is based on *interval estimation*.

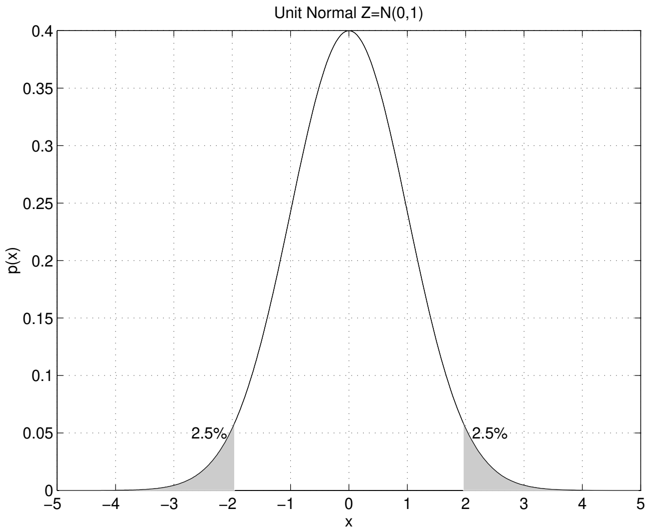

Figure 1: 95% of a unit normal distribution lies between -1.96 and 1.96.

2 Interval Estimation

A point estimator, e.g., maximum likelihood, specifies a value for a parameter θ. In interval estimation, we specify an interval within which θ lies with a certain degree of confidence. To obtain such an interval estimator, we make use of the probability distribution of the point estimator.

2.1 Interval Estimate of the Mean of a Normal Population with Known Variance σ^2

For example, say we are trying to estimate the mean μ of a normal density from a sample $\mathcal{X} = \{x^t\}_{t=1}^N$. $m =$

$\sum_t x^t/N$ is a sum of normals and thus is also normal, $m \sim \mathcal{N}(\mu, \sigma^2/N)$. Thus $\sqrt{N}(m-\mu)/\sigma$ has a unit normal distribution. Knowing that 95% of $\mathcal{Z}$ lies in $(-1.96, 1.96)$, i.e., $P\{-1.96 < \mathcal{Z} < 1.96\} = 0.95$, we have (Fig. 1)

$$P\left\{-1.96 < \sqrt{N}\frac{(m-\mu)}{\sigma} < 1.96\right\} = 0.95$$

or equivalently

$$P\left\{m - 1.96\frac{\sigma}{\sqrt{N}} < \mu < m + 1.96\frac{\sigma}{\sqrt{N}}\right\} = 0.95$$

That is 95 percent of the time, or "with 95 percent confidence," μ will lie within $1.96\sigma/\sqrt{N}$ units of the sample average. With 99 percent confidence, μ will lie in $(m - 2.58\sigma/\sqrt{N}, m + 2.58\sigma/\sqrt{N})$, thus if we want more confidence, the interval gets larger. The interval gets smaller as N, sample size, increases. This is a *two-sided confidence interval*.

This can be generalized as follows: Let us denote z_α such that

$$P\{\mathcal{Z} > z_\alpha\} = \alpha,\ 0 < \alpha < 1$$

where $\mathcal{Z}$ is unit normal. Note that because $\mathcal{Z}$ is symmetric around the mean, $P\{X < -z_{\alpha/2}\} = P\{X > z_{\alpha/2}\} = \alpha/2$. Hence for any specified level of confidence $1-\alpha$

$$P\left\{-z_{\alpha/2} < \mathcal{Z} < z_{\alpha/2}\right\} = 1 - \alpha$$

and

$$P\left\{-z_{\alpha/2} < \sqrt{N}\frac{(m-\mu)}{\sigma} < z_{\alpha/2}\right\} = 1 - \alpha$$

or

$$P\left\{m - z_{\alpha/2}\frac{\sigma}{\sqrt{N}} < \mu < m + z_{\alpha/2}\frac{\sigma}{\sqrt{N}}\right\} = 1 - \alpha \quad (1)$$

Hence a $100(1-\alpha)$ percent two-sided confidence interval for μ is computed.

Similarly, knowing that $P\{\mathcal{Z} < 1.64\} = 0.95$, we have (Fig. 2)

$$P\left\{\sqrt{N}\frac{(m-\mu)}{\sigma} < 1.64\right\} = 0.95$$

or

$$P\left\{m - 1.64\frac{\sigma}{\sqrt{N}} < \mu\right\} = 0.95$$

and a 95 percent *one-sided confidence interval* for μ is $(m - 1.64\sigma/\sqrt{N}, \infty)$. Or in the general case, a $100(1-\alpha)$ percent one-sided confidence interval for μ can be computed from

$$P\left\{m - z_\alpha\frac{\sigma}{\sqrt{N}} < \mu\right\} = 1 - \alpha \quad (2)$$

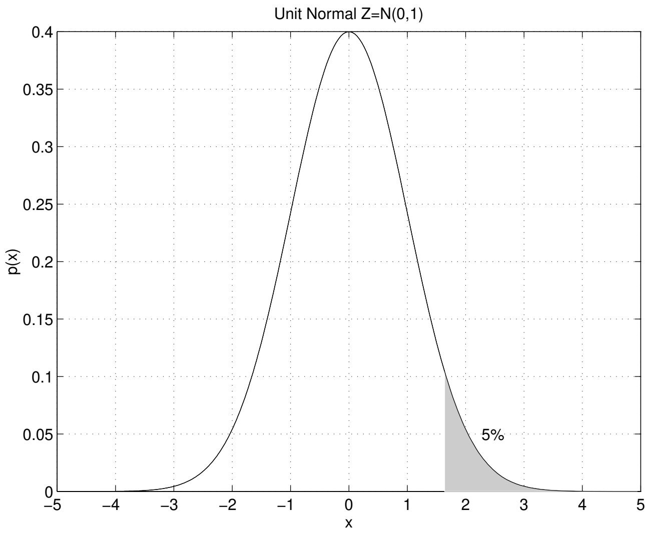

Figure 2: 95% of a unit normal distribution lies before 1.64.

2.2 Interval Estimate of the Mean of a Normal Population with Unknown Variance

If the variance, σ^2, is not known, one can plug the sample variance

$$S^2 = \sum_t (x^t - m)^2/(N-1)$$

instead of σ^2 and make use of the fact that

$$\frac{\sqrt{N}(m-\mu)}{S} \sim t_{N-1}$$

Hence for any $\alpha \in (0, 1/2)$

$$P\left\{t_{1-\alpha/2,N-1} < \sqrt{N}\frac{(m-\mu)}{S} < t_{\alpha/2,N-1}\right\} = 1 - \alpha$$

or using that $t_{1-\alpha/2,N-1} = -t_{\alpha/2,N-1}$

$$P\left\{m - t_{\alpha/2,N-1}\frac{S}{\sqrt{N}} < \mu < m + t_{\alpha/2,N-1}\frac{S}{\sqrt{N}}\right\} = 1-\alpha$$

3 Hypothesis Testing

In certain applications instead of explicitly estimating some parameters, we may want to use the sample to test some particular hypothesis concerning the parameters. For example instead of estimating the mean, we may want to test whether the mean is less than 1. If the random sample is consistent with the hypothesis under consideration, we say that the hypothesis is "accepted," otherwise we say that it is "rejected." But when we make such a decision, we are not really saying that it is true but rather that the resulting data appear to be consistent with it with a given degree of confidence. The procedure is very similar to that used in interval estimates.

3.1 Hypothesis Test of the Mean of a Normal Population with Known Variance σ^2

Let us say we have a sample from a normal distribution with unknown mean μ and known variance σ^2 and we want to test a specific hypothesis about μ, for example whether it is equal to a specified constant μ_0. It is denoted as H_0 and is called the *null hypothesis*

$$H_0 : \mu = \mu_0$$

against the alternative hypothesis

$$H_1 : \mu \neq \mu_0$$

m is the point estimate of μ and it is reasonable to accept H_0 if m is not too far from μ_0. We accept the hypothesis with $(1-\alpha)$ degree of significance if μ_0 lies in the $100(1-\alpha)$ percent confidence interval, i.e., H_0 is accepted if

$$\frac{\sqrt{N}(m-\mu_0)}{\sigma} \in (-z_{\alpha/2}, z_{\alpha/2}) \tag{3}$$

If we reject when the hypothesis is correct, this is a *type I error* and thus α, set before the test, defines how much type I error we can tolerate. *Type II error* is the probability of accepting the null hypothesis when the true mean μ is unequal to μ_0. The probability that H_0 is accepted when the true mean is μ is a function of μ and is given as

$$\beta(\mu) = P_\mu \left\{ -z_{\alpha/2} \leq \frac{m-\mu_0}{\sigma/\sqrt{N}} \leq z_{\alpha/2} \right\} \tag{4}$$

$1-\beta(\mu)$ is called the *power function* of the test and is equal to the probability of rejection when μ is the true value (Table 1).

Table 1: Type I, Type II errors and power of a test.

	Decision	
Truth	Accept	Reject
True	Correct	Type I error
False	Type II error	Correct (Power)

One can also have a *one-sided test* of the form

$$\begin{aligned} H_0 &: \mu \leq \mu_0 \\ H_1 &: \mu > \mu_0 \end{aligned}$$

as opposed to the *two-sided* problem when the alternative hypothesis is $\mu \neq \mu_0$. The $1-\alpha$ degree of significance one-sided test defines the $100(1-\alpha)$ confidence interval bounded on one side in which m should lie for the hypothesis to be accepted. We accept if

$$\frac{\sqrt{N}}{\sigma}(m-\mu_0) \in (-\infty, z_\alpha) \tag{5}$$

3.2 Hypothesis Test of the Mean of a Normal Population with Unknown Variance σ^2

For $H_0 : \mu = \mu_0$ and $H_1 : \mu \neq \mu_0$, if the variance is unknown, just like we did in the interval estimates, we use the sample variance instead of the population variance and the fact that

$$\frac{\sqrt{N}(m-\mu_0)}{S} \sim t_{N-1}$$

Then, we accept if

$$\frac{\sqrt{N}(m-\mu_0)}{S} \in (-t_{\alpha/2,N-1}, t_{\alpha/2,N-1}) \tag{6}$$

This is known as the *two-sided t-test*. A one-sided test can similarly be defined.

4 Assessing Classifier Performance

Let us say a classifier has a probability p of making an error. If we denote by x^t the correctness of the classifier's decision for instance with index t, x^t is Bernoulli distributed where with probability p, it commits an error and x^t takes the value 1 and with probability $1-p$, it is successful and x^t is 0.

We would like to test if the error probability p is less than or equal to some specified value p_0. We have the test

$$H_0 : p \leq p_0 \text{ vs. } H_1 : p > p_0$$

4.1 Binomial Test

Let X denote the number of errors on a sample of size N

$$X = \sum_{t=1}^{N} x^t$$

x^t are independent Bernoullis and X is Binomial. If the probability of error is p, the probability that the classifier commits k errors on a validation sample of size N is

$$P\{X=k\} = \binom{N}{k} p^k (1-p)^{N-k}$$

Under the null hypothesis, we assume that p is p_0 (or less), and the probability that there are e errors or less is

$$P\{X \leq e\} = \sum_{k=1}^{e} \binom{N}{k} {p_0}^k (1-p_0)^{N-k}$$

If this probability is less than the significance level $1-\alpha$, we accept the hypothesis; otherwise we reject it.

4.2 Approximate Normal Test

X is the sum of independent Bernoullis and for large N and small p, X is approximately normally distributed with mean Np and variance $Np(1-p)$. Then

$$\frac{X-Np}{\sqrt{Np(1-p)}}$$

is approximately unit normal distributed. Then an approximate test accepts the null hypothesis if this value is less than or equal to z_α. $z_{0.05}$ is 1.96. This is an approximate test and may give erroneous results; do not use it if $Np_0 > 20$.

4.3 t Test

If we run the algorithm k times, on k training/validation set pairs, we get k error percentages $x^i, i = 1, \ldots, k$. Then their average

$$m = \frac{\sum_{i=1}^k x^i}{k}$$

is also normal distributed with mean Np and variance $Np(1-p)/k$. Thus the following statistic

$$\frac{\sqrt{k}(m - Np)}{\sqrt{Np(1-p)}}$$

is also unit normal. The sample variance of e_i is

$$S = \frac{\sum_{i=1}^k (x^i - m)}{k-1}$$

and

$$\frac{\sqrt{k}(m - \mu_e)}{S} \sim t_{k-1}$$

and we accept the null hypothesis that the classifier has μ_e or less error percentage if this value is less than or equal to $t_{\alpha,k-1}$. Typically k is taken as 10 or 30. $t_{0.05,9} = 1.83$ and $t_{0.05,29} = 1.70$.

5 Comparing Two Classifiers

Given two learning algorithms and a training set, we want to test if the two algorithms construct classifiers that have the same error rate on a new example.

5.1 McNemar's Test

Given a training set and a validation set, we train both classifiers on the training set and test them on the validation set and compute the contigency table of the two classifiers:

N_{00}: Number of examples misclassified by both	N_{01}: Number of examples misclassified by 1 but not 2
N_{10}: Number of examples misclassified by 2 but not 1	N_{11}: Number of examples correctly classified by both

We have

$$\frac{(|N_{01} - N_{10}| - 1)^2}{N_{01} + N_{10}} \sim \mathcal{X}^2$$

and we accept if this value is less than or equal to $\mathcal{X}^2_{1,\alpha}$. $\mathcal{X}^2_{1,0.05} = 3.84$.

5.2 K-Fold Cross-Validated Paired t Test

This set uses k-fold cross-validation to get k training/validation set pairs as follows: The dataset D is randomly divided into k parts. To get each pair, $k-1$ of the parts are combined to form the training set and the remaining one part is the validation set. Therefore the training set is $k-1$ times as large as the validation set. This is done k times where for each pair, another of the k parts is left out as the validation set and we get k training/validation set pairs.

Then given pair $i = 1, \ldots, k$, the two classifiers are trained on the training set and are tested on the validation set, their errors are recorded as x_i^1 and x_i^2. The difference is $x^i = x_i^1 - x_i^2$. When this is done k times, we have a distribution of x^i containing k points.

Given that x_i^1 and x_i^2 are both (approximately) normal, their difference x^i is also normal. The null hypothesis is that this distribution has zero mean:

$$\begin{aligned} H_0 &: \mu = 0 \\ H_1 &: \mu \neq 0 \end{aligned}$$

The estimator for the mean is

$$m = \frac{\sum_{i=1}^k x^i}{k}$$

and the variance is

$$S = \frac{\sum_{i=1}^k (x^i - m)^2}{k-1}$$

Under the null hypothesis that $\mu = 0$, we have

$$\frac{\sqrt{k}(m-0)}{S} = \frac{\sqrt{k} \cdot m}{S} \sim t_{k-1}$$

Thus we accept the hypothesis that two classifiers have the same error with $1-\alpha$ degree of confidence if this value is in the interval $(-t_{\alpha/2,k-1}, t_{\alpha/2,k-1})$. Typically, $t_{0.025,29} = 2.05$.

5.3 5×2 cv Paired t Test

In the 5×2 cv t test, proposed by Dietterich (1998), we perform 5 replications of 2-fold cross-validation. In each replication, the dataset is divided into two equal-sized sets. $p_i^{(j)}$ is the difference between the error rates of the two classifiers on fold $j = 1, 2$ of replication $i = 1, \ldots, 5$. The average on replication i is $\overline{p}_i = (p_i^{(1)} + p_i^{(2)})/2$ and the estimated variance is $s_i^2 = (p_i^{(1)} - \overline{p}_i)^2 + (p_i^{(2)} - \overline{p}_i)^2$.

Under the null hypothesis, $p_i^{(j)}$ is the difference of two identically distributed proportions, and ignoring the fact that these proportions are not independent, $p_i^{(j)}$ can be treated as approximately normal distributed with zero mean and unknown variance σ^2 (Dietterich, 1998). Then $p_i^{(j)}/\sigma$ is approximately unit normal. If we assume $p_i^{(1)}$ and $p_i^{(2)}$ are independent normals (which is not strictly true because their training and test sets are not drawn independently of each other), then s_i^2/σ^2 has a chi-square distribution with one degree of freedom. If each of the s_i^2 are assumed to be independent (which is not true because they are all computed from the same set of available data) then

$$M = \frac{\sum_{i=1}^{5} s_i^2}{\sigma^2} \sim \mathcal{X}_5^2$$

and

$$t = \frac{p_1^{(1)}}{\sqrt{M/5}} = \frac{p_1^{(1)}}{\sqrt{\sum_{i=1}^{5} s_i^2/5}} \sim t_5 \tag{7}$$

We accept the hypothesis if this value is in the interval $(-t_{\alpha/2,5}, t_{\alpha/2,5})$. $t_{0.025,5} = 2.57$.

5.4 5×2 **cv Paired** F **Test**

We note that the numerator $p_1^{(1)}$ is arbitrary and actually there are ten different values that can be placed in the numerator, i.e., $p_i^{(j)}, j = 1, 2, i = 1, \ldots, 5$, leading to ten possible statistics

$$t_i^{(j)} = \frac{p_i^{(j)}}{\sqrt{\sum_{i=1}^{5} s_i^2/5}} \tag{8}$$

Alpaydın (1999) proposed an extension to the 5×2cv t test that combines the results of the ten possible statistics. If $p_i^{(j)}/\sigma \sim \mathcal{Z}$, then $\left(p_i^{(j)}\right)^2/\sigma^2 \sim \mathcal{X}_1^2$ and

$$N = \frac{\sum_{i=1}^{5} \sum_{j=1}^{2} \left(p_i^{(j)}\right)^2}{\sigma^2} \sim \mathcal{X}_{10}^2$$

Then

$$f = \frac{N/10}{M/5} = \frac{\sum_{i=1}^{5} \sum_{j=1}^{2} \left(p_i^{(j)}\right)^2}{2\sum_{i=1}^{5} s_i^2} \sim F_{10,5} \tag{9}$$

We accept the hypothesis that the algorithms have the same error rate if this value is less than $F_{\alpha,10,5}$. $F_{0.05,10,5} = 4.74$.

Acknowledgment

This work is supported by Boğaziçi University Research Funds Grant 00A101D.

References

1. E. Alpaydın (1999) "Combined 5×2cv F Test for Comparing Supervised Classification Learning Algorithms," *Neural Computation*, **11**(8), 1885–1982.

2. T. G. Dietterich (1998) "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Computation*, **10**, 1895–1923.

3. S. M. Ross (1987). *Introduction to Probability and Statistics for Engineers and Scientists*, John Wiley and Sons.