LES NOUVEAUX ÉNONCÉS DE LA MODÉLISATION PRÉDICTIVE À TRÈS GRAND NOMBRE DE VARIABLES

Michel Béra

Membre, Institut des actuaires français
Co-Founder et Chief Scientific Officer, Kxen Inc

Les travaux du mathématicien russe Vladimir Vapnik (AT&T Labs) permettent de reprendre à la base la notion de statistique théorique, abandonnant les paramètres d'un Fisher en faveur des approches générales inaugurées dans les années 30 par Glivenko-Cantelli-Kolmogorov. Il devient aujourd'hui possible de modéliser des dizaines de millions d'événements décrits par des milliers de variables, dans un temps acceptable pour une application concrète. Cela ouvre des perspectives importantes à de nombreux domaines, y compris ceux de l'assurance, où les données sont si nombreuses.

Dans le cadre général de l'évolution des mathématiques et de la physique, les éclairages considérables ont été apportés par la théorie des probabilités, la statistique, et plus récemment par l'analyse des données et la théorie de l'apprentissage. Toutes ces disciplines méritaient un cadre commun de réflexion, basé sur une formalisation commune. A ce jour seuls coexistent une succession de rapports individuellement satisfaisants, mais qui n'éclairent pas uniformément toutes ces disciplines.

Les travaux de Vladimir Vapnik, qui remontent aux années 70, apportent avec la publication de deux ouvrages majeurs ([1][2]), le cadre méthodologique de la " théorie de l'apprentissage statistique ". C'est ce cadre de réflexion nouveau que nous voulons présenter ici succinctement aujourd'hui.

Ces travaux théoriques ont d'ores et déjà apportés un avantage compétitif aux entreprises qui les ont adoptés en apportant une information plus fiable et plus rapide que toutes méthodologies classiques. Leurs domaines d'application dans l'assurance vie et IARD sont nombreux. Nous tâcherons d'en ébaucher certains.

Epistémologie de la notion de « modèle simple » : d'Aristote au rasoir d'Ockham

Si l'on se rapporte à l'approche de Kant, toute théorie scientifique doit comprendre trois éléments :

– l'énoncé du problème ;
– la résolution du problème ;

– les preuves.

Bien avant lui, Aristote [3][4] exposait qu'il était préférable de représenter la nature de la manière la plus simple et concise, à qualité de modèle égale. Guillaume d'Ockham (1285 – 1349AD), avec son fameux " rasoir " institutionnalisait ce principe : si deux théories expliquent des faits avec la même qualité, alors la théorie la plus simple doit être retenue.

Dans les années 30, deux grandes approches conduiront à deux développements fort différents de la modélisation des données, celle de Glivenko-Cantelli et celle de Fisher.

Glivenko et Cantelli analysent la convergence uniforme d'une fonction empirique de distribution F_{emp} d'un échantillon de L variables aléatoires indépendantes identiquement distribuées à valeurs dans $\Re$, $(X_1, .. ,X_L)$ tirées d'une même variable X :

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

vers la fonction de distribution

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

Ils établissent la convergence vers 0 en probabilité de $\sup_x |F(x) - F_{emp}(x)|$, pour un échantillon de taille L. Pour parachever ce travail Kolmogorov et Smirnov établiront une loi limite restée célèbre de cette statistique.

Fisher choisit vers la même époque une approche plus concrète, basée sur une représentation paramétrique des lois de probabilités : il pose alors les bases des théories modernes actuelles de l'analyse de la densité, de l'analyse discriminante et de l'analyse de la régression. Il sépare enfin la statistique théorique, partie de la science statistique qui étudie les problèmes généraux d'inférence, et les statistiques appliquées qui mettent en œuvre des modèles paramétriques particuliers.

La qualité de l'approche et des résultats concrets de Fisher contribua à établir une confiance forte dans les statistiques appliquées, au détriment de la statistique théorique.
Dans les années 60, avec l'arrivée des premiers fichiers de données importants, aux variables multiples et très corrélées, il devint évident que les méthodes "appliquées classiques " ne suffiraient pas à construire une modélisation acceptable pour les ensembles de données à grand nombre de variables : une sorte de " malédiction des problèmes de grande dimension" apparaissait. Ceux qui mettaient en évidence des solutions efficaces mais non démontrées à l'époque (ex : analyse des données, analyse des correspondances, premières vagues de réseaux de neurones) se sont vus rejetés dans un artisanat contesté par la communauté statistique traditionnelle.

Il faudra attendre 25 ans et les premiers résultats tangibles de réseaux de neurones (1990), pour montrer que la " malédiction des problèmes de grande dimension " pouvait être conjurée. Le cadre général de la théorie de l'apprentissage posée par Vapnik en 1995, permet, en remettant en question l'énoncé du problème de la modélisation prédictive, un nouveau mode de résolution. Contrairement aux solutions alors disponibles, cette résolution est assise sur une théorie statistique parfaitement démontrée.

Pour prendre un exemple concret, dans ce nouveau cadre scientifique, il n'est plus dépourvu de sens de modéliser un scoring sur 3000 variables descriptives, à partir d'un échantillon de 100 observations, exemple courant en génétique moderne. Dans le monde de l'Internet, un enregistrement systématique et en temps réel de comportements de prospects sur un site permet souvent d'obtenir des informations brutes relatives à des sessions de dizaines de millions d'individus. Chacune de ses sessions contenant des milliers

de variables, cette information était auparavant inexploitable. Ce besoin a conforté cette nouvelle génération de modèles, qui recherchent à la fois la robustesse, c'est à dire la stabilité du comportement du modèle sur un nouveau jeu de données du même univers, et la vitesse de mise en œuvre, puisqu'il faut réagir en quelques secondes sur une autorisation de crédit ou l'octroi d'une police d'assurance où des dizaines de milliers de variables peuvent intervenir.

Problème principal de l'apprentissage (construction d'un modèle à partir de données existantes)

On dispose d'un ensemble de données, décrits par des lignes (événements) comprenant chacune n paramètres et une dernière colonne (la « question business »). On peut ainsi imaginer chaque ligne sous la forme $[x_1, ..,x_n \,|\, y\,]$, où y est appelé « question business ».

Appelons X le vecteur de $\Re$n : $X = (x_1,..,x_n)$. On va ici chercher à construire un modèle de $\Re^n$ dans $\Re$ (régression) ou de $\Re^n$ dans [0,1] (classification). Pour ce faire, un modèle calcule une fonction f(X,w) qui doit estimer y, où

- w est un paramètre de $\Re^p$ qui définit le modèle,
- $Z_i = (X_i,y)$ sont les valeurs possibles des données
- Q(z,w) est le coût des erreurs faites par le modèle en assimilant f(X,w) à y
- P(z) est la probabilité inconnue des données Z.

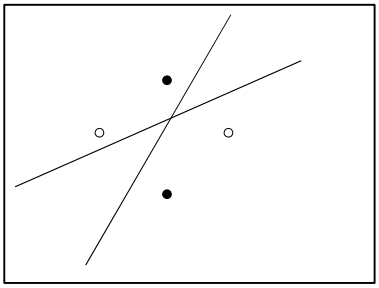
L'objectif est alors de minimiser sur w l'espérance de risque : $R(w) = \int Q(z,w)\, dP(z)$. Pour ce faire, nous disposons seulement de L cas d'apprentissage $(z_1, .. ,z_L)$, qu'on considère tirés suivant la inconnue loi P(z). On tachera donc de minimiser le Risque Empirique

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

La puissance de la théorie de Vapnik est d'aboutir à une majoration du risque R par la somme du risque empirique, mesuré sur l'ensemble d'apprentissage, et d'une quantité déterministe.

Un modèle est dit **consistant** si l'erreur de ce modèle sur des données nouvelles converge vers l'erreur du modèle sur les données d'apprentissage, lorsque la taille de l'ensemble d'apprentissage augmente.

Soit $f \in \Im$ la fonction qui décrit le modèle : Y = f(X,w). Vapnik associe à la famille de fonctions $\Im$ de $\Re^n$ dans $\Re$ un nombre entier h, appelé dimension de Vapnik-Chernovenkis de la famille $\Im$. Ce nombre caractérise pour la famille $\Im$ sa capacité à séparer (« complexité »), à « hacher » des points de l'espace $\Re^n$: une famille $\Im$ de fonctions de $\Re^n$ dans $\Re$ « hache » un jeu de points $(x_1, ..,x_L)$ de $\Re^n$ si quelle que soit la coloration des L points en m points blancs et L-m points noirs (il y en a 2^L possibles) , il existe une fonction particulière f de $\Im$ qui aura des valeurs positives sur les « blancs » et négatives sur les « noirs ». La famille $\Im$ de fonctions de $\Re^n$ dans $\Re$ a alors la VC



Une droite ne peut pas toujours séparer 4 points donc si F est l'ensemble des droites du plan $h_F=3$

dimension h s'il existe un jeu de h points de $\Re^n$ qui peut être « haché », aucun ensemble de h+1 vecteurs ne peut être « haché ». On montre par exemple que si $\Im$ est l'ensemble des droites du plan, alors h=3.

Le théorème majeur de Vapnik est alors le suivant :
- l'apprentissage du modèle f(X,w) est consistant si et seulement si la famille de modèles a une dimension h finie ;
- avec la probabilité 1-q,

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

Cette dernière équation est fondamentale :
- le risque du modèle « lâché dans la nature » est majoré avec la probabilité 1-q (seuil de risque, eg. q=1% ou 0.01) par la somme du risque empirique, mesuré sur l'ensemble d'apprentissage, et d'une quantité déterministe.
- Elle ne fait pas intervenir le nombre de variables du problème : c'est ce théorème qui permet de reprendre complètement l'approche intellectuelle de la modélisation statistique.
- Elle ne fait pas intervenir la loi de probabilité inconnue P(z), pour laquelle aucune hypothèse n'est formulée.
- le terme à la droite de E(w) tend vers zéro lorsque h/L tend vers 0.

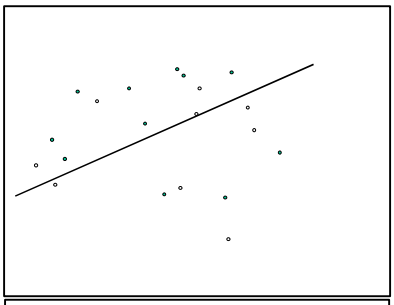
Même si la borne est dans la pratique trop élevée, elle montre qu'il est possible de contrôler l'erreur d'un modèle général « lâché dans la nature », même pour un très grand nombre de paramètres, à condition de choisir le modèle f(X,w) dans une famille de modèles $\Im$ dont la VC dimension h reste petite par rapport à L. Par ailleurs, même si le modèle fait intervenir des millions de variables, si le rapport h/L reste faible (1/20 est une bonne valeur pratique), le modèle est utilisable et robuste : il va donner des résultats comparables en test à ceux observés sur les données dont on dispose pour le bâtir (ensemble d'apprentissage).

Principe de la SRM *(Structured Risk Modelization)*

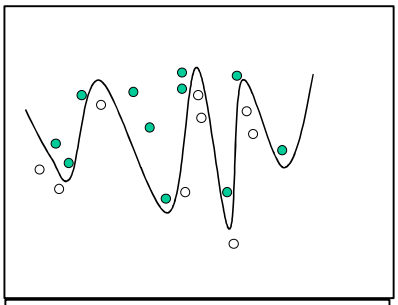
L'idée de la SRM est de construire une approche dans laquelle le modèle f(X,w) va être choisi parmi une famille $\Im_m$ de modèles donnée par un arbitrage entre la précision du modèle (le fit, E(w)) et la robustesse du modèle (caractérisée par l'inverse du terme $\sqrt{[\, h\,(\ln(2L/h) + 1\,) - \ln(q)\,]\,/\,L}$). Pour ce faire, on construit une succession de familles de modèles possibles, de plus en plus « riches » : $\Im_1 \subset \Im_2 \subset \ldots \subset \Im_p$ avec $h_1 < h_2 < .. < h_p$. Les familles de modèles étant de plus en plus « riches », le meilleur modèle de la famille $\Im_q$ aura une meilleure précision que le meilleur modèle de la famille $\Im_p$ si p est inférieur à q. Par contre, comme $h_p < h_q$, la robustesse (consistance) de ce modèle sera moins bonne.

Un concept de « poupées gigognes » de familles de fonctions $\Im_m$ peut s'effectuer de plusieurs manières :
- architecture de réseaux de neurones ;

– degré d'un polynôme ;
– contrôle des poids dans un réseau de neurones ;
– niveau de lissage dans un filtrage des données, etc.



Exemple de séparation d'un nuage de point par une droite



Séparation du même nuage par un polynome de degré plus élevé : le fit est meilleur mais la robustesse est réduite.

Modéliser dans une approche SRM (qui est présente à chaque étape de chacun des composants de Kxen), c'est remplacer le processus classique :

1. Effectuer une hypothèse sur la distribution statistique (inconnue) des données ;
2. Accepter qu'un grand nombre de dimensions du problème impose soit un grand nombre de paramètres et des temps de calcul prohibitif, soit des choix a priori de variables instrumentales avec leurs problèmes de consistance;
3. Chercher le meilleur fit et tester des hypothèses nulles.

Par le processus :

1. Etudier et construire la famille optimale $\Im$ au sens de la SRM, en contrôlant sa dimension de Vapnik-Chernovenkis h ;
2. Retenir tous les paramètres, puisque l'on contrôle par définition la consistance du modèle;
3. Chercher le meilleur compromis entre fit et robustesse / consistance.

Exemples concrets

Un exemple de famille $\Im$ facile à travailler est celui des polynômes à n variables : $Y = P(X_1, .. ,X_n)$.
Vapnik a bâti toute une théorie en écrivant directement les équations de consistance optimale pour des modèles du type : $Y = <w.X> + b$, où w est le paramètre recherché pour fixer le modèle et $<w.X>$ est le produit scalaire du vecteur de paramètres w et du vecteur X.

On est alors ramené à des équations de Lagrange et à un modèle d'optimisation quadratique sous contraintes linéaires : c'est la théorie des Support Vector Machines qui sort du cadre de ce papier, et qui permet de traiter des problèmes de très grande dimension dans un cadre théorique exact.

Une autre approche revient à contrôler la dimension h du polynôme par des approches basées sur la théorie des problèmes mal posés, c'est à dire des problèmes dont la

résolution numérique est fortement instable lorsqu'il y a du bruit dans les données (typique des données humaines). En injectant artificiellement du bruit dans les données, on contrôle le h et donc la robustesse du modèle : trop de bruit donne un modèle parfaitement idiot mais très stable sur de nouvelles données, aucun bruit donne un modèle très précis sur l'apprentissage, mais très instable sur de nouvelles données. On retrouve par l'approche SRM le meilleur arbitrage.

Cette seconde méthode permet de traiter des jeux de données importants (millions de lignes), jusqu'à 3000 paramètres environ. On peut ainsi modéliser un jeu de données brutes (chaînes de caractères en les encodant au vol) extrêmement rapidement : 250 variables descriptives (chaînes de caractères, variables numériques et ordinales) et un million de lignes (individus) en 100 minutes sur un PC ordinaire, contre trois semaines de travail auparavant avec des méthodes statistiques traditionnelles (où il faut tout tester, colonne après colonne).

Pour donner un exemple, une approche SRM (moteurs K2C pour l'encodage des variables, moteur K2R pour une régression robuste) a permis à Kxen de construire à partir des données brutes en 85 secondes un modèle de scoring de clientèle pour une assurance de caravanes, fonction de variables socioprofessionnelles données. Ce modèle est arrivé 5^{e}.[1]

Applications à l'assurance

Les modèles issus de la théorie de Vapnik permettent une exploitation optimale des données historiques disponibles dans les compagnies d'assurance ou dans les organismes professionnels (FFSA). On abordera ici deux exemples liés à l'assurance automobile.

La méthode de la SRM peut avantageusement se substituer aux méthodes de crédibilité utilisées dans la tarification automobile monovéhicule. En effet, une analyse historique des contrats en portefeuille permet de recueillir de nombreuses informations (x_i) comme la marque et le type du véhicule, l'année de souscription, son immatriculation, sa couleur, le lieu de résidence, le coefficient de bonus-malus du souscripteur. L'adjonction à ces informations brutes de la charge de sinistres rattachée à ce contrat (y_i) permet d'obtenir une estimation consistante de la prime pure à appliquer à chaque couple (conducteur, véhicule) et isole notamment les facteurs de risques. Ici c'est la capacité du modèle à réduire seul la dimension du problème qui est déterminante.

Appliquée à la base de données des sinistres, la SRM peut permettre une estimation fiable, à l'ouverture d'un sinistre et de minimiser ainsi les écarts sur les provisions pour sinistres à payer à constituer. L'impact pour les compagnies est loin d'être négligeable, les reprises sur PSAP étant lourdement taxées. Dans ce cas, les informations (x_i) sont constituées des différents éléments connus à l'ouverture du sinistre, telles qu'elles figurent dans les bases de données de la compagnie. (y_i) pourrait représenter le coût total du sinistre ou sa durée de traitement, en fonction de l'information recherchée. Avec l'aide de KXEN, une compagnie pourrait ainsi évoluer de façon fiable et optimale le coût total espéré d'un sinistre et éviter

[1] Ce concours mondial était organisé par P. van der Putten and M. van Someren (EDS) . CoIL Challenge 2000: The Insurance Company Case. Publié par Sentient Machine Research, Amsterdam. Leiden Institute of Advanced Computer Science Technical Report 2000-09. June 22, 2000 (Disponible à l'URL : http://www.liacs.nl/~putten/library/cc2000/.)

d'avoir recours à des méthodes de coût moyen dont les résultats sont souvent insatisfaisants. En effet, la distribution des sinistres étant "à queue épaisse", les sinistres exceptionnels peuvent perturber les calculs et le simple écrêtement donne parfois des résultats peu probants. Dans ce cas, c'est la capacité du modèle à identifier seul les bonnes variables "instrumentales" qui est un atout.

Demain, la possibilité de disposer de bases de données clients permettront de disposer d'information toujours plus fines sur les assurés et d'assurer une tarification toujours plus fine sur des millions de critères. La compagnie qui développera en premier ce savoir faire disposera alors d'un réel avantage concurrentiel.

Conclusion

L'approche de Vladimir Vapnik, en reprenant à la base la notion de modélisation prédictive, permet de conjurer la malédiction vieille de 25 ans des problèmes à très grands nombres de données.

Le jeu de théorèmes basés sur le concept de VC dimension d'une famille de fonction, et la stratégie de SRM (Structured Risk Modelization) donnent un cadre général où la plupart des anciennes méthodes se trouvent reprises, étendues, validées, des méthodes de ridge regression aux séries chronologiques, de l'analyse des correspondances aux problèmes de trade-off et de panels.

C'est pour la statistique un défi constructif, puisque désormais, dans les problèmes des bases de données du CRM, et notamment celles de la finance et de l'assurance, il devient possible d'approcher de nouvelles manières les calculs de comportement, à partir de fichiers qui comptent aujourd'hui des millions de variables et des centaines de millions de lignes dans les cas les plus étendus.

Références :

[1] V.Vapnik - The Nature of Statistical Learning Theory, Springer-Verlag, 1999 (2nd edition)
[2] V.Vapnik – Statistical Learning Theory – Wiley, 1998
[3] Aristote, Livre I, chap. vi
[4] Aristote, Livre VIII, chap vi
Voir aussi pour une étude plus approfondie :
[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 alii, Advances in Large Classifiers, MIT Press, 2000
[7] Bernard Schölkopf, Christopher Burges et alii, Advances in Kernel Methods, MIT Press, 1999

Remerciements :

Je tiens à remercier Sylvain Coriat, Generali France Assurances, qui a accepté de relire ce papier et y apporté sa précieuse contribution.