

# Nouvelles méthodologies de la modélisation: Théorie de Vapnik et mise en oeuvre par KXEN

Michel Bera

KXEN, 25 Quai Gallieni  
92158 Suresnes cedex, France  
Michel.Bera@kxen.com,  
<http://www.kxen.com>

**Résumé.** Nous présentons ici quelques éléments de la théorie de Vladimir Vapnik. Nous montrons comment les concepts de précision et de robustesse permettent de définir un « bon » modèle. Le théorème de Vapnik, que nous présentons, indique les contraintes pour qu'un modèle soit « bon » : la classe de fonctions où l'on recherche ce modèle doit avoir une VC dimension finie et cette classe de fonctions est choisie selon le principe de SRM (minimisation structurelle du risque). Nous décrivons ensuite comment KXEN a implémenté ces éléments théoriques pour réaliser un moteur de modélisation efficace et robuste

## 1 Introduction

En 1995, le mathématicien Vladimir Vapnik donnait une conférence remarquée au sein des laboratoires Bell Labs, présentant une nouvelle approche scientifique de la modélisation basée sur le concept qui devait le rendre célèbre : la VC Dimension (V comme Vapnik, C comme Chernovenkis, autre mathématicien russe). Nous avons déjà exposé, très brièvement, les principes de la théorie de Vapnik dans un article paru dans la revue *Risques* (Bera, 2001).

Le présent exposé est composé de deux parties : une première partie rappelle les principes majeurs de la théorie de Vladimir Vapnik ; la deuxième partie est dévolue à la mise en oeuvre de ces idées par la société KXEN, selon l'esprit du mathématicien Léon Bottou et que reflète l'un de ses propos : *« Il y a un grand nombre de problèmes de modélisation, particulièrement ardu à résoudre. Cependant, il y a un plus grand nombre de problèmes qui ne sont certes pas glorieux, mais dont on sait d'ores et déjà trouver la solution, d'une manière efficace »*.

A ce stade, il convient de définir ce que nous entendons par « modèle » ou « modélisation ». Comme données, on dispose d'un ensemble de  $L$  observations d'un phénomène, décrit par des variables  $X^1, \dots, X^n$  (on parle de données structurées). A chacune des observations, on peut adosser la réponse à une question posée, « la question métier »  $Y$ . Ainsi, le modèle est une fonction permettant de prédire  $Y$ , à partir de la connaissance des variables  $X^1, \dots, X^n$ . Le type de fonction est donné de façon explicite (modèle paramétrique) ou implicite (non paramétrique) par le modélisateur. Quant à la fonction exacte, elle est déterminée ou estimée grace