On-line Confidence Machines are well-calibrated

Volodya Vovk
Computer Learning Research Centre
Department of Computer Science
Royal Holloway, University of London
Egham, Surrey, England
vovk@cs.rhul.ac.uk

"Generalization bound < 0.5"

$\downarrow$

"Error probability < 0.5"

Yellow background means later additions/corrections

My plan:

- TCM and ICM: definitions
- TCM and ICM are well-calibrated
- Experimental results for the USPS data set: TCM/ICM vs PAC(-Bayesian) and hold-out bounds
- If time left: general discussion of TCM/ICM vs PAC/ "predicting on randomly drawn points"

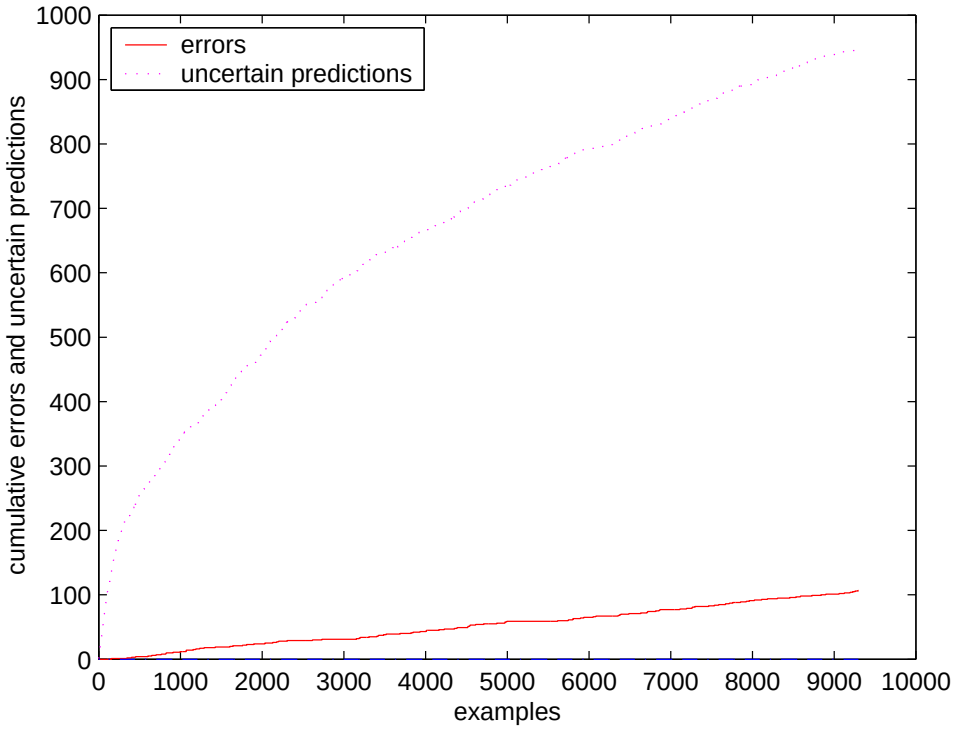
Details:

http://www.cs.rhul.ac.uk/~vovk/cm

Part I: Theory (mainly)

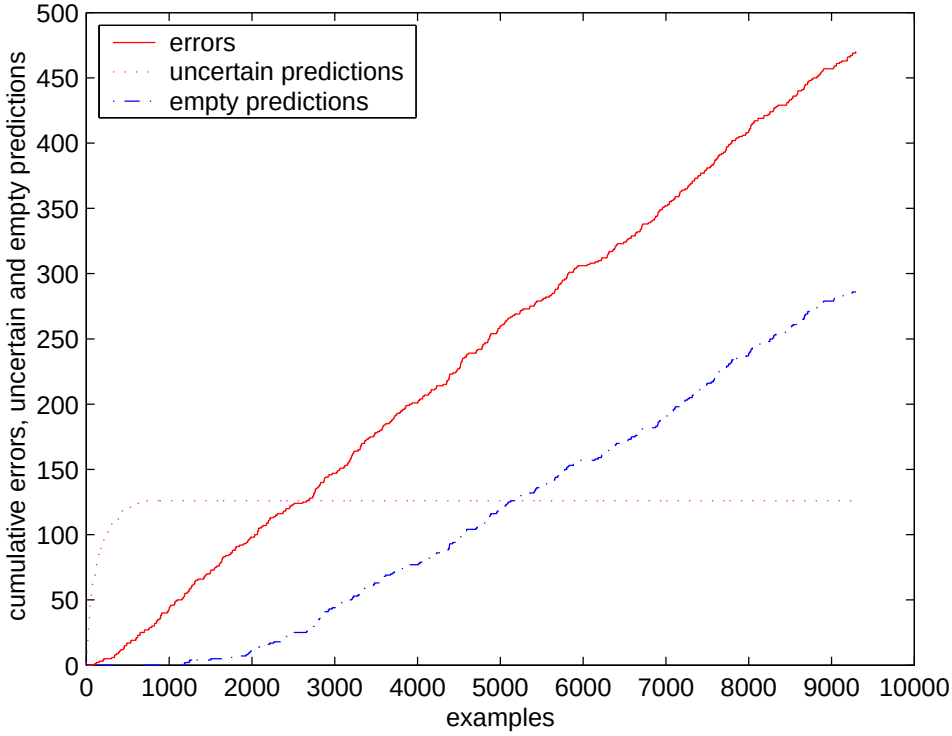
First: an illustration. USPS data set: 9298 hand-written digits (randomly permuted). TCM is given a confidence level, such as 95% or 99% (one minus maximum allowable probability of mistake). For every new hand-written digit TCM predicts a set of possible labels (0 to 9) for this digit (the predictive region).

If 99%:


errors
uncertain predictions
cumulative errors and uncertain predictions
examples
0
100
200
300
400
500
600
700
800
900
1000
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

What happens if the confidence level is more generous?

For 95%, part of the 5% allowance is "wasted" on empty predictions:


errors
uncertain predictions
empty predictions
cumulative errors, uncertain and empty predictions
examples
0
50
100
150
200
250
300
350
400
450
500
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

Examples

Nature outputs pairs

$$(z_1, z_2, \ldots) = ((x_1, y_1), (x_2, y_2), \ldots)$$

(examples); object x_i and label y_i.

$x_i \in \mathbf{X}$: the object space

$y_i \in \mathbf{Y}$: the label space

$z_i \in \mathbf{Z} := \mathbf{X} \times \mathbf{Y}$: the example space.

The data sequence is output according to an iid $P = Q^\infty$ in $\mathbf{Z}^\infty$.

Protocol

$\mathsf{Err}_0 := 0$

FOR $n = 1, 2, \ldots$:

Nature outputs $x_n \in \mathbf{X}$

Learner outputs $\Gamma_n \subseteq \mathbf{Y}$

Nature outputs $y_n \in \mathbf{Y}$

$$\mathsf{err}_n := \begin{cases} 1 & \text{if } y_n \notin \Gamma_n \\ 0 & \text{otherwise} \end{cases}$$

$\mathsf{Err}_n := \mathsf{Err}_{n-1} + \mathsf{err}_n$

END FOR

A region predictor is a family of Learner's strategies such that

$$\Gamma\left(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1-\delta_1\right)$$
$$\subseteq \Gamma\left(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1-\delta_2\right)$$

when $(1-\delta_1) \leq (1-\delta_2)$.

Randomised region predictors: depend, additionally, on an element of an auxiliary probability space.

Transductive Confidence Machine

TCM: a way to define a region predictor from a "bare predictions" algorithm.
Formally: "individual strangeness measure" $\mapsto$ region predictor.

A family of measurable

$$A_n : (z_1, \ldots, z_n) \mapsto (\alpha_1, \ldots, \alpha_n)$$

$(n = 1, 2, \ldots)$ is an individual strangeness measure if every α_i is determined by the bag $\wr z_1, \ldots, z_n \int$ and z_i.

The TCM associated with A_n is the following region predictor:

$$\Gamma\left(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1-\delta\right)$$

is the set of all $y \in \mathbf{Y}$ such that

$$\frac{\#\{i = 1, \ldots, n : \alpha_i \geq \alpha_n\}}{n} > \delta$$

where $(\alpha_1, \ldots, \alpha_n) :=$

$$A_n((x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), (x_n, y))$$

Example

The individual strangeness measure used in producing the figures: αs are defined, in the spirit of the 1-Nearest Neighbour Algorithm, as

$$\alpha_i := \frac{\min_{j \neq i : y_j = y_i} d(x_i, x_j)}{\min_{j \neq i : y_j \neq y_i} d(x_i, x_j)}$$

where d is the Euclidean distance. An object is considered strange if it is in the middle of objects labelled in a different way and is far from the objects labelled in the same way.

Other ways: SVM, inductive algorithms.

TCM result

A set $E \subseteq \{0,1\}^\infty$ is monotonic if

$$\left.\begin{array}{r}(a_1, a_2, \ldots) \in E \\ a_i \le b_i, \forall i\end{array}\right\} \Longrightarrow (b_1, b_2, \ldots) \in E$$

As far as upper bounds on $P\{\text{err} \in E\}$ for monotonic E are concerned, we can assume that the error probability of Γ is δ at every trial and errors happen independently at different trials:

Theorem TCM. For any confidence level $1-\delta$, any iid probability distribution P in $\mathbf{Z}^\infty$, and any monotonic $E \subseteq \{0,1\}^\infty$, any TCM Γ satisfies

$$P\left\{(\text{err}_n(\Gamma, 1-\delta))_{n=1}^\infty \in E\right\} \le B_\delta^\infty(E)$$

where B_δ is the Bernoulli distribution in $\{0,1\}$ with the parameter δ: $B_\delta\{1\} = \delta$ and $B_\delta\{0\} = 1-\delta$.

Corollary. Every TCM Γ is conservatively well-calibrated in the sense that, for any iid probability distribution P in $\mathbf{Z}^\infty$ and any confidence level $1-\delta$,

$$\limsup_{n\to\infty} \frac{\mathrm{Err}_n(\Gamma, 1-\delta)}{n} \le \delta \quad P\text{-a.s.}$$

Other implications. Law of the iterated logarithm:

$$\limsup_{n\to\infty} \frac{\mathrm{Err}_n(\Gamma, 1-\delta) - n\delta}{\sqrt{2\delta(1-\delta)n \ln\ln n}} \le 1 \quad \text{a.s.}$$

Hoeffding's inequality:

$$P\left\{\mathrm{Err}_N(\Gamma, 1-\delta) \ge N(\delta+\epsilon)\right\} \le e^{-2N\epsilon^2}$$

Central limit theorem:

$$\limsup_{N\to\infty} P\left\{\mathrm{Err}_N(\Gamma, 1-\delta) \ge N\delta + c\sqrt{N}\right\}$$

$$\le \frac{1}{\sqrt{2\pi}} \int_{\frac{c}{\sqrt{\delta(1-\delta)}}}^{\infty} e^{-u^2/2} du$$

Randomised TCM associated with A_n: label $y \in \mathbf{Y}$

1. is included in the predictive region if
$$\frac{\#\{i = 1, \ldots, n : \alpha_i > \alpha_n\}}{n} > \delta$$
2. not included if
$$\frac{\#\{i = 1, \ldots, n : \alpha_i \geq \alpha_n\}}{n} \leq \delta$$
3. otherwise, included with probability
$$\frac{\#\{i = 1, \ldots, n : \alpha_i \geq \alpha_n\} - n\delta}{\#\{i = 1, \ldots, n : \alpha_i = \alpha_n\}}$$

Typically: differs little from TCM.

Formalisation

Auxiliary probability space: $([0,1]^\infty, U^\infty)$, where U is the uniform probability distribution in $[0,1]$. $(\tau_1, \tau_2, \ldots) \in [0,1]^\infty$: random numbers for use at trials $1, 2, \ldots$ produced by a random number generator.

rTCM Γ depends on the extra argument $\tau \in [0,1]$ (random number); for concreteness, interpret the last item of the definition of rTCM as: y is included in $\Gamma\left(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1-\delta, \tau_n\right)$ if

$$\tau_n < \frac{\#\{i : \alpha_i \geq \alpha_n\} - n\delta}{\#\{i : \alpha_i = \alpha_n\}}$$

rTCM result

Theorem rTCM. For any rTCM Γ, any confidence level $1-\delta$, and any iid probability distribution P in $\mathbf{Z}^\infty$,

$$(\mathrm{err}_n(\Gamma, 1-\delta))_{n=1}^\infty$$

is distributed as B_δ^∞ (independent Bernoulli trials) provided τ_n are distributed as U^∞ independently of $z_1, z_2, \ldots$.

Too strong to be true? Usually: to make categorical assertions about error probabilities Bayesian-type assumptions are used. For example, in PAC theory an error probability ϵ is only asserted with some probability $1-\delta$.

Explanation: our theorem does not assert that the probability of error, $\mathrm{err}_n = 1$, is δ conditionally on knowing the whole past; it is only asserted that it is δ unconditionally and conditionally on knowing $\mathrm{err}_1, \ldots, \mathrm{err}_{n-1}$.

Actually: the probability conditional on the whole past often $\neq \delta$. Empty predictive region: conditional probability of error is 1. To balance this: the conditional probability that a non-empty predictive region is wrong tends to be $< \delta$.

Implication

Corollary Every rTCM Γ is precisely well-calibrated in the sense that

$$\lim_{n \to \infty} \frac{\mathrm{Err}_n(\Gamma, 1 - \delta)}{n} = \delta \quad \text{a.s.}$$

Idea of the proof of Theorem rTCM

- it is sufficient to show that, for any finite horizon N, $(\mathrm{err}_1, \ldots, \mathrm{err}_N)$ is distributed as B_δ^N
- it is sufficient to show that $(\mathrm{err}_N, \ldots, \mathrm{err}_1)$ is distributed as B_δ^N
- it is sufficient to show that $(\mathrm{err}_N, \ldots, \mathrm{err}_1)$ is distributed as B_δ^N conditionally on knowing $\wr z_1, \ldots, z_N \int$
- ignoring ties and borderline effects: err_N will be 1 if α_N is among the $N\delta$ largest α_i, and the probability of this is δ since all permutations are equiprobable; when z_N is disclosed, the value err_N will be settled; conditionally on knowing $\wr z_1, \ldots, z_N \int$ and z_N (and, therefore, knowing $\wr z_1, \ldots, z_{N-1} \int$), err_{N-1} will also be 1 with probability δ, and so on.

Inductive Confidence Machine

Large data sets: TCMs can be computationally inefficient. ICM: sacrifices (in typical cases) predictive accuracy for computational efficiency.

"individual strangeness measure" $\mapsto$ ("inductive algorithm", "discrepancy measure")

$\widehat{\mathbf{Y}}$: prediction space (often $\widehat{\mathbf{Y}} = \mathbf{Y}$).
Inductive algorithm:

$$D : \wr z_1, \ldots, z_n \int \mapsto (D_{\wr z_1, \ldots, z_n \int} : \mathbf{X} \to \widehat{\mathbf{Y}})$$

($D_{\wr z_1, \ldots, z_n \int}$: decision rule).

Discrepancy measure $\Delta : \mathbf{Y} \times \widehat{\mathbf{Y}} \to \mathbb{R}$.

D and Δ define an individual strangeness measure:

$$\alpha_i := \Delta\left(y_i, D_{\wr(x_1,y_1),\ldots,(x_n,y_n)\int}(x_i)\right)$$

Alternatively (less natural but faster):

$$\alpha_i := \Delta\Big(y_i,$$
$$D_{\wr(x_1,y_1),\ldots,(x_{i-1},y_{i-1}),(x_{i+1},y_{i+1}),\ldots,(x_n,y_n)\int}(x_i)\Big)$$

Inductive algorithms: "proper inductive algorithms" vs "transductive algorithms" (Vapnik).

- Proper inductive algorithms: $D_{\wr z_1,\ldots,z_n\int}$ can be "computed"; after that, computing $D_{\wr z_1,\ldots,z_n\int}(x)$ for a new x is fast.

- Transductive algorithms: little can be done before seeing x.

Fix a finite or infinite sequence $m_1 < m_2 < \cdots$ (training trials); if finite, set $m_i := \infty$ for $i >$ length. ICM based on D, Δ, and $m_1, m_2, \ldots$:

- if $n \le m_1$, $\Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1-\delta)$ is found using TCM;

- otherwise, find the k such that $m_k < n \le m_{k+1}$ and set

$$\Gamma(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, 1-\delta) := \left\{ y : \frac{\#\{j = m_k + 1, \ldots, n : \alpha_j \ge \alpha_n\}}{n - m_k} > \delta \right\}$$

where the αs are defined by

$$\alpha_j := \Delta\left(y_j, D_{\wr(x_1,y_1),\ldots,(x_{m_k},y_{m_k})\int}(x_j)\right),$$

$$j = m_k + 1, \ldots, n-1$$

$$\alpha_n := \Delta\left(y, D_{\wr(x_1,y_1),\ldots,(x_{m_k},y_{m_k})\int}(x_n)\right)$$

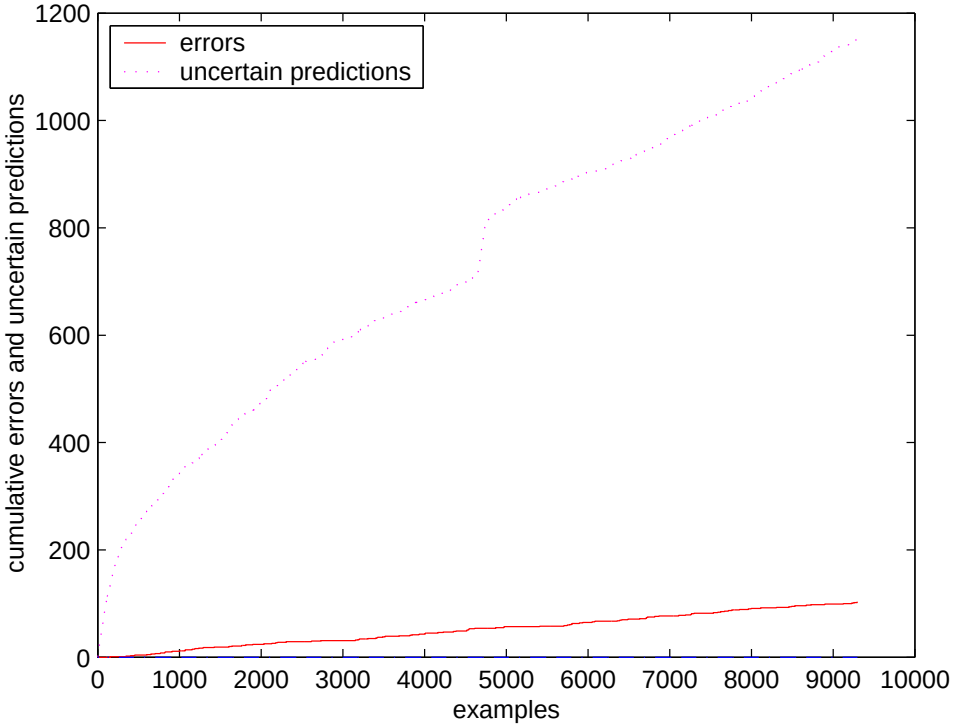
Theorem ICM. Theorems TCM and rTCM continue to hold in the case of ICMs and rICMs (defined analogously to rTCM).

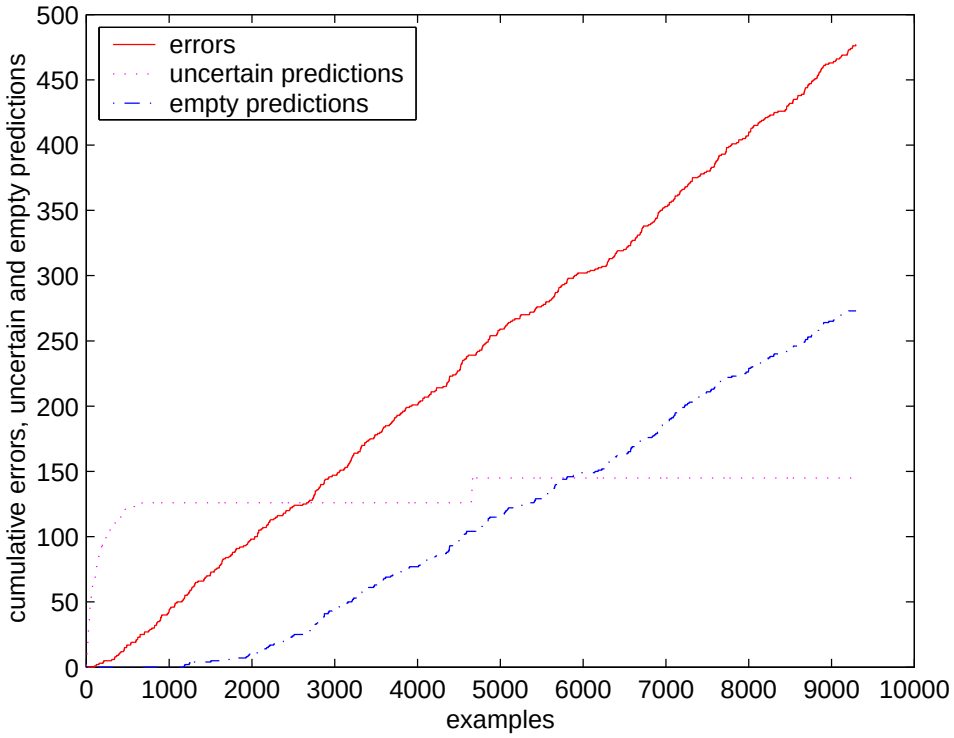
ICM requires recomputing the decision rule at $m_1, m_2, \ldots$; the rate of growth of m_i = the chosen balance between predictive accuracy and computational efficiency. The most important case: only one training trial m_1.

Let a and b be positive numbers such that either ($a \geq 1$ and $b \geq 1$) or $a > 1$.

- If A_n is computable in time $\Theta(n^a \log^b n)$, TCM: time $\Theta(n^{a+1} \log^b n)$.
- If D is computable in time $\Theta(n^a \log^b n)$, Δ is computable in constant time, and m_i is infinite and grows exponentially, ICM: same time $\Theta(n^a \log^b n)$.
- If m_i is finite, the ICM's computation time becomes $\Theta(n \log n)$ (e.g., use red-black trees for storing αs).

The performance of the Nearest Neighbour ICM on the USPS data set with training trial 4649 (the middle of the data set) at 99% and 95%:


errors
uncertain predictions
cumulative errors and uncertain predictions
examples
0
200
400
600
800
1000
1200
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000


errors
uncertain predictions
empty predictions
cumulative errors, uncertain and empty predictions
examples
0
50
100
150
200
250
300
350
400
450
500
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000

ICM's performance (= the number of uncertain predictions) deteriorates after training trials m_i. Short spells of "learning" after each training trial? (ICM is provided with fresh "training examples" and its predictions are not used or evaluated)

"Distributed prediction"

Part II: Experiments

Standard PAC bounds: too loose (a bound of 0.5 on generalization error is worth talking about!)

Littlestone and Warmuth's (1986) theorem (among the tightest): with probability $1 - \delta$ over the training examples, SVM has error probability $\leq$

$$\frac{1}{l-d}\left(d \ln \frac{el}{d} + \ln \frac{l}{\delta}\right)$$

where d is the number of support vectors.

Gives nothing interesting even for the USPS data set!

Using numbers of SVs given in Vapnik (1998), the error bound for one classifier (out of 10) is

$$\frac{1}{l-d}\left(d\ln\frac{el}{d}+\ln\frac{l}{\delta}\right)$$

$$\approx\frac{1}{7291-274}274\ln\frac{7291e}{274}\approx 0.17$$

even if we ignore $\ln\frac{l}{\delta}$ (274 is the average number of support vectors for polynomials of degree 3, which give the best predictive performance).

There are ten classifiers $\therefore$ the bound on the total probability of mistake becomes ≈ 1.7; we knew this already.

I multiply 0.17 by 10 rather than $\log_2 10$ because Vapnik's data (the number of support vectors) is about one-against-the-rest classifiers. For fancy classifiers, such as even-against-odd, the number of support vectors can be expected to be much larger: you need to have a much more complicated decision rule for such an unnatural split of the labels.

If L&W were applicable to multi-class classifiers:

$$\frac{1}{l-d}\left(d\ln\frac{el}{d}+\ln\frac{l}{\delta}\right)$$

$$\approx\frac{1}{7291-1677}1677\ln\frac{7291e}{1677}\approx 0.74$$

(1677 is the total number of support vectors for all ten classifiers for polynomial kernels).

Matthias Seeger's results (EDI-INF-RR-0094)

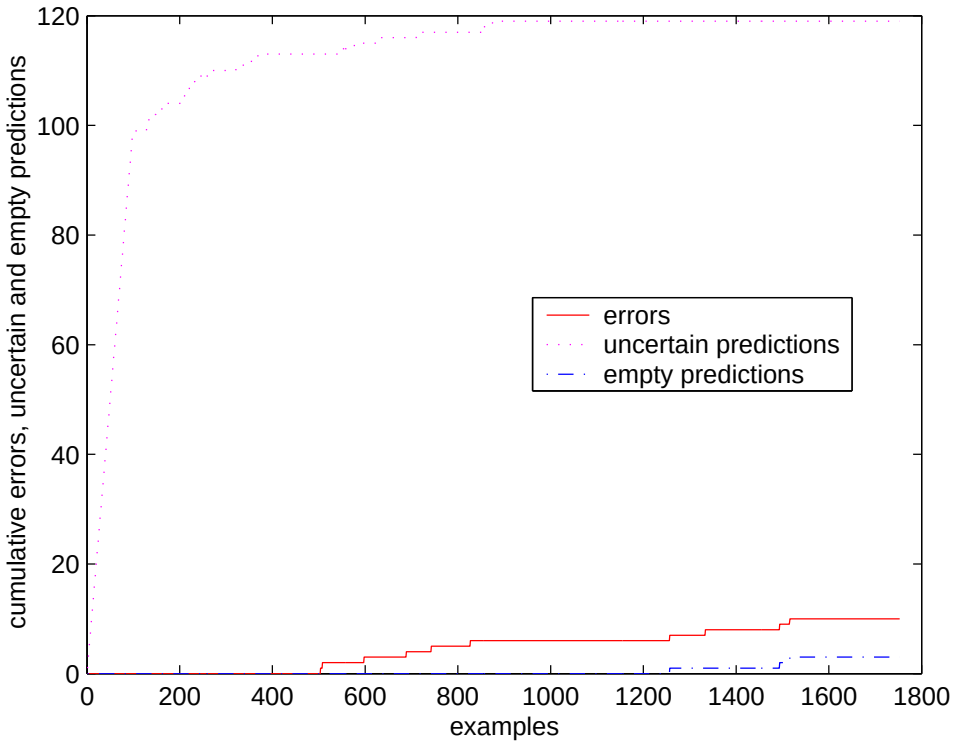
Much better, but roughly the same order of magnitude as L&W:

$$17\% \rightarrow 7.5\%$$

(MS: $\delta = 1\%$; 7K training set; 2 vs 3).

L&W's mitigating circumstances:

- L&W bound is for "one against the rest", not "one against one".
- USPS data set seems to be harder. Vapnik's 1998 data: the best error rate for NIST is 0.7% and for USPS it is 2.5% (humans). USPS is smaller, but for humans it hardly matters.


120
100
80
60
40
20
0
0
200
400
600
800
1000
1200
1400
1600
1800
examples
cumulative errors, uncertain and empty predictions
errors
uncertain predictions
empty predictions

Separating 2 and 3 is simple for USPS at confidence level 99% (there are empty predictions):

1753 examples; all errors after trial 1333 are due to empty predictions; just one wrong non-empty prediction after trial 826. No uncertain prediction after trial 874.

Hold-out estimate

Much better: assuming that the first half (4649 examples) gives a decision rule whose error rate on the second half is 4%, the generalization error bound is at most

$$4\% + z_{1\%}\sqrt{\frac{4\% \times 96\%}{4649}} \approx 4.9\%$$

with probability 99%.

Results for 1-NN TCM

USPS data set, randomly permuted, with different seeds ("initial states" in MatLab).

seed	unc pred	unc pred 2nd half	err rate
0	10.17%	5.05%	1.15%
1	9.95%	4.99%	1.14%
2	11.54%	5.36%	0.96%
3	13.96%	6.73%	0.87%
4	10.97%	6.37%	0.95%
mean	11.32%	5.70%	1.01%
σ			0.10%

Part III: Relations to other approaches

PAC learning

Batch learning: the training set

$$(x_1, y_1, \ldots, x_l, y_l)$$

and the test set

$$(x_{l+1}, y_{l+1}, \ldots, x_{l+n}, y_{l+n})$$

The standard approach in PAC theory: given the training set, upper bound the error probability $L_l :=$

$$Q\left\{(x, y) \in (\mathbf{X} \times \mathbf{Y}) : D_{\wr x_1, y_1, \ldots, x_l, y_l \int}(x) \neq y\right\}$$

where $Q^\infty = P$. We cannot guarantee that such an upper bound, say ϵ, will hold always, but quite often it is possible to show that an upper bound of ϵ will hold with probability $1 - \delta$ over the random choice of the training set from Q^l.

"PAC objection"

HLW94 studies $\mathbb{E}\, L_l$ (expectation is over the random choice of the training set), but it is generally believed that knowing $\mathbb{E}\, L_l$ would only be marginally useful. DGL1996:

> this number would indicate the quality of an average data sequence, not your data sequence

Let n be large. One pragmatic interpretation of this "PAC objection":

- If $L_l < \epsilon$, we can be sure that the percentage of errors D makes on the test set $\lessapprox \epsilon$.
- If $\mathbb{E}\, L_l < \epsilon$, the percentage of errors D makes on the test set can be considerably greater than ϵ.

PAC objection and ICM

PAC objection is also applicable to batch ICM. The training set is split into: proper training set

$$(x_1, y_1, \ldots, x_m, y_m)$$

calibration set

$$(x_{m+1}, y_{m+1}, \ldots, x_l, y_l)$$

For every test object x_i, $i = l+1, \ldots, l+n$, compute the predictive region

$$\left\{ y : \frac{\#\{j = m+1, \ldots, l, i : \alpha_j \geq \alpha_i\}}{l - m + 1} > \delta \right\}$$

where

$$\alpha_j := \Delta\left(y_j, D_{\wr(x_1,y_1),\ldots,(x_m,y_m)\int}(x_j)\right),$$

$$j = m+1, \ldots, l$$

$$\alpha_i := \Delta\left(y, D_{\wr(x_1,y_1),\ldots,(x_m,y_m)\int}(x_i)\right)$$

Semi-batch learning

To meet (our interpretation of) PAC objection: after processing each test example (x_i, y_i), $i = l+1, \ldots, l+n$, the corresponding α_i should be added to the pool of αs used in generating the predictive regions for the following test examples. Formally, redefine the predictive region as

$$\left\{ y : \frac{\#\{j = m+1, \ldots, i : \alpha_j \geq \alpha_i\}}{i-m} > \delta \right\}$$

where

$$\alpha_j := \Delta\left(y_j, D_{\wr(x_1,y_1),\ldots,(x_m,y_m)\int}(x_j)\right)$$
$$j = m+1, \ldots, i-1$$

$$\alpha_i := \Delta\left(y, D_{\wr(x_1,y_1),\ldots,(x_m,y_m)\int}(x_i)\right)$$

Theorem ICM $\therefore$ the error rate on the test set will not exceed δ, up to statistical fluctuations.

Advantage:

- In the semi-batch mode the long-term percentage of wrong predictions is guaranteed to be at most the chosen threshold δ
- In the PAC mode, there is always a non-zero probability δ that the generalization error exceeds the chosen ϵ (and so the long-term percentage of wrong predictions will exceed the chosen threshold)

When $n \ll k$: the semi-batch mode differs slightly from the batch mode of applying ICM; are the errors on the test set produced in the batch mode "nearly independent"?

Computational efficiency

- Bare predictions:

$$O\left(T_{\text{train}} + nT_{\text{appl}}\right)$$

  where T_{train} is the time required for computing $D_{\wr x_1,y_1,\ldots,x_l,y_l\int}$ and T_{appl} is the time needed to apply it to a new object

- Batch ICM:

$$O\Big(T^{\dagger}_{\text{train}} + (l-m+n)T^{\dagger}_{\text{appl}} \\ + (l-m)\log(l-m) + n\log(l-m)\Big)$$

- Semi-batch ICM:

$$O\Big(T^{\dagger}_{\text{train}} + (l-m+n)T^{\dagger}_{\text{appl}} \\ + (l-m)\log(l-m) + n\log(l-m+n)\Big)$$

Predicting on randomly chosen points

Precursor of TCM: HLW1994. Let F be a class of $\{0,1\}$-valued functions on $\mathbf{X}$. F-TCM associated with A_n outputs as predictive region the set of all y such that

- the sample $(x_1, y_1, \ldots, x_{n-1}, y_{n-1}, x_n, y)$ is consistent with F
- y would be in the predictive region output by the usual TCM

Theorem HLW Let F be a class of $\{0,1\}$-valued functions on $\mathbf{X}$ of VC dimension V; consider the variable confidence level $1 - \delta_n := 1 - V/n$. There exists an F-TCM Γ such that for any sequence $(x_1, y_1, x_2, y_2, \ldots)$ of examples,

$$|\Gamma(x_1, y_1, \ldots, x_n, 1 - V/n)| \leq 1, \quad \forall n$$

Added as compared to HLW94:

the independence of the errors at different trials allows us to use the machine of probability theory.

Two examples. Fix the variable confidence level $1 - V/n$; let P be consistent with F. No uncertain predictions;

$$\mathrm{Err}_n := \mathrm{Err}_n(\Gamma, 1 - V/n)$$

Law of the iterated logarithm: P-almost surely,

$$\limsup_{n\to\infty} \frac{\mathrm{Err}_n - V \ln n}{\sqrt{2V \ln n \ln\ln\ln n}} \le 1$$

Poisson's theorem: the distribution of $\mathrm{Err}_N - \mathrm{Err}_{\lceil N/2 \rceil}$ is asymptotically dominated, as $N \to \infty$, by the Poisson distribution with parameter $\ln 2$.

Conclusion

Main advantages of this talk's approach:

- As compared to the standard theory of PAC learning, our error bounds are practically meaningful.
- As compared to the theory of Bayesian learning, we do not assume anything beyond iid.
- The usual justification ("validity") of TCM with confidence level $1 - \delta$ is the fact that the error probability at any trial does not exceed δ. This paper adds the observation that, in the case of rTCM, the events "error at trial n", $n = 1, 2, \ldots$, not only have probability δ, but also are independent of each other. Very little can be said about a sequence of events of probability δ, but combined with independence this gives us a plethora of known properties for a random walk.