

Kolmogorov's algorithmic statistics and Transductive Confidence Machine

Vladimir Vovk

Computer Learning Research Centre
Department of Computer Science
Royal Holloway, University of London
Egham, Surrey, UK

vovk@cs.rhul.ac.uk

Centennial Seminar on Kolmogorov
Complexity and Applications

This talk: a new way of statistical modelling, **on-line compression modelling**; closely connected with Kolmogorov's programme for applications of probability.

My **plan**:

1. Kolmogorov's programme
2. Repetitive structures
3. On-line compression modelling
4. Three specific models

The first 2 items: history; the other 2: maths.

Standard statistical modelling

The standard approach to modelling uncertainty: choose a family of probability distributions (**statistical model**) one of which is believed to be the true distribution generating, or explaining in a satisfactory way, the data.

Some applications of **probability theory**: the true distribution is assumed to be known, and so the statistical model is a one-element set.

Bayesian statistics: the statistical model is complemented by a prior distribution on the distributions in the model.

All modern applications of probability depend on this scheme.

Kolmogorov's programme

Andrei Kolmogorov: a different approach, to provide a more direct link between the theory and applications of probability (1963+).

Main idea: “practical conclusions of probability theory can be substantiated as implications of hypotheses of **limiting**, under given constraints, complexity of the phenomena under study” (K83).

практические выводы теории вероятностей могут быть обоснованы в качестве следствий гипотез о предельной при данных ограничениях сложности изучаемых явлений

My sources on the Kolmogorov programme:

K83: Andrei Kolmogorov, Combinatorial foundations of information theory and the calculus of probabilities, [Russian Mathematical Surveys](#) **38**, 29–40 (1983). Main source; written for the 1970 International Mathematical Congress (Nice).

K68: Andrei Kolmogorov, Logical basis for information theory and probability theory, [IEEE Transactions on Information Theory](#) **IT-14**, 662–664 (1968).

A: Eugene Asarin, Individual random signals: complexity approach, PhD thesis (1988).

Kolmogorov's 1963 paper in [Sankhya](#) was a precursor of his information-theoretic programme.

The main features of Kolmogorov's programme:

C (Compression): One fixes a “sufficient statistic” for the data. This is a function of the data that extracts, intuitively, all useful information from the data. This can be

- the number of ones in a binary sequence (**Bernoulli model**, [K68] and [M66]),
- the number of ones after ones, ones after zeros, zeros after ones and zeros after zeros in a binary sequence (**Markov model**, [K83]),
- the sample average and sample variance of a sequence of real numbers (the **Gaussian model**, [A]).

AC (Algorithmic Complexity): If the value of the sufficient statistic is known, the information left in the data is noise. This is formalized in terms of Kolmogorov complexity: the complexity of the data under the constraint given by the value of the sufficient statistic should be maximal.

U (Uniformity): Semantically, this requirement of algorithmic randomness means that the conditional distribution of the data given the sufficient statistic is uniform.

DI (Direct Inference): It is preferable to deduce properties of data sets directly from the assumption of limiting complexity, without a detour through standard statistical models (as in [A]; hinted at in [K83]).

Martin-Löf's development

Martin-Löf spent 1964–1965 in Moscow as Kolmogorov's PhD student. After 1965: he and Kolmogorov worked independently but arrived at similar concepts and definitions.

In 1973 Martin-Löf introduced the notion of [repetitive structure](#), later studied by Lauritzen. The theory of repetitive structures has features C and U of Kolmogorov's programme but not features AC and DI.

Extra feature of repetitive structures: their [on-line character](#). The conditional probability distributions are required to be consistent and the sufficient statistic can usually be computed recursively.

The absence of AC (algorithmic complexity and randomness) from Martin-Löf's theory: a manifestation of a general phenomenon?

The absence of DI: the goal is to derive standard statistical models from repetitive structures; to apply repetitive structures to reality one still needs to go through statistical models.

Stochastic sequences

Typically: the sufficient statistic discards a lot of information (noise); there is not so much useful information. For example:

- In the Bernoulli case, the sequence itself contains $\approx n$ bits of information, the summary contains $\approx \log n$ bits.
- In the Markov case, the sequence itself contains $\approx n$ bits of information, the summary contains $\approx 4 \log n \approx \log n$ bits.
- In the Gaussian case, n real numbers are replaced with just 2 numbers.

Therefore: if a sequence is described by a Kolmogorov complexity model, it is **stochastic** (i.e., is a random element of a simple set). Formally: a binary sequence x is **(α, β) -stochastic** if there exists a finite set $A \ni x$ such that

$$K(A) \leq \alpha, \quad K(x | A) \geq \log |A| - \beta.$$

On-line compression modelling

Direct methods of prediction for an on-line version of Kolmogorov-type models ([on-line compression models](#); akin to repetitive structures).

Our framework does not have feature AC of Kolmogorov's programme; our general theory also does not need feature U, although our specific examples have it. Feature C (in its on-line version): the key element.

The main difference of our approach from Martin-Löf's: DI.

On-line prediction protocol

$\text{Err}_0 := 0$

$\text{Unc}_0 := 0$

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

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

Forecaster 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.
- An on-line compression model (OCM) is an automaton (usually infinite) for summarizing statistical information efficiently. Lossy compression, but all useful information retained.

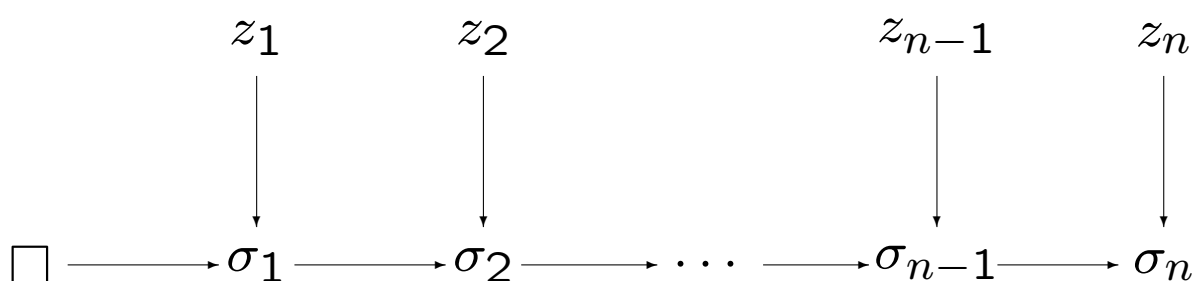
Forecaster's strategy: [region predictor](#).

An **on-line compression model** is a 5-tuple $M = (\Sigma, \square, \mathbf{Z}, (F_n), (B_n))$, where:

1. Σ is a measurable space called the **summary space**; $\square \in \Sigma$ is the **empty summary**.
2. $\mathbf{Z} (= \mathbf{X} \times \mathbf{Y})$ is a measurable **example space**.
3. $F_n : \Sigma \times \mathbf{Z} \rightarrow \Sigma$, $n = 1, 2, \dots$, are **forward functions**.
4. B_n , $n = 1, 2, \dots$, are kernels of the type $\Sigma \rightarrow \Sigma \times \mathbf{Z}$ called **backward kernels**; it is required that B_n be a reverse to F_n :

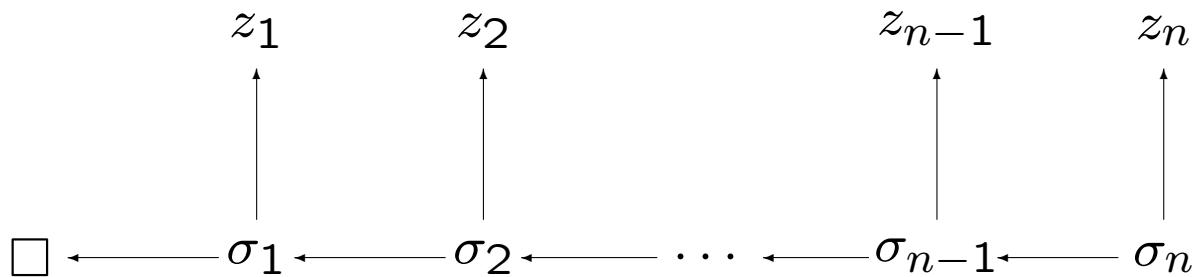
$$B_n \left(F_n^{-1}(\sigma) \mid \sigma \right) = 1, \quad \forall \sigma \in \Sigma.$$

Intuitively: an on-line compression model is a way of summarizing statistical information. At the beginning we do not have any information, $\sigma_0 := \square$. When the first example z_1 arrives, we update our summary to $\sigma_1 := F_1(\sigma_0, z_1)$, etc.; when example z_n arrives, we update the summary to $\sigma_n := F_n(\sigma_{n-1}, z_n)$.



$$t_n : (z_1, \dots, z_n) \mapsto \sigma_n$$

We can also compute the distribution $P_n(dz_1, \dots, dz_n \mid \sigma_n)$ of the data sequence z_1, \dots, z_n from σ_n :



$$\begin{aligned}
 P_n(A_1 \times \dots \times A_n \mid \sigma_n) &:= \\
 &\int \dots \int B_1(A_1 \mid \sigma_1) B_2(d\sigma_1, A_2 \mid \sigma_2) \dots \\
 &B_{n-1}(d\sigma_{n-2}, A_{n-1} \mid \sigma_{n-1}) B_n(d\sigma_{n-1}, A_n \mid \sigma_n)
 \end{aligned}$$

Region prediction in OCM

Any sequence of measurable functions $A_n : \Sigma \times \mathbf{Z} \rightarrow \mathbb{R}$, $n = 1, 2, \dots$, is called a **local strangeness measure** w.r. to the OCM $M = (\Sigma, \square, \mathbf{Z}, (F_n), (B_n))$. The **TCM** associated with (A_n) and a significance level $\delta > 0$ is the region predictor defined to be the set of all $y \in \mathbf{Y}$ such that

$$p_n(y) > \delta,$$

where

$$p_n(y) := B_n\left(\left\{(\sigma, z) \in \Sigma \times \mathbf{Z} : A_n(\sigma, z) \geq A_n(\sigma_{n-1}, (x_n, y))\right\} \mid \sigma_n\right)$$

and

$$\begin{aligned}\sigma_n &:= t_n(z_1, \dots, z_{n-1}, (x_n, y)), \\ \sigma_{n-1} &:= t_{n-1}(z_1, \dots, z_{n-1}).\end{aligned}$$

The randomised version:

$$p_n(y) := B_n\left(\left\{(\sigma, z) \in \Sigma \times \mathbf{Z} : \right. \right. \\ \left. \left. A_n(\sigma, z) > A_n(\sigma_{n-1}, (x_n, y))\right\} \mid \sigma_n\right) \\ + \theta_n B_n\left(\left\{(\sigma, z) \in \Sigma \times \mathbf{Z} : \right. \right. \\ \left. \left. A_n(\sigma, z) = A_n(\sigma_{n-1}, (x_n, y))\right\} \mid \sigma_n\right).$$

In practice: difference negligible.

We say that a probability distribution P in \mathbf{Z}^∞ agrees with the on-line compression model $(\Sigma, \square, \mathbf{Z}, (F_n), (B_n))$ if, for each n , $B_n(A | \sigma)$ is a version of the conditional probability, w.r. to P , that $(t_{n-1}(z_1, \dots, z_{n-1}), z_n) \in A$ given $t_n(z_1, \dots, z_n) = \sigma$ and given the values of z_{n+1}, z_{n+2}, \dots .

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

Corollary Every (r)TCM is well-calibrated in the sense that

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

Exchangeability model

For defining specific OCM, we will specify their statistics t_n and conditional distributions P_n ; these will uniquely determine F_n and B_n .

The exchangeability model has statistics

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

given the value of the statistic, all orderings have the same probability $1/n!$.

Example of a local strangeness measure

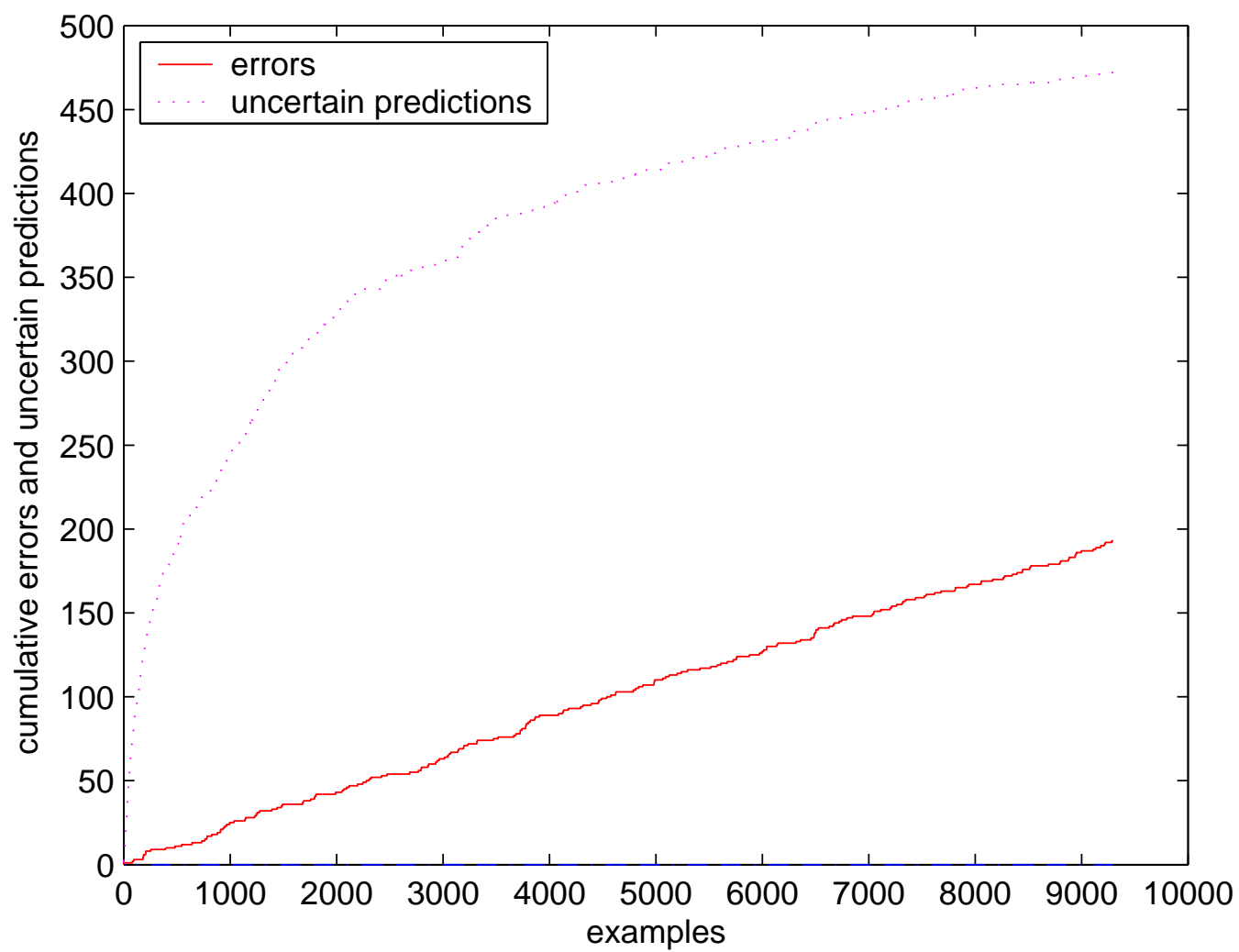
In the spirit of the Nearest Neighbour Algorithm:

$$A_n (\{z_1, \dots, z_{n-1}\}, z_n) := \frac{\min_{i \in \{1, \dots, n-1\}: y_i = y_n} d(x_i, x_n)}{\min_{i \in \{1, \dots, n-1\}: y_i \neq y_n} d(x_i, x_n)}$$

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, Ridge Regression,

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).



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 **universal** TCM that makes asymptotically as few uncertain predictions as any other well-calibrated prediction algorithm.

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.

PAC-Bayesian results I tried: even worse.

For the “Bayesian” region predictor:

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

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

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

provided z_n are distributed as P , θ_n are distributed as U , and all are independent.

Universal classification TCM

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

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

Theorem For any confidence level $1 - \delta$ and any probability distribution P in \mathbf{Z} , the K_n -Nearest Neighbours rTCM satisfies

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

provided z_n are distributed as P , θ_n are distributed as U , and all are independent.

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

Gaussian model

In the Gaussian model, $\mathbf{Z} := \mathbb{R}$,

$$t_n(z_1, \dots, z_n) := (\bar{z}_n, R_n),$$

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

$$R_n := \sqrt{(z_1 - \bar{z}_n)^2 + \dots + (z_n - \bar{z}_n)^2}$$

and $P_n(dz_1, \dots, dz_n | \sigma)$ is the uniform distribution in $t_n^{-1}(\sigma)$.

Local strangeness measure:

$$A_n(t_{n-1}, z_n) = A_n((\bar{z}_{n-1}, R_{n-1}), z_n) := z_n - \bar{z}_{n-1}.$$

If t_δ is defined by $\mathbb{P}\{|t_{n-2}| > t_\delta\} = \delta$ (where t_{n-2} has Student's t -distribution with $n - 2$ d.f.), the predictive interval is

$$\left\{ z : |z - z_n| \leq t_\delta \sqrt{\frac{n}{(n-1)(n-2)} R_{n-1}} \right\}.$$

The usual predictive regions based on the t -test; new property: the errors are independent.

Markov model

Markov model: goes beyond exchangeability. The example space \mathbf{Z} is finite (often $\mathbf{Z} = \{0, 1\}$).

The **Markov summary** σ_n of $z_1 \dots z_n$ is this digraph:

- the set of vertices is the state space \mathbf{Z} ;
- the vertex z_1 is marked as the **source** and the vertex z_n is marked as the **sink**;
- the arcs of the digraph are the transitions $z_i z_{i+1}$, $i = 1, \dots, n - 1$; the arc $z_i z_{i+1}$ has z_i as its tail and z_{i+1} as its head.

All vertices v satisfy $\text{in}(v) = \text{out}(v)$ with the possible exception of the source and sink (unless they coincide), for which we then have $\text{out}(\text{source}) = \text{in}(\text{source}) + 1$ and $\text{in}(\text{sink}) = \text{out}(\text{sink}) + 1$. We call a digraph with this property a **Markov graph** if the arcs with the same tail and head are indistinguishable.

Markov model: P_n is the uniform distribution in the set of Eulerian trails from the source to the sink in σ_n .

The local strangeness measure:

$$A_n(\sigma_{n-1}, z_n) := -B_n(\{(\sigma_{n-1}, z_n)\} | F_n(\sigma_{n-1}, z_n)).$$

An efficient (explicit in the binary case) representation of the corresponding region predictor is obtained from the BEST theorem (de Bruijn, van Aardenne-Ehrenfest, Smith and Tutte) and the Matrix-Tree theorem.

Lemma (BEST). In any Markov graph $\sigma = (V, E)$ the number of Eulerian trails from the source to the sink equals

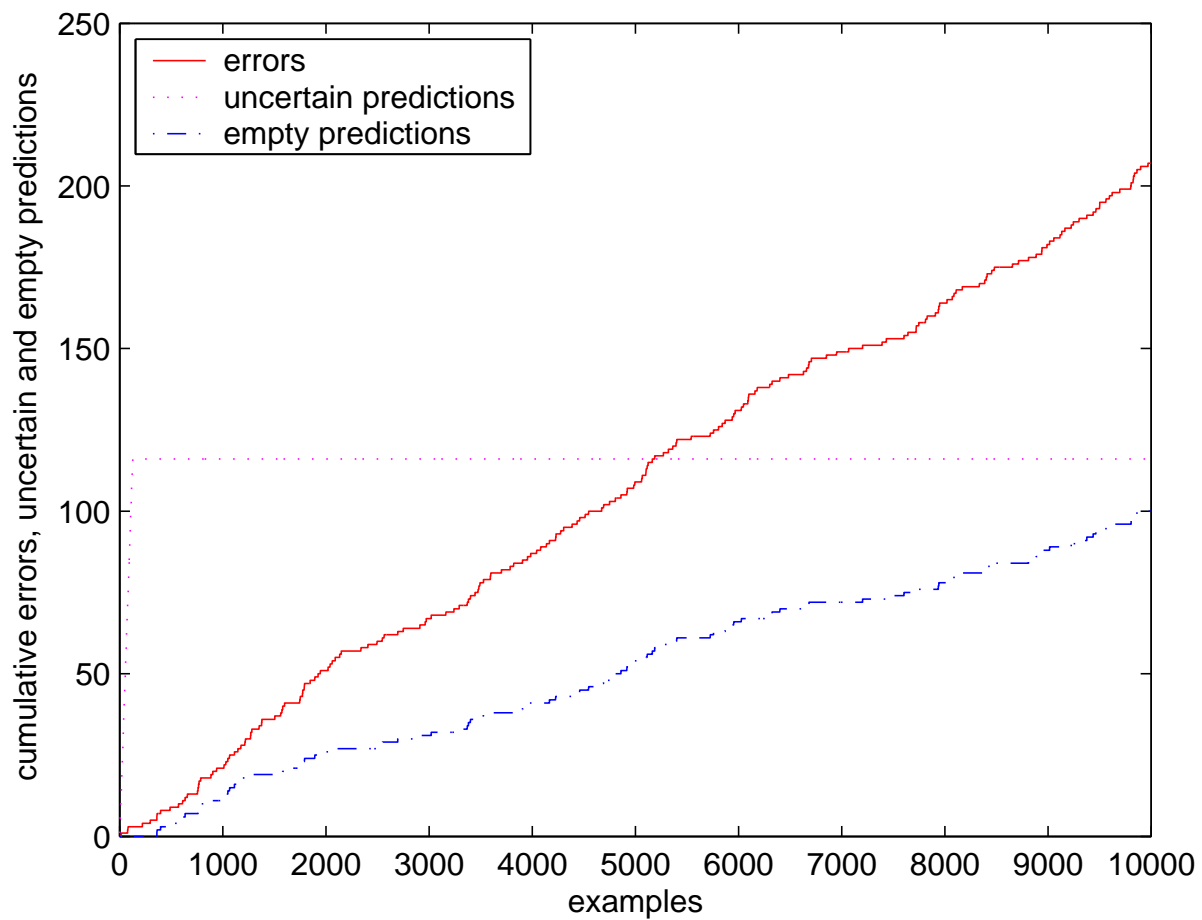
$$T(\sigma) \frac{\text{out}(\text{sink}) \prod_{v \neq \text{sink}} (\text{out}(v) - 1)!}{\prod_{u, v \in V} n_{u, v}},$$

where $T(\sigma)$ is the number of spanning out-trees in the underlying digraph centred at the source and $n_{u, v}$ is the number of arcs from u to v .

Lemma (MTT). To find the number $T(\sigma)$ of spanning out-trees rooted at the source in the underlying digraph of a Markov graph σ with vertices z_1, \dots, z_n (z_1 being the source),

- create the $n \times n$ matrix with the elements $a_{i, j} = -n_{z_i, z_j}$;
- change the diagonal elements so that each column sums to 0;
- compute the co-factor of $a_{1, 1}$.

TCM predicting the binary Markov chain
with transition probabilities
 $\mathbb{P}(1 | 0) = \mathbb{P}(0 | 1) = 1\%$ at significance level
2%:



Other models: Poisson, uniform, geometric, etc. (see Bernardo and Smith (2000), Chapter 4). Especially interesting: Bayes nets.

The usual scheme: find all probability measures P in \mathcal{Z}^∞ that agree with OCM. The extreme points of the set of all such P will form a statistical model.

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

The basic on-line scenario (the true label is disclosed immediately after the prediction is made) is not realistic: typically there is a delay before the true label becomes known, and some (or most) labels are never disclosed. Daniil Ryabko and Ilia Nouretdinov: calibration continues to hold even if only a small fraction of labels is eventually disclosed; huge delays allowed.

Conclusion

The main idea:

Statistical models \rightarrow OCM

Some advantages:

- New kind of guarantees, such as:
 $\text{Err}_n - \delta n$ is a random walk.
- Practically meaningful error bounds under the exchangeability assumption.
- Universality results.

Details: <http://vovk.net/kp>.