Abstract/Résumé

Expliquer ou prédire? Les nouveaux défis

G. Saporta[1]
[1] CEDRIC-CNAM, 292 rue Saint Martin, 75003 Paris, gilbert.saporta@cnam.fr

Mots clés: modèles, prévision, machine learning.

1 Introduction

Le développement de la *Data Science* suscite fréquemment des controverses entre statistique et *machine learning* et conduit à repenser le débat entre expliquer et prédire initié par Leo Breiman [1] en 2001. Breiman distinguait deux cultures dans la modélisation statistique : celle dominante jusqu'alors des modèles génératifs qui suppose que les données ont été engendrées par un modèle probabiliste qu'il faut ajuster et estimer, et la culture des modèles algorithmiques ou prédictifs qui ne se préoccupe que d'obtenir des prévisions précises et fiables et que les statisticiens avaient eu tort de négliger. Comme l'écrit David Donoho [2] qui a récemment repris ce thème, l'hypothèse implicite dans la première culture est qu'il existe un modèle « vrai », ce qu'a contesté George Box [3] avec sa phrase célèbre « *Essentially, all models are wrong, but some are useful* ». Dans l'intervalle le débat a fait l'objet de plusieurs publications : *cf.* Gilbert Saporta [4] qui relevait l'ambigüité du terme de modèle utilisé aussi bien comme une représentation de la réalité que comme un algorithme et Galit Shmueli [5] qui analysait la dualité explicatif/prédictif.

2 Critères

Si on considère le modèle usuel $y = f(x) + \varepsilon$, l'approche générative s'intéresse à des critères d'ajustement comme le R^2, ou le taux de bons classements, éventuellement pénalisés par la complexité du modèle tels l'*AIC* ou le *BIC* pour obtenir des modèles parcimonieux, mais toujours calculés sur l'ensemble des données disponibles ce qui conduit à des biais : on prédit le passé. L'approche prédictive mesure la qualité de prédiction sur de nouvelles observations (généralisation) et utilise systématiquement un partitionnement des données en apprentissage et test ou apprentissage, test et validation : c'est un des apports fondamentaux de la théorie de l'apprentissage statistique [6].

3 Paradigmes et paradoxes

Un modèle génératif se doit en général d'être compréhensible, donc simple, ce qui ne conduit pas nécessairement à de bonnes prévisions au niveau individuel. Bien des modèles épidémiologiques se contentent de détecter des facteurs de risque sans prétendre à une bonne précision au niveau individuel. Il existe également des modèles simples mais non génératifs, comme les arbres. Le paradigme de la boîte noire en apprentissage consiste à imiter le comportement du modèle $y = f(x) + \varepsilon$ plutôt que de chercher à identifier la fonction f, donc à prédire sans comprendre! Aux modèles simples viennent se rajouter les meta-modèles ou modèles d'ensemble qui combinent linéairement ou non les prévisions de différents algorithmes.

Selon Breiman [1] « *Modern statistical thinking makes a clear distinction between the statistical model and the world. The actual mechanisms underlying the data are considered unknown. The statistical models do not need to reproduce these mechanisms to emulate the observable data* ». Et Vapnik [7] de renchérir : « *Better models are sometimes obtained by deliberately avoiding to reproduce the true mechanisms* ».

D'ailleurs, même dans un contexte classique, un modèle biaisé (régularisé, avec moins de variables, etc.) peut donner, dans certaines conditions, de meilleures prévisions que le vrai modèle.

4 Défis

Les méthodes du Machine Learning ont fait leurs preuves dans de nombreux domaines impliquant données massives, modèles complexes et décision. La détermination de l'influence d'une variable sur une réponse reste cependant délicate, même dans des modèles aussi simples que le modèle linéaire [8]. L'usage des corrélations n'est pas suffisant, à l'encontre de ce que prétendait Chris Anderson [9] pour savoir comment on peut influer sur une réponse. Comprendre pour mieux prédire devient la nouvelle frontière et implique de revenir à la causalité, en procédant par exemple par expérimentation, ou à l'aide de méthodes adaptées aux données d'observation, comme les scores de propension [10] et le raisonnement contrefactuel [11]. L'inférence causale pour les données de grande dimension est également cruciale [12].

5 Réferences

[1] L.Breiman : Statistical Modeling: The Two Cultures, *Statistical Science*, 16, **3**, 199–231, 2001.

[2] D.Donoho : 50 years of Data Science, *Tukey Centennial workshop*, https://dl.dropboxusercontent.com/u/23421017/50YearsDataScience.pdf, 2015.

[3] G.Box & N.Draper : *Empirical Model-Building and Response Surfaces*, Wiley, 1987.

[[4] G.Saporta : Models for Understanding versus Models for Prediction, In P.Brito, ed., *Compstat Proceedings*, Physica Verlag, 315-322, 2008.

[5] G.Shmueli : To explain or to predict ? *Statistical science*, 25, **3**, 289–310, 2010.

[6] T.Hastie, R.Tibshirani, J. Friedman : *The elements of statistical learning : data mining, inference, and prediction*. Springer. Second edition, 2009.

[7] V.Vapnik : *Estimation of dependences based on empirical data*, second edition. Springer, 2006.

[10] H.Wallard : Using Explained Variance Allocation to analyse Importance of Predictors, in C.Skiadas, ed., *16th ASMDA Conference Proceedings*, 1043-1054, 2015.

[9] C.Anderson : The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, *Wired*, http://www.wired.com/2008/06/pb-theory/, 2008.

[10] P.R.Rosenbaum & D. Rubin : The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*. 70, **1**, 41–55, 1983.

[11] L.Bottou *et al.* : Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising, *Journal of Machine Learning Research*, 14, 3207–3260, 2013

[12] P.Bühlmann : Causal statistical inference in high dimensions, *Mathematical Methods of Operations Research*, 77, 357-370, 2013.