Fundamenta Informaticae 70 (2006) 1–15
IOS Press

Bounds for Validation

Wojciech Jaworski*
Faculty of Mathematics, Computer Science and Mechanics
Warsaw University, Banacha 2, 02-07 Warsaw, Poland
wjaworski@mimuw.edu.pl

Abstract. In this paper we derive the bounds for Validation (known also as Hold-Out Estimate and Train-and-Test Method). We present the best possible bound in the case of 0-1 valued loss function. We also provide the tables where the least sample size is calculated that is necessary for obtaining the bound for a given estimation rate and reliability of estimation. For an arbitrary bounded loss function we present the optimal bound approximation with any given accuracy.

Keywords: computational learning theory, model assessment, hold-out estimate, train-and-test, validation, Bernoulli sums

1. Introduction

We present the evaluation classifiers error rate method, known as Validation. Its purpose is to estimate the difference between the error value returned by Validation and the real error rate of the classifier, further called the generalization error.

We derive the best possible bounds in the case of 0-1 valued loss function. The bounds are presented in the following forms:

$$\mathrm{P}(\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f) < \varepsilon) > 1 - \delta,$$

and

$$\mathrm{P}(|\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)| < \varepsilon) > 1 - \delta,$$

where f is the classifier, $\mathcal{E}$ — the generalization error, $\mathcal{E}_{\mathbf{z}}$ — the empirical error, and $1-\delta$ is the reliability.

We also calculate the least sample size, necessary for obtaining the bound for a given estimation accuracy and reliability, and we illustrate the results in tables.

*Address for correspondence: Faculty of Mathematics, Computer Science and Mechanics, Warsaw University, Banacha 2, 02-07 Warsaw, Poland

For an arbitrary loss function we propose the algorithm based on concept of Fourier series that allows us to approximate the optimal error bound from both sides.

The paper is organised as follows: In Section 2, we recall the basic concepts of the learning theory in order to define Validation, generalization error, and empirical error. In Section 3 and 4, we deal with Validation for 0-1 loss function and in Section 5 we derive bound for Validation for arbitrary loss function.

2. The basic definitions

In this section, the fundamental concepts of the learning theory are introduced.

Let X be the *set of examples (attribute value vectors)*, Y be the *set of decisions (labels)*, and ρ be a Borel probability measure on $Z = X \times Y$. ρ plays an important role in sampling as for a given set it describes the probability of getting a sample that belongs to that set as well as distribution of decision for any example. Unfortunately, ρ is unknown to us.

It is given a finite sequence $\mathbf{z} = \big((x_1, y_1), \ldots, (x_m, y_m)\big)$, where x_i is an example and y_i – a decision for $i = 1, \ldots, m$. The sequence $\mathbf{z}$ will be called a *sample* of the length m; $\mathbf{z}$ is randomly generated by m independent draws according to the probability measure ρ; $\mathbf{z}$ represents all our knowledge about ρ.

An algorithm $A_m : Z^m \to (X \to Y)$ is a function such that for each sample $\mathbf{z}$ of the length m, A_m yields a *classifier* (i.e., a function) $f_{\mathbf{z}} : X \to Y$.

Having a classifier, we want to evaluate its quality. The quality of a classifier f is determined by its *generalization error* defined by

$$\mathcal{E}(f) = \int_Z V\big(y, f(x)\big)\, d\rho(x, y),$$

where $V : Y \times Y \to \mathbb{R}_+$ is called the *loss function*. For example, the loss function can be defined by:

$$V\big(y, f(x)\big) = (y - f(x))^2,$$

$$V\big(y, f(x)\big) = |y - f(x)|,$$

or

$$V\big(y, f(x)\big) = \begin{cases} 0 & \text{if } y = f(x) \\ 1 & \text{if } y \neq f(x). \end{cases}$$

For a finite set of decisions $Y = \{d_1, d_2, \ldots, d_l\}$, the last case may be generalized to

$$V\big(d_i, d_j\big) = a_{i,j},$$

where $a_{i,i} = 0$ and $0 \leq a_{i,j} \leq 1$. Such a loss function allows us to express the fact that we prefer one type of the classifier error to another.

We want to estimate $\mathcal{E}(f_{\mathbf{z}})$, which cannot be calculated directly. To this end, we use the generalization error evaluators such as Validation.

The idea of *Validation* is to divide a given sample $\mathbf{z}$ into two distinct parts $\mathbf{z}_1, \mathbf{z}_2$. The sample $\mathbf{z}_1$ will be used to learn the classifier and the sample $\mathbf{z}_2 = \big((x'_1, y'_1), \ldots, (x'_{m'}, y'_{m'})\big)$ to test it by calculation of

$$\mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1}) = \frac{1}{m'} \sum_{i=1}^{m'} V\big(y'_i, f_{\mathbf{z}_1}(x'_i)\big).$$

$\mathcal{E}_{\mathbf{z}}(f)$ is called the *empirical error* of the function f on the sample $\mathbf{z}$. Having calculated $\mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1})$, we claim that its value is similar to the value of the *generalization error* of $f_{\mathbf{z}_1}$.

$$\mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1}) \sim \mathcal{E}(f_{\mathbf{z}_1}).$$

In the next sections, we will try to express this similarity by numeric means.

3. Optimal bounds for 0-1 loss function when ε is given

The simplest way to obtain the quality of estimation is to assess

$$|\mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1}) - \mathcal{E}(f_{\mathbf{z}_1})|$$

or at least

$$\mathcal{E}(f_{\mathbf{z}_1}) - \mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1}),$$

if we are interested only in how bad the estimation can be.

In case of 0-1 loss function $\mathcal{E}_{\mathbf{z}}(f)$ is the random variable with the binomial distribution we have:

$$\mathrm{P}(m\mathcal{E}_{\mathbf{z}}(f) = i) = \binom{m}{i} p^i (1-p)^{m-i},$$

where $p = \mathcal{E}(f)$ is the probability that the decision for an example is not consistent with the function f.

The following two lemmas are crucial for obtaining the bound:

Lemma 3.1. Let $\varepsilon > 0$.

$$\mathrm{P}(\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f) > \varepsilon) = \begin{cases} 0 & \text{if } p \in [0, \varepsilon] \\ \sum_{i=0}^{k} \binom{m}{i} p^i (1-p)^{m-i} & \text{if } p \in (\varepsilon + \frac{k}{m}, \varepsilon + \frac{k+1}{m}], \end{cases}$$

and

$$\mathrm{P}(\mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}(f) > \varepsilon) = \begin{cases} 0 & \text{if } p \in [1-\varepsilon, 1] \\ \sum_{i=l}^{m} \binom{m}{i} p^i (1-p)^{m-i} & \text{if } p \in [\frac{l-1}{m} - \varepsilon, \frac{l}{m} - \varepsilon). \end{cases}$$

Lemma 3.2. Let $\varepsilon > 0$. The function

$$\sum_{i=0}^{k} \binom{m}{i} p^i (1-p)^{m-i},$$

is monotonically decreasing in the interval $(\varepsilon + \frac{k}{m}, \varepsilon + \frac{k+1}{m}]$. The function

$$\sum_{i=l}^{m} \binom{m}{i} p^i (1-p)^{m-i},$$

is monotonically increasing in the interval $[\frac{l-1}{m} - \varepsilon, \frac{l}{m} - \varepsilon)$.

Theorem 3.1. Let m denote the size of $\mathbf{z}_2$, and let $\varepsilon > 0$. If $V\big(y, f_{\mathbf{z}_1}(x)\big) \in \{0, 1\}$, then the least δ such that

$$\mathrm{P}\big(\mathcal{E}(f_{\mathbf{z}_1}) - \mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1}) > \varepsilon\big) < \delta \tag{1}$$

has the value

$$\delta = \max_{k<(1-\varepsilon)m} \sum_{i=0}^{k} \binom{m}{i} \left(\varepsilon + \frac{k}{m}\right)^i \left(1 - \varepsilon - \frac{k}{m}\right)^{m-i}$$

Proof:
We split $[0, 1]$ into the distinct intervals by the sequence $\{\varepsilon + \frac{k}{m}\}_0^m$ and we apply for each interval Lemmas 3.1 and 3.2. □

We show the behaviour of the bound in Table 1.

It was the monotonicity of $p^i(1-p)^{m-i}$ that allowed us to find the maximum on the intervals by values on their borders. In case of two side bound we cannot use monotonicity. However, we can use convexity.

Lemma 3.3. Let $p_1, p_2 \in [\varepsilon + \frac{k}{m}, \varepsilon + \frac{k+1}{m}] \cap [\frac{l-1}{m} - \varepsilon, \frac{l}{m} - \varepsilon]$ and $p_1 < p_2$. If

$$\frac{1}{4\varepsilon^2} + 1 \leq m,$$

then the function

$$\sum_{i=0}^{k} \binom{m}{i} p^i(1-p)^{m-i} + \sum_{i=l}^{m} \binom{m}{i} p^i(1-p)^{m-i},$$

is convex on the interval $[p_1, p_2]$.

Proof:
If $i \leq k$, then the function $p^i(1-p)^{m-i}$ is convex on the interval

$$\left[\frac{i}{m} + \frac{\sqrt{i(m-i)}}{m\sqrt{m-1}}, 1\right],$$

and when $i \geq l$ the function $p^i(1-p)^{m-i}$ is convex on the interval

$$\left[0, \frac{i}{m} - \frac{\sqrt{i(m-i)}}{m\sqrt{m-1}}\right].$$

For each $i \leq k$ it must hold

$$\frac{i}{m} + \frac{\sqrt{i(m-i)}}{m\sqrt{m-1}} \leq \varepsilon + \frac{k}{m},$$

and for each $i \geq l$ it must hold

$$\frac{i}{m} - \frac{\sqrt{i(m-i)}}{m\sqrt{m-1}} \geq \frac{l}{m} - \varepsilon.$$

For above to hold it is enough to have

$$\frac{\sqrt{i(m-i)}}{m\sqrt{m-1}} \leq \varepsilon,$$

for each i. Left side has supremum at $i = \frac{m}{2}$, which gives us a condition on m. □

Table 1. Number of samples needed for inequality (1) to hold for given ε and δ

$\varepsilon \backslash \delta$	0.1000	0.0500	0.0200	0.0100	0.0050	0.0020	0.0010	0.0005	0.0002	0.0001
0.005	16624	27255	42379	54319	66549	83038	95695	108475	125522	138510
0.010	4206	6864	10645	13630	16687	20809	23974	27169	31430	34677
0.015	1891	3073	4753	6080	7439	9271	10677	12097	13991	15434
0.020	1076	1741	2686	3432	4197	5227	6018	6817	7882	8694
0.025	697	1122	1727	2205	2694	3353	3860	4371	5053	5572
0.030	489	785	1205	1537	1876	2334	2686	3041	3514	3875
0.035	364	581	889	1133	1383	1719	1977	2238	2586	2851
0.040	281	448	684	871	1062	1319	1517	1717	1983	2186
0.045	225	356	543	690	841	1045	1201	1359	1569	1729
0.050	184	291	442	561	683	848	975	1103	1273	1403
0.055	154	242	367	465	566	703	807	913	1054	1161
0.060	131	205	310	392	477	592	680	768	887	977
0.065	112	175	265	336	408	505	580	656	757	833
0.070	98	152	230	290	353	437	501	566	653	720
0.075	86	134	201	254	308	381	438	494	570	628
0.080	77	118	177	224	272	336	385	435	502	552
0.085	69	105	158	199	241	298	342	386	445	490
0.090	62	95	141	178	216	267	306	345	398	438
0.095	56	86	127	160	194	240	275	310	357	393
0.100	51	78	115	145	176	217	249	280	323	355
0.105	47	71	105	132	160	197	226	255	293	323
0.110	43	65	96	121	146	180	206	233	268	294
0.115	40	60	88	111	134	165	189	213	245	270
0.120	37	55	82	102	123	152	174	196	226	248
0.125	34	51	76	95	114	140	161	181	208	229
0.130	32	48	70	88	106	130	149	168	193	212
0.135	30	45	65	82	98	121	138	156	179	197
0.140	28	42	61	76	92	113	129	145	167	183
0.145	26	39	57	71	86	105	120	135	156	171
0.150	25	37	54	67	80	99	113	127	146	160
0.155	24	35	50	63	75	93	106	119	137	150
0.160	22	33	48	59	71	87	99	112	128	141
0.165	21	31	45	56	67	82	94	105	121	133
0.170	20	29	42	53	63	77	88	99	114	125
0.175	19	28	40	50	60	73	84	94	108	118
0.180	18	27	38	47	57	69	79	89	102	112
0.185	17	25	36	45	54	66	75	84	97	106
0.190	17	24	35	43	51	63	71	80	92	101
0.195	16	23	33	41	49	60	68	76	87	96
0.200	15	22	31	39	46	57	65	72	83	91

Lemma 3.4. Let $p_1, p_2 \in [\varepsilon + \frac{k}{m}, \varepsilon + \frac{k+1}{m}] \cap [\frac{l-1}{m} - \varepsilon, \frac{l}{m} - \varepsilon]$ and $p_1 < p_2$. If

$$\delta = \sup_{p \in [p_1, p_2]} \sum_{i=0}^{k} \binom{m}{i} p^i (1-p)^{m-i} + \sum_{i=l}^{m} \binom{m}{i} p^i (1-p)^{m-i},$$

and $\frac{1}{4\varepsilon^2} + 1 \leq m$, then

$$\delta = \max\Bigg(\sum_{i=0}^{k} \binom{m}{i} p_1^i (1-p_1)^{m-i} + \sum_{i=l}^{m} \binom{m}{i} p_1^i (1-p_1)^{m-i},$$

$$\sum_{i=0}^{k} \binom{m}{i} p_2^i (1-p_2)^{m-i} + \sum_{i=l}^{m} \binom{m}{i} p_2^i (1-p_2)^{m-i} \Bigg).$$

Theorem 3.2. Let m denote the size of $\mathbf{z}_2$, and let $\varepsilon > 0$. If $V\big(y, f_{\mathbf{z}_1}(x)\big) \in \{0, 1\}$ and m be such that $\frac{1}{4\varepsilon^2} + 1 \leq m$, then the least δ such that

$$\mathrm{P}\big(|\mathcal{E}(f_{\mathbf{z}_1}) - \mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1})| > \varepsilon\big) < \delta \tag{2}$$

satisfies

$$\delta = \max_{0 \leq k < m(1-\varepsilon)} \sum_{i=0}^{k} \binom{m}{i} \left(\varepsilon + \frac{k}{m}\right)^i \left(1 - \varepsilon - \frac{k}{m}\right)^{m-i} +$$

$$+ \sum_{i=k+\lfloor 2m\varepsilon \rfloor + 1}^{m} \binom{m}{i} \left(\varepsilon + \frac{k}{m}\right)^i \left(1 - \varepsilon - \frac{k}{m}\right)^{m-i}.$$

Proof:
Analogously to the proof of Theorem 3.1 we we split $[0, 1]$ into distinct intervals by sequences $\{\varepsilon + \frac{k}{m}\}_{k=0}^{m}$ and $\{\frac{l}{m} - \varepsilon\}_{l=0}^{m}$.

$$\mathrm{P}\big(|\mathcal{E}(f_{\mathbf{z}_1}) - \mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1})| > \varepsilon\big) = \mathrm{P}\big(\mathcal{E}(f_{\mathbf{z}_1}) - \mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1}) > \varepsilon\big) + \mathrm{P}\big(\mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1}) - \mathcal{E}(f_{\mathbf{z}_1}) > \varepsilon\big).$$

Hence from Lemma 3.1 and Lemma 3.4 it is enough to calculate maximum in the proximity of $\{\varepsilon + \frac{k}{m}\}_{k=0}^{m}$ and $\{\frac{l}{m} - \varepsilon\}_{l=0}^{m}$. Points $\frac{l}{m} - \varepsilon$ can be eliminated using the symmetry of the problem. Now we observe that the probability is always bigger on the right side of $\varepsilon + \frac{k}{m}$ point than on its left side. □

Having the bound derived, we can see in Table 2 how it behaves. One can observe that the necessary number of samples for the two side bound is only slightly larger than the number of samples in the one side bound.

Remark 3.1. Luckily the condition

$$\frac{1}{4\varepsilon^2} + 1 \leq m,$$

is satisfied in the interesting cases.

Table 2. Number of samples needed for inequality (2) to hold for given ε and δ

$\varepsilon \backslash \delta$	0.1000	0.0500	0.0200	0.0100	0.0050	0.0020	0.0010	0.0005	0.0002	0.0001
0.005	27100	38500	54200	66400	78800	95500	108300	121200	138400	151400
0.010	6800	9650	13550	16600	19700	23900	27100	30300	34600	37850
0.015	3034	4300	6034	7400	8767	10634	12034	13467	15400	16834
0.020	1700	2425	3400	4150	4925	5975	6775	7575	8650	9475
0.025	1100	1540	2180	2660	3160	3820	4340	4860	5540	6060
0.030	767	1084	1517	1850	2200	2667	3017	3367	3850	4217
0.035	558	786	1115	1358	1615	1958	2215	2486	2829	3100
0.040	425	613	850	1038	1238	1500	1700	1900	2163	2375
0.045	345	478	678	823	978	1189	1345	1500	1712	1878
0.050	280	390	550	670	790	960	1090	1220	1390	1520
0.055	228	328	455	555	655	791	900	1010	1146	1255
0.060	192	275	384	467	550	667	759	842	967	1050
0.065	162	231	324	400	470	570	647	724	824	900
0.070	143	200	279	343	408	493	558	622	708	772
0.075	127	174	247	300	354	427	487	540	620	674
0.080	113	157	213	263	313	375	425	475	544	594
0.085	100	136	189	236	277	336	377	424	483	524
0.090	89	123	173	206	245	300	339	378	428	467
0.095	79	111	153	190	222	269	300	337	385	422
0.100	70	100	140	170	200	240	275	305	345	380
0.105	67	91	124	153	181	220	248	277	315	343
0.110	60	82	114	141	164	200	228	250	287	314
0.115	57	74	105	127	153	183	209	231	261	287
0.120	50	71	96	117	138	167	192	213	242	263
0.125	44	64	88	108	128	156	176	196	224	244
0.130	43	58	81	100	120	143	162	181	204	227
0.135	41	56	78	93	112	134	152	167	189	208
0.140	36	50	72	86	104	125	140	158	179	193
0.145	35	49	66	80	97	114	132	145	166	180
0.150	34	47	64	77	90	107	120	137	154	170
0.155	33	42	59	71	84	100	113	126	146	159
0.160	29	41	57	66	79	94	107	119	135	147
0.165	28	37	52	64	73	88	100	113	128	140
0.170	27	36	50	59	71	86	95	106	121	133
0.175	23	35	46	58	66	80	89	100	115	123
0.180	23	31	45	53	62	75	84	95	109	117
0.185	22	30	41	52	60	71	82	90	103	111
0.190	22	29	40	48	56	69	77	85	98	106
0.195	21	29	36	47	54	65	72	80	93	100
0.200	20	25	35	43	50	60	68	75	88	95

4. Classification Bounds for a given δ

Using Theorems 3.1 and 3.2 we can solve the following problems:

- find the least δ for given ε and m,
- find the least m for given ε and δ.

Now, we find the least ε for given δ and m.

Definition 4.1. For a given p, let k_p be the largest k satisfying

$$\sum_{i=0}^{k} \binom{m}{i} p^i (1-p)^{m-i} < \delta.$$

Lemma 4.1. The following inequalities hold

$$\mathrm{P}(\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f) > p - \frac{k_p + 1}{m}) < \delta,$$

and

$$\mathrm{P}(\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f) \geq p - \frac{k_p}{m}) < \delta.$$

Lemma 4.2. If we assume that

$$\frac{k_p}{m} < p,$$

(i.e. $\varepsilon > 0$) and

$$p' > p,$$

then we have

$$k_{p'} \geq k_p.$$

Theorem 4.1. Let m denote the size of $\mathbf{z}$, let $\delta > 0$ and $V\big(y, f_{\mathbf{z}_1}(x)\big) \in \{0, 1\}$. Assume that the least ε such that

$$\mathrm{P}(\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f) > \varepsilon) < \delta,$$

is positive. Then for each $n \in \mathbb{N}$

$$\max_{i \leq n} \frac{i}{n} - \frac{k_{\frac{i}{n}} + 1}{m} \leq \varepsilon \leq \max_{i \leq n} \frac{i+1}{n} - \frac{k_{\frac{i}{n}} + 1}{m}.$$

Theorem 4.2. Let m denote the size of $\mathbf{z}$, let $\delta > 0$ and $V\big(y, f_{\mathbf{z}_1}(x)\big) \in \{0, 1\}$. Assume that the least ε such that

$$\mathrm{P}(\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f) \geq \varepsilon) < \delta,$$

is positive. Then for each $n \in \mathbb{N}$

$$\max_{i \leq n} \frac{i}{n} - \frac{k_{\frac{i}{n}}}{m} \leq \varepsilon \leq \max_{i \leq n} \frac{i+1}{n} - \frac{k_{\frac{i}{n}}}{m}.$$

Definition 4.2. For a given p let k_p and l_p be such that

$$\delta > \sum_{i=0}^{k_p} \binom{m}{i} p^i (1-p)^{m-i} + \sum_{i=l_p}^{m} \binom{m}{i} p^i (1-p)^{m-i},$$

and

$$\max\left(p - \frac{k_p}{m}, \frac{l_p}{m} - p\right),$$

is least.

Lemma 4.3. The following inequalities are valid

$$\mathrm{P}(|\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)| > \max\left(p - \frac{k_p + 1}{m}, \frac{l_p - 1}{m} - p\right)) < \delta,$$

and

$$\mathrm{P}(|\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)| \geq \max\left(p - \frac{k_p}{m}, \frac{l_p}{m} - p\right)) < \delta.$$

When $p = \frac{1}{2}\frac{m+k}{m}$ for some $k \in \{-m, \ldots, 0, \ldots, m\}$, for every k exists l such that $p - \frac{k}{m} = \frac{l}{m} - p$. In this case is satisfied

$$p - \frac{k_p}{m} = \frac{l_p}{m} - p,$$

when $p \neq \frac{1}{2}\frac{m+k}{m}$ above equality does not hold.

As a result of such a behaviour of k_p and l_p the dependence of ε from p is discontinuous from both sides in $p = \frac{1}{2}\frac{m+k}{m}$ and the ε found for such a p is greater than the ones in its proximity.

Definition 4.3. Let $p_1 < p_2$. Let k_{p_1,p_2}, l_{p_1,p_2} be such that

$$\delta > \sum_{i=0}^{k_{p_1,p_2}} \binom{m}{i} p_1^i (1-p_1)^{m-i} + \sum_{i=l_{p_1,p_2}}^{m} \binom{m}{i} p_2^i (1-p_2)^{m-i},$$

and they are minimizing

$$\max\left(p_2 - \frac{k_{p_1,p_2}}{m}, \frac{l_{p_1,p_2}}{m} - p_1\right).$$

Lemma 4.4. Let $p_1 < p_2$, $p_1, p_2 \in [\frac{1}{2}\frac{m-k-1}{m}, \frac{1}{2}\frac{m-k}{m}]$ and $p \in (p_1, p_2)$. If $\frac{k_{p_1,p_2}}{m} < p_1$ and $\frac{l_{p_1,p_2}}{m} > p_2$ then

$$\max\left(p - \frac{k_p}{m}, \frac{l_p}{m} - p\right) \leq \max\left(p_2 - \frac{k_{p_1,p_2}}{m}, \frac{l_{p_1,p_2}}{m} - p_1\right).$$

The following theorems finalize our reflections concerning two sides bounds as they provide estimation of ε value for given probability in both sharp and not sharp case:

Theorem 4.3. Let m denote the size of $\mathbf{z}$, let $\delta > 0$ and $V\big(y, f_{\mathbf{z}_1}(x)\big) \in \{0,1\}$. Assume that the least ε such that

$$\mathrm{P}(|\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)| > \varepsilon) < \delta$$

is positive. Let x_n be finite sequence such that $0 = x_0 < x_1 < \cdots < x_n = \frac{1}{2}$, such that

$$\frac{1}{2}\frac{m+k}{m} \in \{x_n\} \text{ for every } k,$$

then for each $n \in \mathbb{N}$

$$\max_{k\in\{-m,-m+1,\ldots,0\}} \frac{1}{2}\frac{m+k}{m} - \frac{k_{\frac{1}{2}\frac{m+k}{m}}+1}{m} \le \varepsilon \le$$

$$\le \max\left(\max_{k\in\{-m,-m+1,\ldots,0\}} \frac{1}{2}\frac{m+k}{m} - \frac{k_{\frac{1}{2}\frac{m+k}{m}}+1}{m}, \max_{i<n} x_{i+1} - \frac{k_{x_i,x_{i+1}}+1}{m}, \max_{i<n} \frac{l_{x_i,x_{i+1}}-1}{m} - x_i\right).$$

Theorem 4.4. Let m denote the size of $\mathbf{z}$, let $\delta > 0$, and $V\big(y, f_{\mathbf{z}_1}(x)\big) \in \{0,1\}$. Assume that the least ε such that

$$\mathrm{P}(|\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f)| \ge \varepsilon) < \delta,$$

is positive. Let x_n be a finite sequence such that $0 = x_0 < x_1 < \cdots < x_n = \frac{1}{2}$, such that

$$\frac{1}{2}\frac{m+k}{m} \in \{x_n\} \text{ for every } k,$$

then for each $n \in \mathbb{N}$

$$\max_{k\in\{-m,-m+1,\ldots,0\}} \frac{1}{2}\frac{m+k}{m} - \frac{k_{\frac{1}{2}\frac{m+k}{m}}}{m} \le \varepsilon \le$$

$$\le \max\left(\max_{k\in\{-m,-m+1,\ldots,0\}} \frac{1}{2}\frac{m+k}{m} - \frac{k_{\frac{1}{2}\frac{m+k}{m}}}{m}, \max_{i<n} x_{i+1} - \frac{k_{x_i,x_{i+1}}}{m}, \max_{i<n} \frac{l_{x_i,x_{i+1}}}{m} - x_i\right).$$

5. Bounds for arbitrary loss function

Now we deal with the case of arbitrary loss function, i.e., $0 \le V\big(y, f(x)\big) \le 1$. We use approximation of *signum* function by Fourier series and the fact that trigonometric functions can be easily calculated for sum of independent random variables. Let θ be the Heaviside function

$$\theta(x) = \begin{cases} 0 & \text{when } x < 0 \\ 1 & \text{when } x \ge 0. \end{cases}$$

Let $X \in [0,1]$ be a random variable. We estimate $\mathrm{P}(X \ge \varepsilon) = E\theta(X - \varepsilon)$.

Consider the function

$$f_\lambda(x) = \begin{cases} \frac{-1}{\lambda}x + \frac{1}{\lambda} & \text{when } x \in [-1, -1+\lambda] \\ -1 & \text{when } x \in [-1+\lambda, -\lambda] \\ \frac{1}{\lambda}x & \text{when } x \in [-\lambda, \lambda] \\ 1 & \text{when } x \in [\lambda, 1-\lambda] \\ \frac{-1}{\lambda}x + \frac{1}{\lambda} & \text{when } x \in [1-\lambda, 1]. \end{cases}$$

f_λ can be expanded on [-1,1] into Fourier series as follows:

$$f_\lambda(x) = \frac{4}{\pi^2\lambda}\sum_{k=0}^{\infty}\frac{\sin\big(\pi(2k+1)\lambda\big)\sin\big(\pi(2k+1)x\big)}{(2k+1)^2}.$$

Now we take a finite part of the series:

$$f_{\lambda,N}(x) = \frac{4}{\pi^2\lambda}\sum_{k=0}^{N}\frac{\sin\big(\pi(2k+1)\lambda\big)\sin\big(\pi(2k+1)x\big)}{(2k+1)^2},$$

and we choose α_N and β_N such that for each $x \in [-1, 1-2\alpha_N]$

$$\theta(x) \le \frac{1}{2}f_{\lambda,N}(x+\alpha_N) + \frac{1}{2} + \beta_N.$$

Hence

$$\mathrm{P}(X \ge \varepsilon) = E\theta(X-\varepsilon) \le E\frac{1}{2}f_{\lambda,N}(X-\varepsilon+\alpha_N) + \frac{1}{2} + \beta_N + P(X > \varepsilon + 1 - 2\alpha_N).$$

$P(X > \varepsilon + 1 - 2\alpha_N)$ is usually small, so it doesn't play an important role.

$$E\frac{1}{2}f_{\lambda,N}(X-\varepsilon+\alpha_N) = \frac{2}{\pi^2\lambda}\sum_{k=0}^{N}\frac{\sin\big(\pi(2k+1)\lambda\big)E\sin\big(\pi(2k+1)(X-\varepsilon+\alpha_N)\big)}{(2k+1)^2} =$$

$$= \sum_{k=0}^{N} a^+_{\varepsilon,k}E\sin\big(\pi(2k+1)X\big) + b^+_{\varepsilon,k}E\cos\big(\pi(2k+1)X\big),$$

where

$$a^+_{\varepsilon,k} = \frac{2}{\pi^2\lambda}\frac{\sin\big(\pi(2k+1)\lambda\big)\cos\big(\pi(2k+1)(\varepsilon-\alpha_N)\big)}{(2k+1)^2},$$

$$b^+_{\varepsilon,k} = -\frac{2}{\pi^2\lambda}\frac{\sin\big(\pi(2k+1)\lambda\big)\sin\big(\pi(2k+1)(\varepsilon-\alpha_N)\big)}{(2k+1)^2}.$$

Similarly:

$$\mathrm{P}(X \ge \varepsilon) \ge E\frac{1}{2}f_{\lambda,N}(X-\varepsilon-\alpha_N) + \frac{1}{2} - \beta_N - P(X < \varepsilon - 1 + 2\alpha_N),$$

$$E\frac{1}{2}f_{\lambda,N}(X-\varepsilon-\alpha_N) = \sum_{k=0}^{N} a^-_{\varepsilon,k}E\sin\big(\pi(2k+1)X\big) + b^-_{\varepsilon,k}E\cos\big(\pi(2k+1)X\big).$$

Where

$$a^-_{\varepsilon,k} = \frac{2}{\pi^2\lambda}\frac{\sin\big(\pi(2k+1)\lambda\big)\cos\big(\pi(2k+1)(\varepsilon+\alpha_N)\big)}{(2k+1)^2},$$

$$b^-_{\varepsilon,k} = -\frac{2}{\pi^2\lambda}\frac{\sin\big(\pi(2k+1)\lambda\big)\sin\big(\pi(2k+1)(\varepsilon+\alpha_N)\big)}{(2k+1)^2}.$$

Remark 5.1. We have chosen $f_{\lambda,N}$ for approximating θ because α_N does not converge to 0 in the Fourier series that converges to θ.

Theorem 5.1. Let $X_i \in [0,1]$ be iid. (independent identically distributed random variables), with same distribution as X. Let

$$\mathrm{P}(\frac{1}{m}\sum_{i=1}^{m} X_i - EX \geq \varepsilon) = \delta_{m,\varepsilon}$$

then

$$\delta_{m,\varepsilon} \leq \sum_{k=0}^{N} a^{+}_{\varepsilon+EX,k} \Im\Big(E\exp\big(i\pi(2k+1)\frac{1}{m}X\big)\Big)^m +$$

$$+ b^{+}_{\varepsilon+EX,k} \Re\Big(E\exp\big(i\pi(2k+1)\frac{1}{m}X\big)\Big)^m + \frac{1}{2} + \beta_N + P(X > \varepsilon + EX + 1 - 2\alpha_N)$$

and

$$\delta_{m,\varepsilon} \geq \sum_{k=0}^{N} a^{-}_{\varepsilon+EX,k} \Im\Big(E\exp\big(i\pi(2k+1)\frac{1}{m}X\big)\Big)^m +$$

$$+ b^{-}_{\varepsilon+EX,k} \Re\Big(E\exp\big(i\pi(2k+1)\frac{1}{m}X\big)\Big)^m + \frac{1}{2} - \beta_N - P(X < \varepsilon + EX - 1 + 2\alpha_N).$$

Proof:
Observe that:

$$E\cos\big(i\pi(2k+1)\frac{1}{m}\sum_{i=1}^{m} X_i\big) = E\Re\exp\big(i\pi(2k+1)\frac{1}{m}\sum_{i=1}^{m} X_i\big) = \Re E\prod_{i=1}^{m}\exp\big(i\pi(2k+1)\frac{1}{m}X_i\big) =$$

$$= \Re\prod_{i=1}^{m} E\exp\big(i\pi(2k+1)\frac{1}{m}X_i\big) = \Re\Big(E\exp\big(i\pi(2k+1)\frac{1}{m}X\big)\Big)^m$$

$$E\sin\big(i\pi(2k+1)\frac{1}{m}\sum_{i=1}^{m} X_i\big) = \Im\Big(E\exp\big(i\pi(2k+1)\frac{1}{m}X\big)\Big)^m.$$

□

Theorem 5.1 solves the problem for a certain random variable. Now we deal with the case when we have a family of random variables. First we reduce a general bounded random variable to the case of random variable with a finite set of values.

Lemma 5.1. Let $X_i \in [0,1]$ be iid. with the distribution μ, let μ^+ be a probability measure on $\{\frac{0}{k}, \frac{1}{k}, \ldots, \frac{k}{k}\}$ such that

$$\mu^+(\{\frac{i}{k}\}) = \mu((\frac{i-1}{k}, \frac{i}{k}])$$

and let μ^- be a probability measure on $\{\frac{0}{k}, \frac{1}{k}, \ldots, \frac{k}{k}\}$ such that

$$\mu^-(\{\frac{i}{k}\}) = \mu([\frac{i}{k}, \frac{i+1}{k})).$$

Let $X_i^+ \in [0,1]$ be iid. with the distribution μ^+ such that $X \leq X^+$. Let $X_i^- \in [0,1]$ be iid. with the distribution μ^- such that $X \geq X^-$. Let

$$\mathrm{P}(EX - \frac{1}{m}\sum_{i=1}^{m} X_i \geq \varepsilon) = \delta_{m,\varepsilon}.$$

Then

$$P(EX^+ - \frac{1}{m}\sum_{i=1}^{m} X_i^+ \geq \varepsilon + \frac{1}{k}) < \delta_{m,\varepsilon} < P(EX^- - \frac{1}{m}\sum_{i=1}^{m} X_i^- \geq \varepsilon - \frac{1}{k}).$$

Proof:
Observe that

$$EX^+ = \sum_{i=0}^{k} \frac{i}{k}\mu((\frac{i-1}{k}, \frac{i}{k}]) = \frac{1}{k} + \sum_{i=0}^{k} \frac{i-1}{k}\mu((\frac{i-1}{k}, \frac{i}{k}]) < EX + \frac{1}{k}.$$

Hence

$$EX^+ - \frac{1}{m}\sum_{i=1}^{m} X_i^+ - \frac{1}{k} < EX - \frac{1}{m}\sum_{i=1}^{m} X_i$$

and

$$P(EX^+ - \frac{1}{m}\sum_{i=1}^{m} X_i^+ \geq \varepsilon + \frac{1}{k}) < \mathrm{P}(EX - \frac{1}{m}\sum_{i=1}^{m} X_i \geq \varepsilon).$$

The second case is analogous. □

Lemma 5.2. Let $X_i \in \{x_0, \ldots, x_n\}$, where $x_0 = 0, x_i < x_{i+1}, x_1 = 1$, be iid. with the distribution μ, let μ^+ be a probability measure on $\{x_0, \ldots, x_n\}$ such that

$$\mu^+(\{x_i\}) = \frac{k_i}{l},\ k_i \in \mathbb{N},$$

where k_i is such that

$$\mu^+(\{x_i, \ldots, x_n\}) \geq \mu(\{x_i, \ldots, x_n\}) > \mu^+(\{x_i, \ldots, x_n\}) - \frac{1}{l},$$

let μ^- be a probability measure on $\{x_0, \ldots, x_n\}$ such that

$$\mu^-(\{x_i\}) = \frac{k_i}{l},\ k_i \in \mathbb{N},$$

where k_i is such that

$$\mu^-(\{x_i, \ldots, x_n\}) \leq \mu(\{x_i, \ldots, x_n\}) < \mu^-(\{x_i, \ldots, x_n\}) + \frac{1}{l}.$$

Let $X_i^+ \in [0,1]$ be iid. with the distribution μ^+ such that $X \leq X^+$. Let $X_i^- \in [0,1]$ be iid. with the distribution μ^- such that $X \geq X^-$. Let

$$\mathrm{P}(EX - \frac{1}{m}\sum_{i=1}^{m} X_i \geq \varepsilon) = \delta_{m,\varepsilon},$$

then

$$P(EX^+ - \frac{1}{m}\sum_{i=1}^{m} X_i^+ \geq \varepsilon + \frac{1}{l}) < \delta_{m,\varepsilon} < P(EX^- - \frac{1}{m}\sum_{i=1}^{m} X_i^- \geq \varepsilon - \frac{1}{l}).$$

Proof:
Observe that

$$EX^+ = \sum_{i=0}^{n} x_i \mu^+(\{x_i\}) = \sum_{i=0}^{n} (x_i - x_{i-1})\mu^+(\{x_i, \ldots, x_n\}) <$$

$$< \sum_{i=0}^{n} (x_i - x_{i-1})(\mu(\{x_i, \ldots, x_n\}) + \frac{1}{l}) = EX + \frac{1}{l}.$$

□

Let

$$\mathcal{X}_{k,l} = \{\{X_j\}_{i=1}^{\infty} : \{X_j\}_{i=1}^{\infty} \text{ are iid. and } X_j \sim \{(\frac{i}{k}, \frac{k_i}{l})\}_{i=0}^{k}, k_i \in \mathbb{N}\}.$$

Theorem 5.2. Let m denote the size of $\mathbf{z}_2$, and let $\varepsilon > 0$. If $V\big(y, f_{\mathbf{z}_1}(x)\big) \in [0, 1]$, then the least δ such that

$$\mathrm{P}\big(\mathcal{E}(f_{\mathbf{z}_1}) - \mathcal{E}_{\mathbf{z}_2}(f_{\mathbf{z}_1}) \geq \varepsilon\big) < \delta,$$

satisfies

$$\delta > \max_{\{X_i\} \in \mathcal{X}_{k,l}} \mathrm{P}(EX_1 - \frac{1}{m}\sum_{i=1}^{m} X_i \geq \varepsilon + \frac{1}{k} + \frac{1}{l}),$$

$$\delta < \max_{\{X_i\} \in \mathcal{X}_{k,l}} \mathrm{P}(EX_1 - \frac{1}{m}\sum_{i=1}^{m} X_i \geq \varepsilon - \frac{1}{k} - \frac{1}{l}).$$

Proof:
The theorem is a straightforward consequence of the Lemmas 5.1 and 5.2. □

Theorem 5.2 combined with Theorem 5.1 provides the bound which for a given ε and an arbitrary bounded loss function determines the probability that the generalization error is lower than empirical error plus ε.

6. Remarks

Inequalities for estimating the distance of the empirical mean to the expected value of random variables are widely used in the probability theory and statistics. The advantage of the ones presented here is that they are optimal (in case of 0-1 loss function) or can be as tight as necessary for the certain applications.

This fact has is of great importance when number of available samples is limited and creating new samples is expensive. The inequalities allow us to tell how many samples is needed for a given estimation rate and reliability, and moreover assures that the smaller number of samples will not be enough without adding more assumptions on the classifier.

The estimations depends only on number of samples, shape of the loss function and measure ρ (i.e., probability distribution of samples). In case of the 0-1 valued loss functions this means that the only way of improving the inequality is to posses some a priori knowledge about possible values of $\mathcal{E}(f)$. So if we know that our classifier is good we can proof that it is even better, but without this information we cannot obtain anything more than Theorem 3.1 and Theorem 3.2 say.

There exists a possibility, that more information about $\mathcal{E}(f)$ can be achieved by replacing the Validation estimator:

$$P(\mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f) < \varepsilon) > 1 - \delta,$$

by

$$P(\mathcal{E}(f) < \varepsilon \mathcal{E}_{\mathbf{z}}(f)) > 1 - \delta,$$

or

$$P(\mathcal{E}(f) < g(\mathcal{E}_{\mathbf{z}}(f))) > 1 - \delta,$$

where g is a given function. With those new estimators we would be able to obtain tighter bound for some $\mathcal{E}(f)$ values and weaker for the others. It would reflect the conviction that the classifier is quite good.

Acknowledgements

The research has been supported by the grant 3T11C00226 from Ministry of Scientific Research and Information Technology of the Republic of Poland.

References

[1] Cucker, F., Smale, S.: On the mathematical foundations of learning, *Bull. Am. Math. Soc., New Ser.*, **39**(1), 2002, 1–49.

[2] Duda, R. O., Hart, P. E., Stork, D. G.: *Pattern classification. 2nd ed*, Chichester: Wiley-Interscience, 2001.

[3] Fukunaga, K., Hayes, R.: Effects of sample size in classifier design, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**(8), 1989, 873–885.

[4] Fukunaga, K., Hayes, R.: Estimation of classifier performance, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**(10), 1989, 1087–1101.

[5] Guyon, I., Makhoul, J., Schwartz, R., Vapnik, V.: What size test set gives good error rate estimates?, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**(1), 1998, 52–64.

[6] Hastie, T., Tibshirani, R., Friedman, J.: *The elements of statistical learning. Data mining, inference, and prediction*, Springer Series in Statistics. New York, NY: Springer, 2001.

[7] Hoeffding, W.: Probability inequalities for sums of bounded random variables, *J. Am. Stat. Assoc.*, **58**, 1963, 13–30.

[8] Michie, D., Spiegelhalter, D. J., Taylor, C. C.: *Machine Learning, Neural and Statistical Classification*, Ellis Horwood, 1994.

[9] Raudys, S., Jain, A.: Small sample size effects in statistical pattern recognition: recommendations for practitioners, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **13**(3), 1991, 252–264.

[10] Vapnik, V. N.: *Statistical learning theory*, Adaptive and Learning Systems for Signal Processing, Communications, and Control. Chichester: Wiley, 1998.