

THE NEW APPROACHES TO PREDICTIVE MODELING WITH A VERY GREAT NUMBER OF VARIABLES

Michel Bera

*Member, French Institute of Actuaries
Co-Founder and Chief Scientific Officer, Kxen Inc*

The work of the Russian mathematician Vladimir Vapnik (AT&T Labs) enables us to go back to the roots of theoretical statistics, leaving behind Fisher's parameters in favour of the general approaches started in the 1930s by Glivenko-Cantelli-Kolmogorov. Nowadays, it has become possible to model millions of events described by thousands of variables, within a reasonable time for a specific application. This opens up great prospects in numerous fields, including insurance, in which there is such an enormous amount of data.

Within the general scope of the development of mathematics and physics, substantial clarification has been provided by the theory of probability, statistics and, more recently, data analysis and learning theory. All these disciplines deserve to be given a common framework for reflection, based on common formalization. At present there are only a few analyses that are individually satisfactory, but these do not uniformly clarify all these disciplines.

Vladimir Vapnik's works date back to the 1970s, when he published two major works ([1][2]), which provide a methodological framework for the "Statistical Learning Theory". It is this new framework for reflection that we would like to briefly focus on here.

These theoretical works have already afforded companies that have adopted them a competitive advantage, contributing to more reliable and faster information than ever before by any other classic methodology. There are many applicable fields within life insurance, Properties and Casualties. We will endeavour to describe some of them below.

Epistemology of the "simple model" concept: from Aristotle to Ockham's razor

If we apply Kant's approach, all scientific theories must consist of three elements:

- The description of a problem;
- The Solution;
- The proof.

Long before him, Aristotle [3][4] stated that it was better to represent nature in the simplest and most concise manner, using a model of equal complexity. William of Ockham (1285 – 1349AD), with his famous "razor" established the following principle: if two theories explain the facts equally well, then the simplest theory is to be preferred.

In the 1930s, two broad approaches led to two very different developments of data modelling, that of Glivenko-Cantelli and that of Fisher.

Glivenko and Cantelli analysed the uniform convergence of an empirical function of distribution F_{emp} with a sample of L random independent variables identically distributed at values within \mathfrak{R} , (X_1, \dots, X_L) based on a single variable X :

$$F_{\text{emp}}(x) = 1/L \{\text{Card}\{X_i \mid X_i < x\}\}$$

Tending towards the distribution function

$$F(x) = P(X < x).$$

They establish convergence towards 0 with a probability of $\sup_x |F(x) - F_{\text{emp}}(x)|$, for a sample of size L . Finishing off this work, Kolmogorov and Smirnov established a limit rule for this statistic that remains famous today.

About the same time, Fisher opted for a more specific approach, based on a parametric representation of the laws of probability: he proposed the basis of current modern theories of density analysis, discriminant analysis and regression analysis. He actually separated theoretical statistics, a part of statistical science that studies general inference problems, and applied statistics, which implement particular parametric models.

The quality of Fisher's approach and its specific results help to establish great confidence in applied statistics, moving away from theoretical statistics.

In the 1960s, with the arrival of the first large data files, with numerous and highly correlated variables, it became clear that the "traditional applied" methods would not be sufficient to create acceptable modelling for sets of data with a large number of variables: the "curse of dimensionality" was discovered. Those that created efficient solutions that were not proven at the time (e.g.: data analysis, PCA, first rings of neural networks) were relegated to a mere amateurism disputed by the traditional statistics community.

It took another 25 years and the first tangible results of neural networks (1990) to prove that the "curse dimensionality" could be warded off. The general scope of the learning theory raised by Vapnik in 1995, created a new system for resolution, by questioning the description of the problem of predictive modelling. Unlike the solution available up to then, this resolution was based on a perfectly proven statistical theory.

To take a specific example, in this new scientific framework, it is no longer the case that it does not make sense to create modelling of a scoring on 3000 descriptive variables, based on a sample of 100 observations, as in the current example of modern genetics. In the world of the Internet, systematic recording in real time of behaviour patterns on a site often allows us to obtain bulk information related to access by tens of millions of people. Each of their sessions contains thousands of variables. Previously, this information was unable to be processed. This need has reinforced the new generation of models, which in turn look for robustness, in other words the behavioural stability of the model on a new set of data in the same universe and speed of implementation, as we must react in just a few seconds for credit authorization or to grant an insurance policy where tens of thousands of variables could be involved.

Main problem of learning (building a model from existing data)

We have a set of data, described by rows (events) each comprising n parameters and a final column (the “business question”). Thus we can imagine each row in the form of $[x_1, \dots, x_n | y]$ where y is called “business question”.

Let X be a vector of \mathcal{R}^n : $X = (x_1, \dots, x_n)$. We will now try to create a model of \mathcal{R}^n in \mathcal{R} (regression) or \mathcal{R}^n in $[0,1]$ (classification). In order to do this, we use a model to calculate function $f(X,w)$, the result of which is y , where

- w is a parameter of \mathcal{R}^p that defines the model,
- $Z_i = (X_i, y)$ are the possible data values
- $Q(z,w)$ is the error rate of the model when $f(X,w)$ equals y
- $P(z)$ is the unknown probability of the Z data.

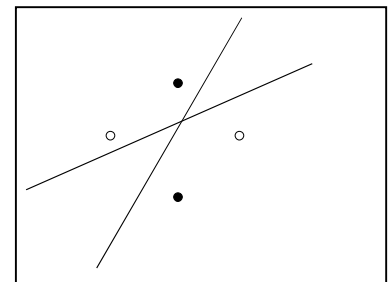
The aim is thus to minimize the risk probability of w : $R(w) = \int Q(z,w) dP(z)$. In order to do this, we only have L learning cases (z_1, \dots, z_L) , which we consider are based on the unknown law $P(z)$. We therefore try to minimize the Empirical Risk

$$E(w) = (1/L) \sum \{ Q(z_i, w) \mid i=1, \dots, L \}.$$

The strength of the Vapnik theory leads to a R risk increased by the sum of the empirical risk, measured on the whole learning and of a deterministic quantity.

A model is said to be consistent if the error in such model on the new data converges towards the error of a model based on learning data, as the size of the whole learning increases.

Let $f \in \mathcal{F}$ be the function that describes the model: $Y = f(X,w)$. Vapnik associates a whole number h to the family of functions \mathcal{F} of \mathcal{R}^n in \mathcal{R} , called Vapnik-Chervonenkis dimension of the family \mathcal{F} . For family \mathcal{F} , this number characterizes its capacity to separate (“complexity”), to “slice” the points of the space \mathcal{R}^n : a family \mathcal{F} of functions of \mathcal{R}^n in \mathcal{R} “slices” a set of points (x_1, \dots, x_L) of \mathcal{R}^n if whatever the colour of the L points divided in m white points and $L-m$ black points (there are 2^L possible), there is a particular function f of \mathcal{F} that has positive values on the “white” ones and negative on the “black” ones. The family \mathcal{F} of functions of \mathcal{R}^n in \mathcal{R} thus has the VC dimension h if: there is a set of h points of \mathcal{R}^n which could be “sliced”, no group of $h+1$ vectors can be “sliced”. We show for example that if \mathcal{F} is the group of straight lines in the plane, then $h=3$.



A straight line may not always separate 4 points so if \mathcal{F} is the set of straight lines in the plane then $h_{\mathcal{F}}=3$

The major Vapnik theorem is thus the following:

- learning of the model (X,w) is consistent if, and only if, the family of models has a finite h dimension;
- with the probability $1-q$,

$$R(w) < E(w) + \sqrt{[h (\ln(2L/h) + 1) - \ln(q)] / L}.$$

The equation above is fundamental:

- The risk of the model “applied to nature” is increased with the probability $1-q$ (risk threshold, i.e. $q=1\%$ or 0.01) by the sum of the empirical risk, measure on the whole learning, and a deterministic quantity.
- It does not involve the number of variables of the problem: this theorem allows the intellectual approach to statistical modelling to be reformulated.
- It does not involve the unknown probability law $P(z)$, for which no hypothesis is proposed.
- The term to the right of $E(w)$ tends towards zero when h/L tends towards 0.

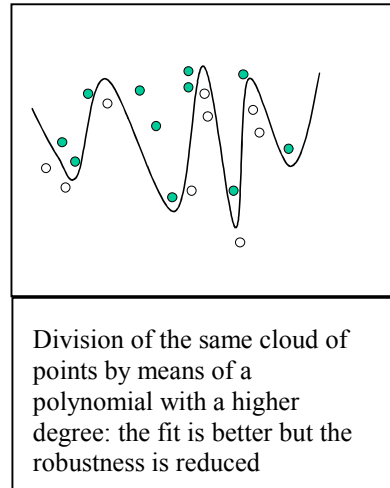
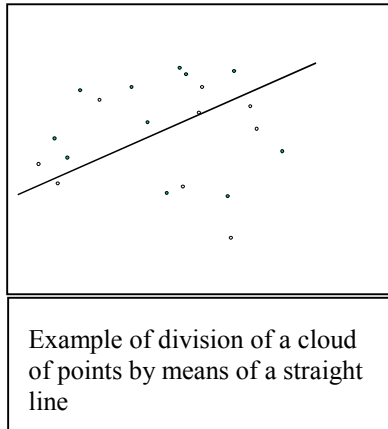
Even when the limit is practically too high, this equation shows that the error rate of a general model can be controlled when “applied to nature”, even for a great amount of parameters, providing that the $f(X,w)$ model is chosen from a family of models \mathfrak{S} where the VC dimension h remains low compared to L . Besides, even when the model involves millions of variables, if the h/L ratio remains low ($1/20$ is a good practical value), the model is useful and robust, as it will provide test results comparable to those observed on the data available to build up the model (learning set).

Principle of SRM (*Structured Risk Minimization*)

The idea of SRM is to create an approach in which the model $f(X,w)$ will be chosen among a given \mathfrak{S}_m family of models by assessing the accuracy of the model (the fit, $E(w)$) and the robustness of the model (characterized by the inverse of the term $\sqrt{[h(\ln(2L/h) + 1) - \ln(q)] / L}$). In order to do this, we build a succession of possible model families, more and more “enriched”: $\mathfrak{S}_1 \subset \mathfrak{S}_2 \subset \dots \subset \mathfrak{S}_p$ with $h_1 < h_2 < \dots < h_p$. The families of models being the most “enriched”, the best model for the family \mathfrak{S}_q will be more precise than the best model for the family \mathfrak{S}_p if p is lower than q . However, as $h_p < h_q$, this model will be less robust (consistent).

A concept of families of “Russian dolls” of functions \mathfrak{S}_m can be carried out in many manners:

- architecture of neural networks;
- polynomial degree;
- weight control in a neural network;
- smoothing level in data filtering, etc.



In a SRM approach, modelling (which is present at each stage of each of the K components) means replacing the traditional process consisting of:

1. making a hypothesis on the (unknown) statistical distribution of data;
2. accepting that a large number of problem dimensions implies either a large number of parameters and excessive calculation times, or choosing a priori instrumental variables with their relevant problems of consistency;
3. finding a better fit and testing null hypotheses.

By the process consisting of:

1. Studying and finding out the optimal \mathfrak{S} family from a SRM point of view, controlling its Vapnik-Chervonenkis h dimension;
2. Bearing in mind all the parameters, as, by definition, we control the consistency of the model;
3. Looking for the best balance between robustness and consistency.

Specific examples

An example of a \mathfrak{S} family that is easy to work with is a polynomial with n variables: $Y = P(X_1, \dots, X_n)$. Vapnik has designed a whole theory by directly writing the optimal consistency equations for models of the type: $Y = \langle w, X \rangle + b$, where w is the parameter sought to determine the model and $\langle w, X \rangle$ is the scalar product of the parameter vector w and vector X .

This leads us to Lagrange equations and to a quadratic optimisation model under linear constraints: this is the Support Vector Machines theory, which is beyond the scope of this document, but that allows large scale problems to be dealt with in an exact theoretical scope.

Another approach comes in again to control the h dimension of the polynomial by means of approaches based on the theory of poorly approached problems, in other words those problems with a highly unreliable numerical solution when there is noise data (that is typical in human data). By artificially adding noise in data, we can control the h and therefore the robustness of the model: an excessive amount of noise results in a useless

model but on which would be very robust when dealing with new data; no noise at all results in a highly precise learning model, but that would be very unstable when dealing with new data. The SRM approach provides the most balanced solution.

This second method allows huge data sets to be dealt with (millions of rows), up to an average of 3000 parameters. Thus, we can model a set of bulk data (encoding character strings on the fly) extremely fast: 250 descriptive variables (character strings, numerical variables and ordinal values) and one million rows (subjects) in only 100 minutes using a normal PC, compared with what used to be three weeks of work with traditional statistical methods (when everything had to be checked column by column).

To give just one example, the SRM approach (K2C engines for encoding variables, K2R engine for robust regression) has allowed Kxen to build a customer scoring model for insuring caravans based on bulk data in 85 seconds and given social professional variables. This model came in fifth place.¹

Applications to insurance

The models stemming from the Vapnik theory allow optimal use of historical data for insurance companies or professional institutions. Here we show two examples related to car insurance.

The SRM method can be advantageously replaced by the reliability methods used in rating a mono-vehicle automobile. In fact, a historical analysis of contracts in a portfolio allows a great deal of information to be gathered (x_i), such as the brand and type of vehicle, the date of the contract, license plate, colour, address of the insurance holder. Adding this bulk information of claim expenses attached to the contract (y_i) allows a consistent estimate to be obtained as regards the pure premium to be applied to each pair (driver, vehicle) and particularly isolated risk factors. In this case, the capacity of a model alone to reduce the scale of the problem is the determining factor.

When applied to the claim database, SRM can provide a reliable estimate for opening a claim and thus minimize the differences in the required provisions for pending claim. The impact that this could have on companies is far from being negligible, as the impact on the provision for pending claim is subject to heavy taxation. In this case, the information (x_i) is composed of different elements that are known when opening the claim, so that they are recorded in the company databases. (y_i) could represent the total cost of the claim and the time it takes to deal with, depending on the information sought. With the aid of KXEN, a company may thus develop the cost of a claim in the most reliable and optimal manner and avoid resorting to average cost methods, the results of which are very often unsatisfactory. In fact, as the distribution of the claims has a "high consequence/risk ratio", exceptional claims could distort the calculations and the simple definition of peaks sometimes provides rather inconclusive results. In this case, the capacity of the model to identify only good "instrumental" variables is an asset.

¹ This world-wide competition was organized by P. van der Putten and M. van Someren (EDS) . CoIL Challenge 2000: The Insurance Company Case. Published by Sentient Machine Research, Amsterdam. Leiden Institute of Advanced Computer Science Technical Report 2000-09. June 22, 2000 (Available at URL : [http://www.liacs.nl/~putten/library/cc2000/.](http://www.liacs.nl/~putten/library/cc2000/))

In the future, the possibility of having customer databases available will allow us to obtain more precise information regarding the insured parties and to be sure of a more accurate rating system based on millions of criteria. The company that is the first to develop this know-how will have a real advantage over the competition.

Conclusion

The Vladimir Vapnik approach, by using the concept of predictive modelling, allows to ward off the curse of 25 years of problems involving a great deal of data.

The set of theorems based on the VC dimension concept for a function family, and SRM strategy (Structured Risk Minimization) provides a general framework where most of the old methods have been reused, understood and validated, ridge regression methods with time series, analysis of correspondence to trade-off and panel problems.

This means a constructive challenge for statistics, as already, in the CRM database problems, and in particular those of finance and insurance, it has become possible to approach new behaviour assessment methods, based on files that nowadays contain millions of variables and even hundreds of millions of rows in the most usual cases.

References:

- [1] V.Vapnik - The Nature of Statistical Learning Theory, Springer-Verlag, 1999 (2nd edition)
 - [2] V.Vapnik – Statistical Learning Theory – Wiley, 1998
 - [3] Aristotle, Book I, chap. vi
 - [4] Aristotle, Book VIII, chap vi
- For further detailed information, please refer to:
- [5] Nello Cristianini – John Shawe Taylor; Support Vector Machines and other kernel-based learning methods, Cambridge University Press, 2000
 - [6] Alexander Smola, Peter Bartlett et al., Advances in Large Classifiers, MIT Press, 2000
 - [7] Bernard Schölkopf, Christopher Burges et al., Advances in Kernel Methods, MIT Press, 1999

Acknowledgements:

I would like to thank Sylvain Coriat, Generali France Assurances, who agreed to re-read this paper and make his very valuable contribution.