

On-line classification with confidence

Vladimir Vovk

Department of Computer Science
Royal Holloway, University of London
Egham, Surrey, England

vovk@cs.rhul.ac.uk

Workshop “Statistical Learning in
Classification and Model Selection”

Two senses of “model” (computer science vs. statistics):

- **computational model**: your “learning machine” (neural nets, SVM, etc.); the class of functions (e.g., polynomials of degree 3) you decided to fit

$$\{F_\theta : \theta \in \Theta\},$$

where

$$F_\theta : \mathbf{X} \rightarrow \mathbf{Y}$$

(objects to labels)

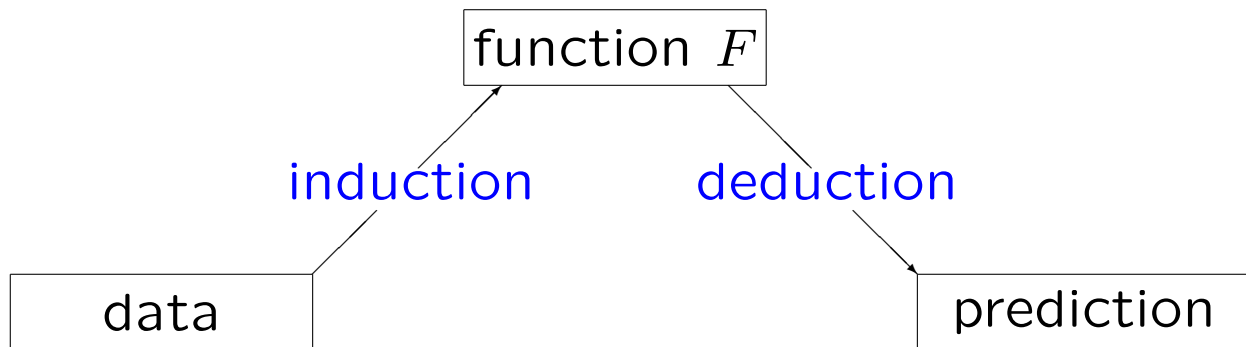
- **statistical model**: something you believe in (or provisionally accept and can test)

$$\{P_\theta(dz) : \theta \in \Theta\}$$

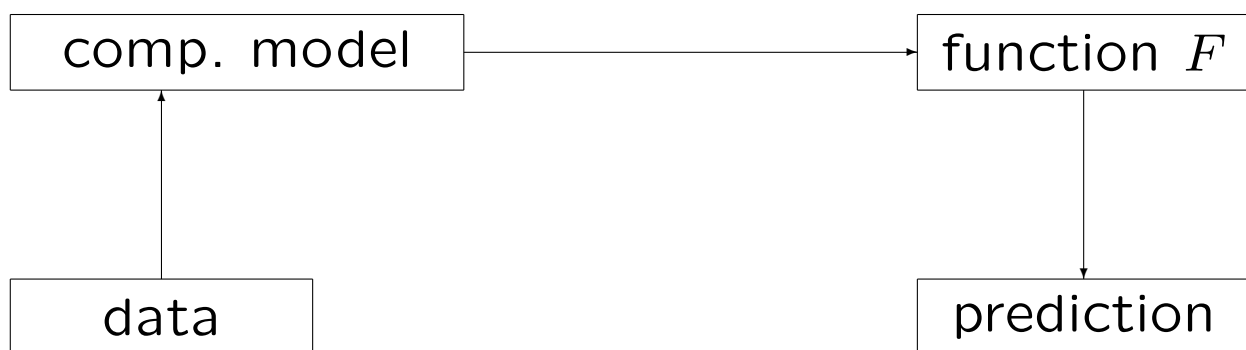
(Bayesians: plus $\mu(d\theta)$)

Early years of COLT: computational model = statistical model.

Induction and deduction according to Vapnik:



More sophisticated picture:



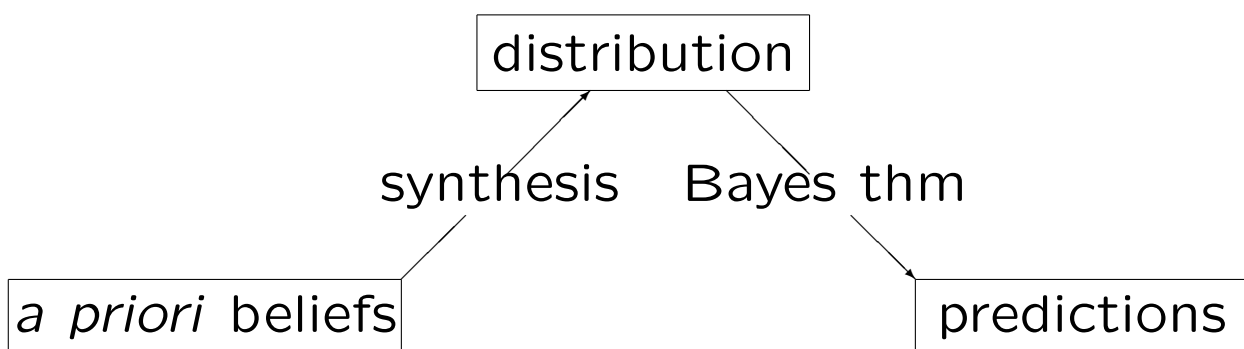
In the opposite direction, shortcut (Vapnik):



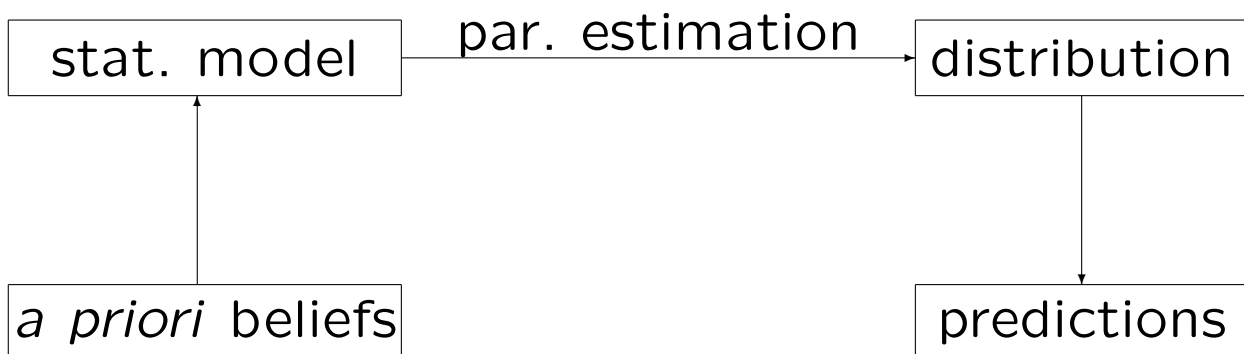
Examples: Nearest Neighbours, local Least Squares.

My main topic today: [statistical modelling](#).

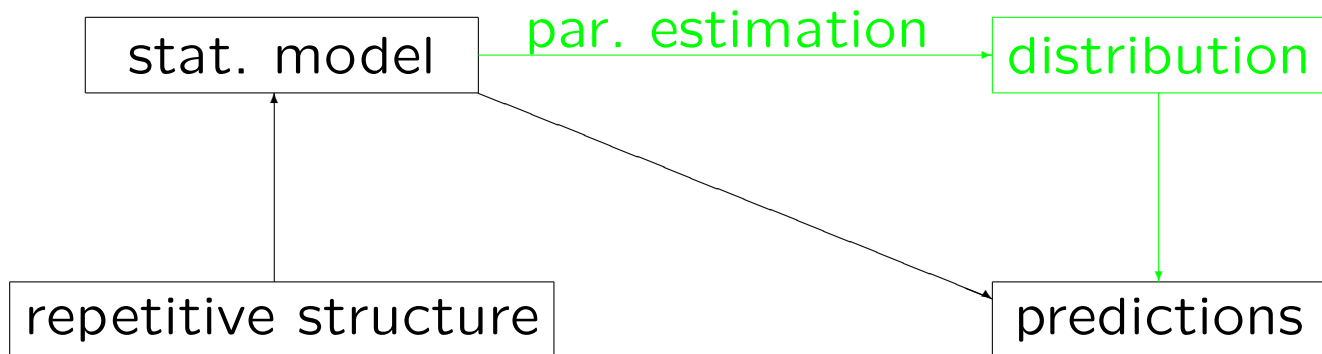
In [Bayesian statistics](#):



[Non-Bayesian statistics](#):



Another kind of statistical modelling:
repetitive structures.



Kolmogorov, Martin-Löf, Lauritzen, . . .

Core of Bayesian modelling (Bernardo and Smith)

On-line prediction protocol

$\text{Err}_0 := 0$

$\text{Unc}_0 := 0$

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

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

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

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

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

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

$\text{unc}_n := \begin{cases} 1 & \text{if } |\Gamma_n| > 1 \\ 0 & \text{otherwise} \end{cases}$

$\text{Unc}_n := \text{Unc}_{n-1} + \text{unc}_n$

END FOR

What is Reality's strategy? One possibility:
statistical model $(P_\theta(dz))^\infty$, $z \in \mathbf{Z} := \mathbf{X} \times \mathbf{Y}$.

Repetitive structure: “sufficient statistics”
 $t_n : \mathbf{Z}^n \rightarrow T$, $n = 1, 2, \dots$, and “conditional
distributions”: for each $t \in t_n(\mathbf{Z}^n)$, $P^{t,n}$ is a
probability distribution in \mathbf{Z}^n concentrated
on $t_n^{-1}(t)$ (“conditional distribution given
 $t_n = t$ ”).

The sufficient statistics must be **on-line**:
there is a measurable function U (**update
function**) such that

$$t_{n+1}(z_1, \dots, z_n, z_{n+1}) = U(t_n(z_1, \dots, z_n), z_{n+1})$$

Plus a natural condition of **consistency**.

Two examples

The **exchangeability model**:

$$t_n(z_1, \dots, z_n) := \{z_1, \dots, z_n\},$$

with the uniform distribution on all orderings.

The **Gaussian model**: \mathbf{X} is trivial (contains only one element), $\mathbf{Z} = \mathbf{Y}$ is \mathbb{R} ,

$$t_n(z_1, \dots, z_n) = (\bar{z}, \sqrt{(z_1 - \bar{z})^2 + \dots + (z_n - \bar{z})^2}),$$

$$\bar{z} := \frac{1}{n} \sum_{i=1}^n z_i,$$

and $P^{t,n}$ is the uniform distribution on the sphere.

Other models: Poisson, uniform, geometric, etc.

The usual scheme: find all probability measures P in \mathbf{Z}^∞ that agree with t_n and $P^{t,n}$ in the sense that

under P , $P^{t,n}$ are conditional distributions given $t_n = t$

The extreme points of the set of all such P will form a statistical model.

Works OK for the Gaussian model. Does not work for the exchangeability model (from the practical point of view).

Another shortcut:



We can do prediction and testing without statistical models using [Transductive Confidence Machine](#) (TCM).

P_n^t : the image of $P^{t,n}$ under the mapping $(z_1, \dots, z_n) \mapsto z_n$.

Partial measurable $F : T \times \mathbf{Z} \rightarrow \mathbb{R}$ is an **individual strangeness measure** if

$$\alpha := F(t_n(z_1, \dots, z_n), z_n)$$

is always defined. The **TCM** associated with F and a significance level $\delta > 0$ is the region predictor defined to be the set of all $y \in \mathbf{Y}$ such that

$$P_n^t \{z \in \mathbf{Z} : F(t, z) \geq F(t, (x_n, y))\} > \delta,$$

where

$$t := t_n(x_1, y_1, \dots, x_n, y).$$

The randomised version:

$$P_n^t \{z \in \mathbf{Z} : F(t, z) > F(t, (x_n, y))\} + \tau_n P_n^t \{z \in \mathbf{Z} : F(t, z) = F(t, (x_n, y))\} > \delta.$$

Theorem Suppose the examples $z_n \in \mathbf{Z}$, $n = 1, 2, \dots$, are generated from a probability distribution P that agrees with a repetitive structure $(t_n, P^{t,n})$. Any rTCM (defined as above from the repetitive structure and a significance level δ) will produce independent δ -Bernoulli errors err_n .

TCM: the statistical model is irrelevant: the inference is done in terms of t_n and P_n^t only.

Corollary Every TCM is well-calibrated in the sense that

$$\limsup_{n \rightarrow \infty} \frac{\text{Err}_n}{n} \leq \delta \quad \text{a.s.}$$

Example for the exchangeability model

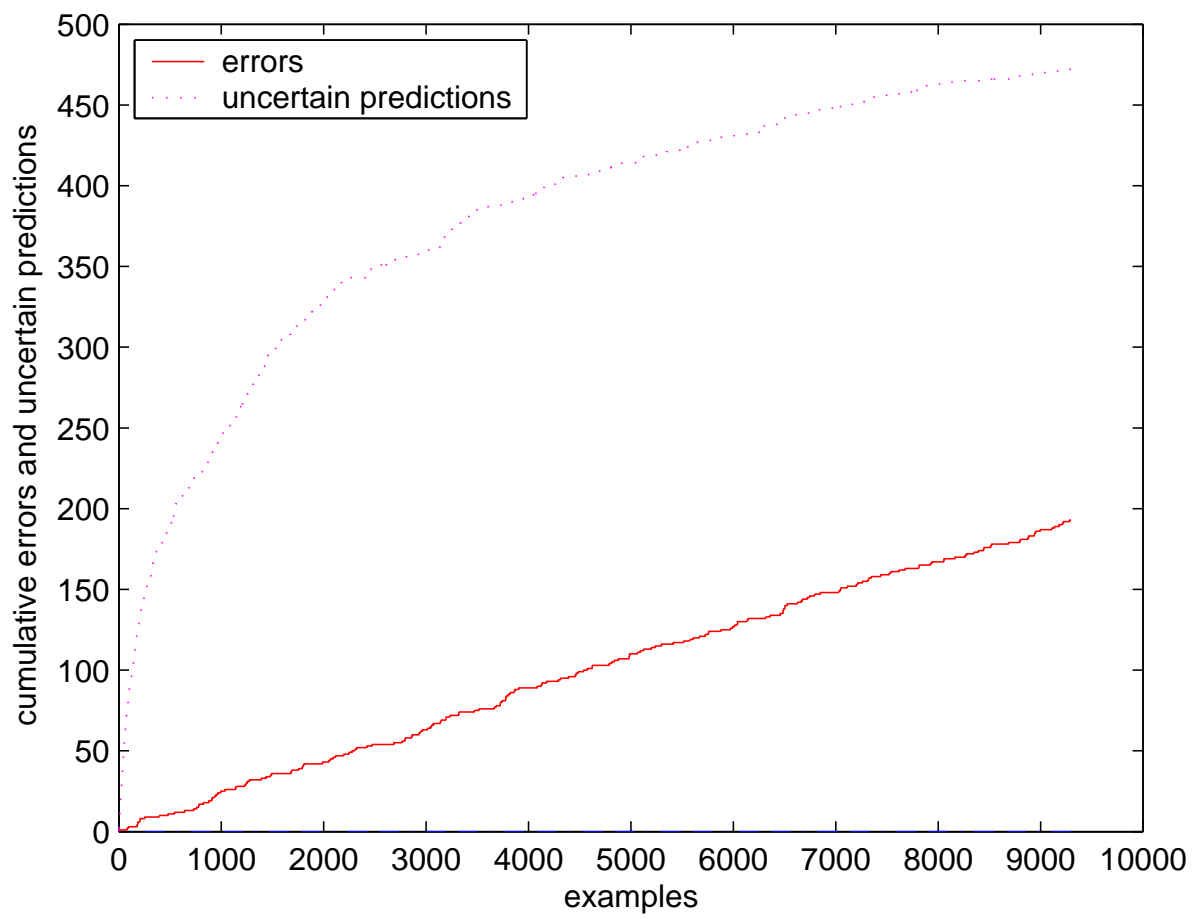
Natural individual strangeness measure: α_s are defined, in the spirit of the Nearest Neighbours Algorithm, as

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

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.

USPS data set: 9298 hand-written digits (randomly permuted). Confidence level (1 – significance level) is 98%. For every new hand-written digit TCM predicts a set of possible labels (0 to 9) for this digit (the predictive region).



Idea of the proof

- it is sufficient to show that, for any finite horizon N , $(\text{err}_1, \dots, \text{err}_N)$ is distributed as B_δ^N
- it is sufficient to show that $(\text{err}_N, \dots, \text{err}_1)$ is distributed as B_δ^N
- it is sufficient to show that $(\text{err}_N, \dots, \text{err}_1)$ is distributed as B_δ^N conditionally on knowing $t(z_1, \dots, z_N)$
- ignoring ties and borderline effects: err_N will be 1 if z_N is in a region of (conditional) probability δ ; when z_N , τ_N and $t(z_1, \dots, z_{N-1})$ are disclosed (the on-line property: info increases), the value err_N will be settled; conditionally on knowing $t(z_1, \dots, z_{N-1})$ (and z_N), err_{N-1} will also be 1 with probability δ , and so on.

It is easy to be well-calibrated.

Why should we care about TCM being well-calibrated?

Two reasons:

Practical: It gives reasonable probabilities in practice. For comparison: state-of-the-art PAC methods often give error bound > 1 (even for the “probably” parameter δ set to 1).

Theoretical: There is a TCM that makes asymptotically as few uncertain predictions as any other well-calibrated prediction algorithm.

From now on: the exchangeability model (=iid model, by de Finetti’s theorem, if \mathbf{Z} is Borel).

Practical

Littlestone and Warmuth's (1986) theorem (the tightest I know): in the binomial case, 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 (relatively clean) 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.

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).

Theoretical

TCM makes the fewest possible number of uncertain predictions.

Suppose you know the true distribution P ; \mathbf{Y} is finite.

Predictability:

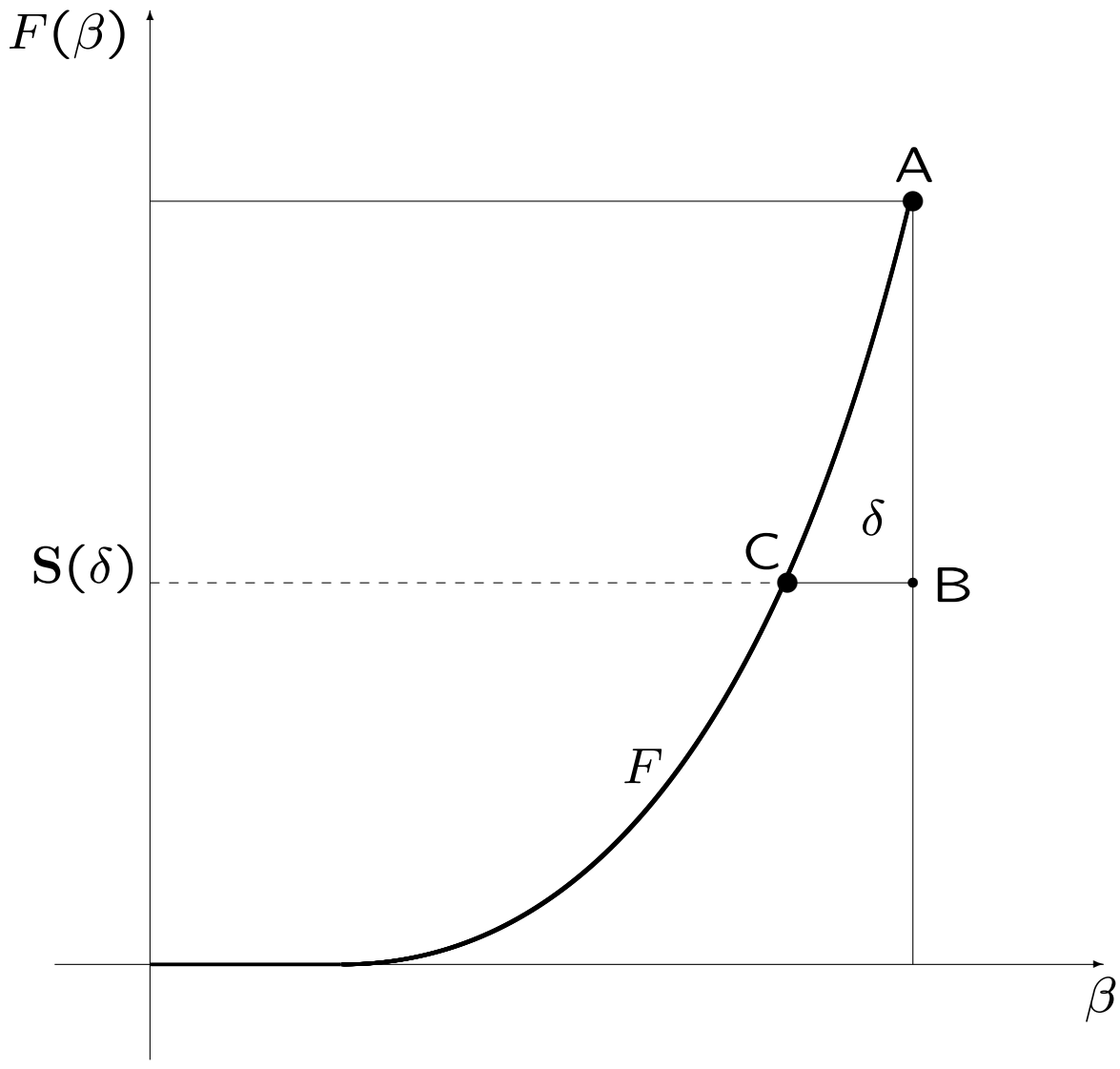
$$f(x) := \max_{y \in \mathbf{Y}} P(y | x)$$

Predictability distribution function:

$$F(\beta) := P\{x : f(x) \leq \beta\}$$

Success curve:

$$S(\delta) = \inf \left\{ B \in [0, 1] : \int_0^1 (F(\beta) - B)^+ d\beta \leq \delta \right\}$$



For the “Bayesian” region predictor:

$$\lim_{n \rightarrow \infty} \frac{\text{Unc}_n}{n} = S(\delta) \quad \text{a.s.}$$

Theorem For any well-calibrated region predictor, any significance level δ , and any probability distribution P in $\mathbf{X} \times \mathbf{Y}$,

$$\liminf_{n \rightarrow \infty} \frac{\text{Unc}_n}{n} \geq S(\delta) \quad \text{a.s.}$$

provided (x_n, y_n) are distributed as P , τ_n are distributed as U , and all are independent.

Choice function $\hat{y} : \mathbf{X} \rightarrow \mathbf{Y}$:

$$\forall x \in \mathbf{X} : f(x) = P(\hat{y}(x) | x)$$

The individual strangeness measure

$$\alpha_i := \begin{cases} 0 & \text{if } y_i = \hat{y}(x_i) \\ P(\hat{y}(x_i) | x_i) & \text{otherwise} \end{cases}$$

gives the P -TCM.

Theorem For any confidence level $1 - \delta$ and any probability distribution P in $\mathbf{X} \times \mathbf{Y}$, the P -TCM satisfies

$$\limsup_{n \rightarrow \infty} \frac{\text{Unc}_n}{n} \leq S(\delta) \quad \text{a.s.}$$

provided (x_n, y_n) are distributed as P , τ_n are distributed as U , and all are independent.

This theorem remains true if you replace P by its Nearest Neighbours approximation.

For details, see

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

Universal classification TCM

For every **extended** example (x_i, σ_i, y_i) in the data sequence,

$$P_n^\neq(y | x_i, \sigma_i) := N^\neq(x_i, \sigma_i, y) / K_n,$$

where $N^\neq(x_i, \sigma_i, y)$ is the number of $j = 1, \dots, n$ such that $y_j = y$ and (x_j, σ_j) is one of the K_n nearest neighbours of (x_i, σ_i) in the sequence $((x_1, \sigma_1), \dots, (x_{i-1}, \sigma_{i-1}), (x_{i+1}, \sigma_{i+1}), \dots, (x_n, \sigma_n))$.

The **empirical predictability function**:

$$f_n^\neq(x_i, \sigma_i) := \max_{y \in \mathbf{Y}} P_n^\neq(y | x_i, \sigma_i).$$

For each (x_i, σ_i) fix some

$$\hat{y}_n(x_i, \sigma_i) \in \arg \max_y P_n^\neq(y | x_i, \sigma_i)$$

and define the individual strangeness values

$$\alpha_i := \begin{cases} -f_n^\neq(x_i, \sigma_i) & \text{if } y_i = \hat{y}_n(x_i, \sigma_i) \\ f_n^\neq(x_i, \sigma_i) & \text{otherwise.} \end{cases}$$

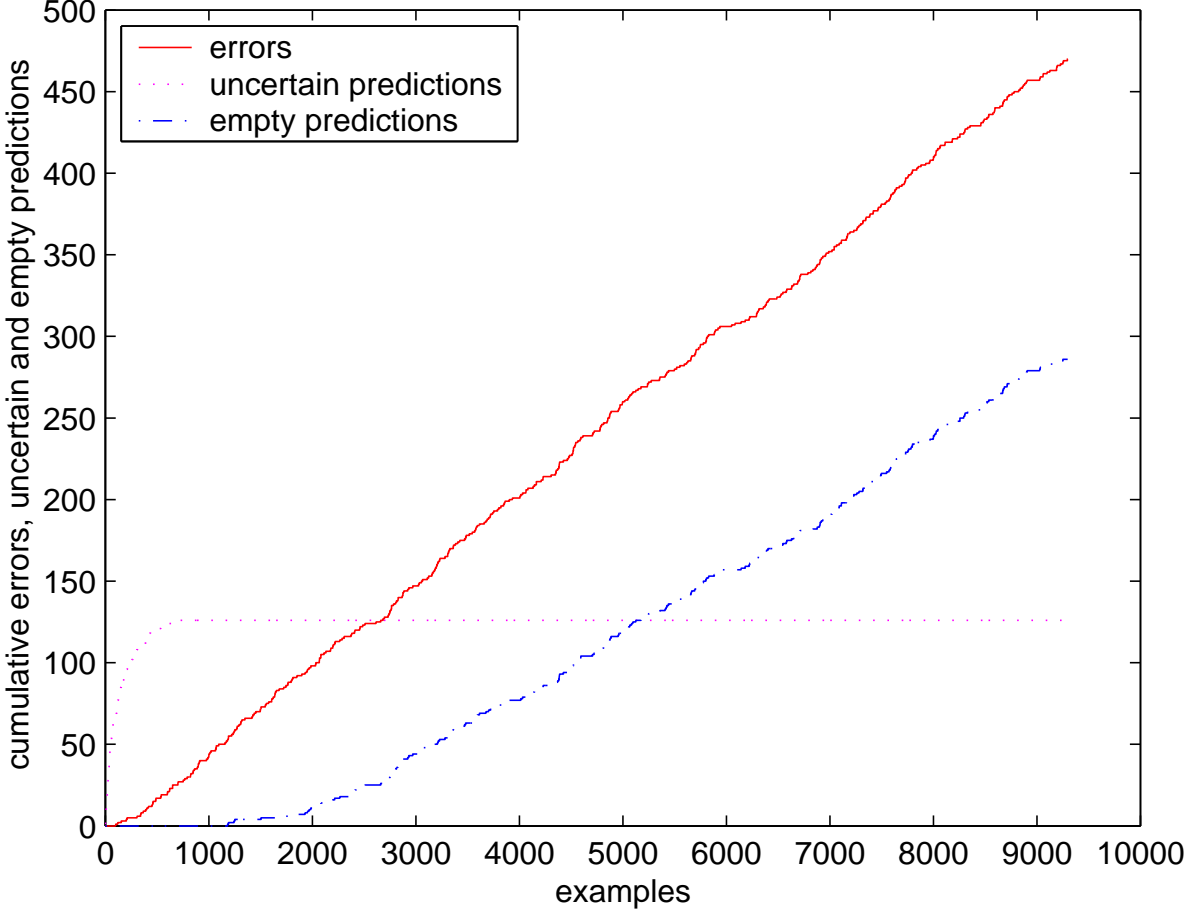
We assume: \mathbf{X} is Borel and \mathbf{Y} is finite;

$$K_n \rightarrow \infty, \quad K_n = o(n / \ln n)$$

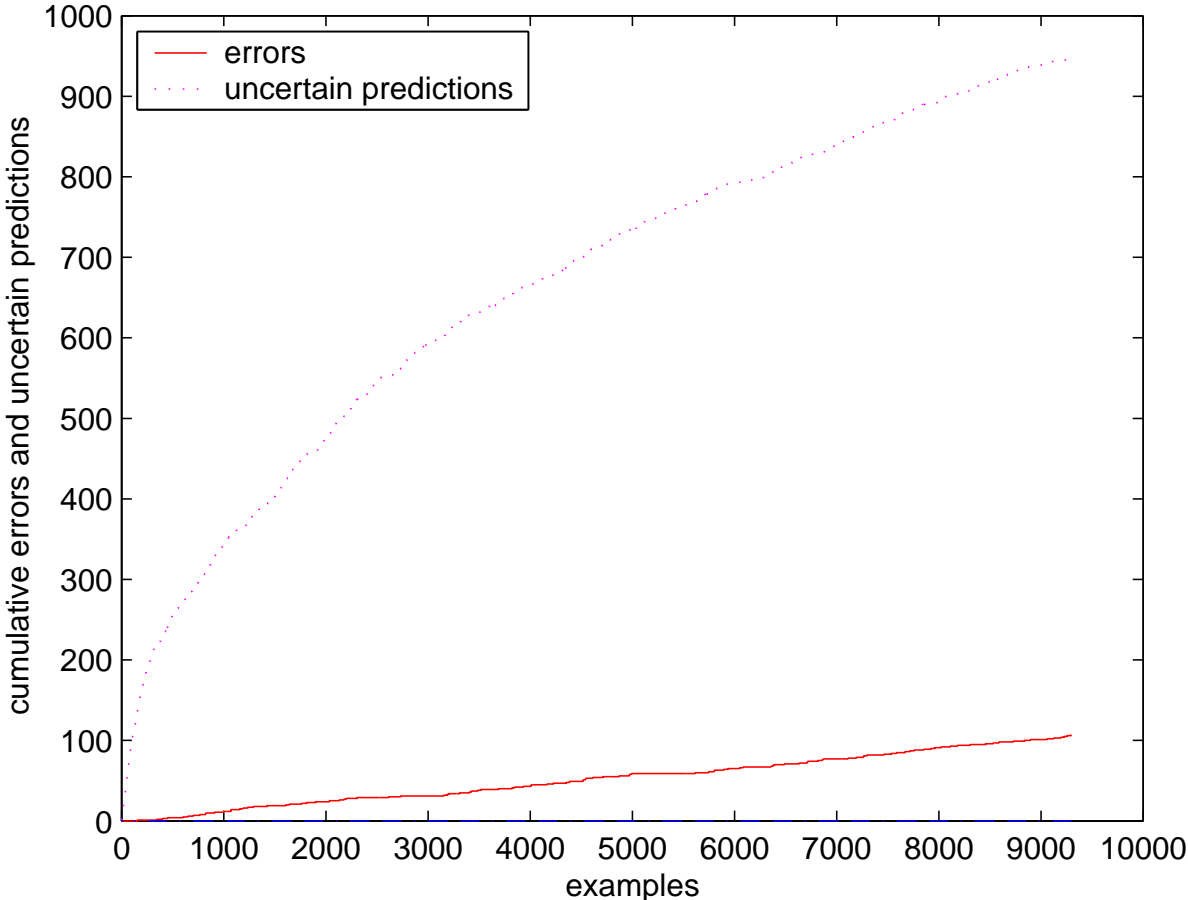
The rTCM defined by these α s is universal:
the numbers of uncertain and empty
predictions are optimal.

Proposition If $\mathbf{X} = [0, 1]$ and $K_n \rightarrow \infty$
sufficiently slowly, the Nearest Neighbours
TCM can be implemented so that
computations at step n are performed in
time $O(\log n)$.

Why the number of empty predictions is important:

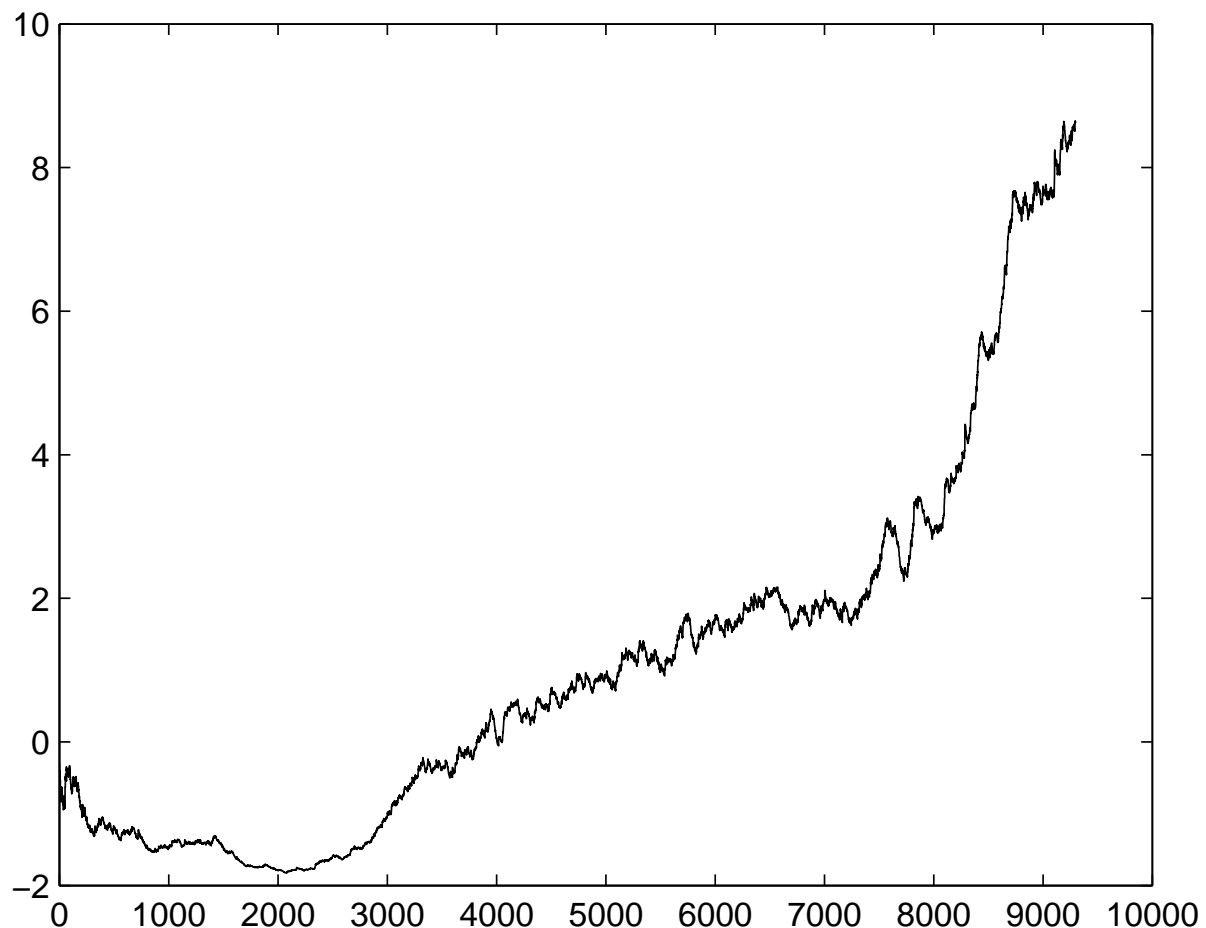


For smaller significance levels we never see empty predictions:



TCM can also be used for testing.

These are the values (on the log scale) taken by a non-negative [exchangeability martingale](#) starting from 1 and constructed from the Nearest Neighbours rTCM on the original USPS data set:



Final value: $\approx 400M$.

Idea of the construction

TCM produces p-values

$$p_n := P_n^t \{z \in \mathbf{Z} : F(t_n, z) > F(t_n, z_n)\} \\ + \tau_n P_n^t \{z \in \mathbf{Z} : F(t_n, z) = F(t_n, z_n)\}$$

Distributed independently as U . Set $S_0^{(\epsilon)} := 1$ and

$$S_n^{(\epsilon)} := S_{n-1}^{(\epsilon)} \epsilon p_n^{\epsilon-1};$$

it is an exchangeability martingale since

$$\int \epsilon p^{\epsilon-1} dp = 1$$

Finally,

$$S := \int_0^1 S^{(\epsilon)} d\epsilon$$

Conclusion

The TCM approach:

Computational models } \implies Repetitive structures
Statistical models }

Main advantages:

- New kind of guarantees, such as:
 $\text{Err}_n - \delta n$ is a random walk.
- 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.