Models for Data or Models for Prediction?

Gilbert Saporta
Chaire de Statistique Appliquée & CEDRIC
CNAM
292 rue Saint Martin, F-75003 Paris

gilbert.saporta@cnam.fr
http://cedric.cnam.fr/~saporta

Outline

1. Introduction
2. Model for data: a few problems
3. Model choice
4. Looking for the « true » model
5. Models for prediction
6. Concluding remarks

1.Introduction

- Statistical modelling aims at:
  - Providing some understanding of data and of its underlying mechanism through a parsimonious representation of a random phenomenon. Usually needs both a statistician and an expert of the application field.
  - Predicting new observations with a high accuracy.

- Understanding vs Prediction does not overlap with Unsupervised vs Supervised
  - « Understanding » could mean either a parametric distribution model for a random vector or an explanatory or a regression model involving a response $y= f(x;\theta)+\varepsilon$
  - A model should be simple, and parameters should be interpretable in terms of the application field : elasticity, odds-ratio, etc.
    - Hence the preference for logistic regression instead of discriminant analysis.

- Paradox 1
  - A « good » statistical model should give insights in the nature of a stochastic phenomenon, not necessarily gives accurate predictions. In epidemiology eg, it is more important to find risk factors than having a prediction of getting some disease at an individual level.
  - Different from physics where a good model must give good predictions, otherwise it is replaced by an other one.
  - Is statistics a science or only technology? (cf C.R.Rao)

- Paradox 2

  Data Mining and KDD prove daily that automatic prediction is possible without understanding.

  - In Customer Relationship Management or pattern recognition, understanding is often a vain task: a banker does not need a theory for predicting if a loan will at risk or not, but only a good score function
  - Here models are just algorithms, even black-boxes, and the quality of a model is assessed by its performance for predicting new observations.

Same formula: $y= f(x;\theta)+\varepsilon$

- **Classical framework**
  - Underlying theory
  - Narrow set of models
  - Focus on parameter estimation and goodness of fit
  - Error: white noise

- **Predictive modelling**
  - Models come from data
  - Algorithmic models
  - Focus on control of generalization error
  - Error: minimal

2. Models for data: a few problems

- The statistician and the scientist
  - A naive view: the scientist (economist, biologist, etc.) first specifies a model, then the statistician estimates the parameters, and (or) refutes the model according to some goodness of fit test. If the model is rejected, the scientist should think of an other one .

- Estimation is not that easy:
  - Needs a general technique
    - Maximum likelihood won over the method of moments and the minimum chi-squared but only recently (see Berkson, 1980 for logistic regression)
    - Least squares still popular since: often more robust, needs less assumptions (eg PLS)
  - Needs data!
    - In order to apply asymptotic results
    - To have unique and good estimates

- If few data:
  - Use constrained estimators: eg ridge regularization
  - Become a Bayesian!
    - Using ridge makes you a non-conscious bayesian
  - Both solutions lead to difficulties in goodness of fit tests:
    - Which degree of freedom?
    - Overparametrized models cannot be rejected

3. Model choice

- When the "expert" hesitates between several formulations
  - Inside a common family
  - The major use is for variable selection
- Parsimony
  - Ockham's* razor: a scientific principle against unnecessary hypotheses

* Or Occam

?

William of Ockham
(1285-1348)

from wikipedia

An English Franciscan friar and scholastic philosopher. He was summoned before the Papal court of Avignon in 1324 under charges of heresy and was excommunicated for leaving Avignon, but his philosophy was never officially condemned.
William of Ockham has inspired in U.Eco's The Name of the Rose, the monastic detective William of Baskerville, who uses logic in a similar manner.

lex parsimoniae:

entia non sunt multiplicanda praeter necessitatem

entities should not be multiplied beyond necessity

Ockham's razor states that the explanation of any phenomenon should make as few assumptions as possible, eliminating, or "shaving off", those that make no difference in the observable predictions of the explanatory hypothesis or theory.

- Many criteria for model selection :
  - Minimum description length
  - Mallow's C_p
  - Penalized likelihood: trade-off between the fit (measured by the likelihood) and the complexity (measured by the number of parameters)

- The likelihood principle (Fisher, 1920)
  - sample of n iid observations:

$$L\left(x_1,..,x_n;\theta\right)=\prod_{i=1}^{n} f\left(x_i;\theta\right)$$

  - Given f, the best estimation for θ is the one which maximizes the likelihood, ie the probability of having observed the data. Hence the best model should have the same property.
  - But likelihood increases with the number of parameters..

penalized likelihood

Akaike : $AIC = -2 \ln(L) + 2k$

Schwartz : $BIC = -2 \ln(L) + k \ln(n)$

- Rule: choose the model which minimizes AIC or BIC
- BIC favourises more parsimonious models than AIC due to its penalization
  - Generalizations such as $AIC3 = -2 \ln(L) + 3k$ or ICOMP (Bozdogan)
- AIC and BIC have similar formulas but originates from different theories: there is no rationale to use simultaneously AIC and BIC

- AIC : approximation of Kullback-Leibler divergence between the true model and the best choice inside the family

$$I(f;g)=\int f(t)\ln\frac{f(t)}{g(t)}dt=E_f(\ln(f(t))-E_f(\ln(g(t))$$

$$E_{\hat{\theta}}E_f(\ln(g(t;\hat{\theta}))\square\ln(L(\hat{\theta}))-k$$

- BIC : bayesian choice between m models M_i . For each model $P(\theta_i / M_i)$. The posterior probability of M_i knowing the data **x** is proportional to P(Mi) $P(\mathbf{x}/M_i)$. With equal priors $P(M_i)$:

$$\ln(P(\mathbf{x}/M_i)) \approx \ln(P(\mathbf{x}/\hat{\theta}_i, M_i)) - \frac{k}{2}\ln(n)$$

- The model with the highest posterior probability is the one with minimal *BIC*

Penalized likelihood: some limitations (1)

The likelihood or (and) the number of parameters could be hard to define

Cannot be simply applied to constrained estimation like ridge or PLS regression:

- Right number of parameters?
  - p parameters constrained by $\|\beta\| \leq c$
  - Should be much less than p if c is low but the exact formula is unknown.

Penalized likelihood: some limitations (2)

- What is the AIC of a decision tree or a neural network? How to choose between them?
  - However may be not relevant in a classical modelling framework: different from the data generating process.
- Is it realistic to assume an uniform prior on all potential models for BIC ?

4. Looking for the « true » model

- AIC is biased : if the true model M_i belongs to the family, the probability that AIC chooses M_i does not tend to 1 when the number of observations goes to infinity. However BIC converges.
- When n is finite, everything is possible!
- A parsimonious model may not be the true one: statistical variable selection may discard truly influential variables. But is it the question?

"The Truth Is Out There" (X-Files, 1993)

- No simple parsimonious model can fit to large data sets and tests are in general useless.
  - For millions of observations, a correlation of 0.01 is significantly different from zero. Is it useful?
  - Most standard models are wrong for real data like the multinormal one often used in PCA to make inference in contradiction with the goal of exploratory data analysis, of revealing the underlying structure of heterogenous data.

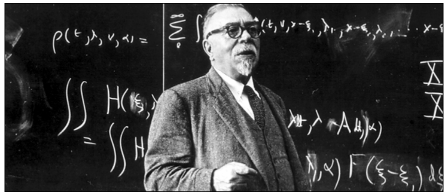
"Essentially, all models are wrong, but some are useful " G.Box (1987)

  - Fortunately for categorical data the multinomial distribution applies provided that the sampling scheme is simply random

5. Models for prediction

- In Data Mining applications (CRM, credit scoring etc.) models are used to make predictions.
- Model efficiency: capacity to make good predictions (predicting the future) and not only to fit to the data (predicting the past)

5.1 The black-box problem and supervised learning (N.Wiener, V.Vapnik)

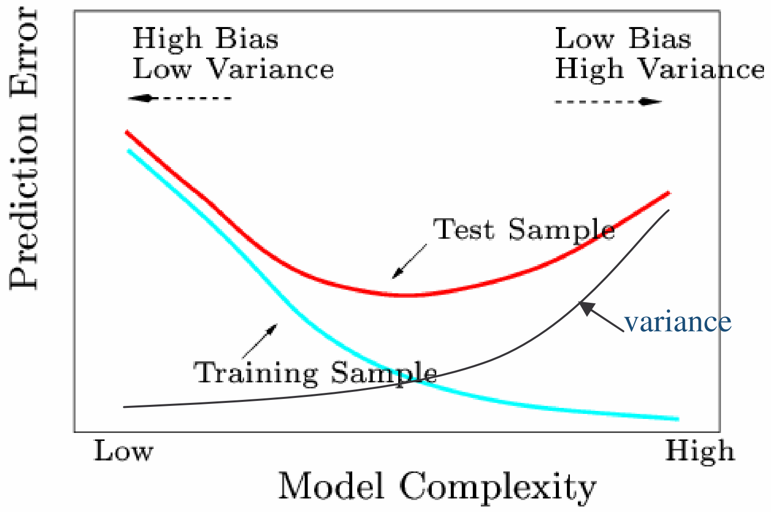


- Given an input x, a non-deterministic system gives a response variable y = f(x)+e. From n pairs (x_i,y_i) one looks for a function $\hat{f}$ which approximates the unknown function f.
- Two conceptions:
  - A good approximation is a function $\hat{f}$ close to the true f (model of data)
  - A good approximation is a function $\hat{f}$ which has an error rate close to the black box, ie which performs as well as f (model for prediction)

5.2 Model complexity and the error of prediction

$$E\left(y_0-\hat{y}_0\right)^2 = \sigma^2 + E\left(f(x_0)-\hat{f}(x_0)\right)^2 =$$

$$\sigma^2 + \underbrace{\left(E\left(\hat{f}(x_0)\right)-f(x_0)\right)^2}_{\text{bias}} + \underbrace{V\left(\hat{f}(x_0)\right)}_{\text{variance}}$$



Adapted from Hastie et al.

Model complexity (cont.)

- The more complex a model, the better the fit but with a high prediction variance.
- Optimal choice: trade-off
- But how can we measure the complexity of a model?
  - Not by the number of parameters
  - For binary classification Vapnik's SLT proposes the VC dimension. VC is connected to the maximum number of points which can be separated by the family of functions (the model) whatever are their labels ±1

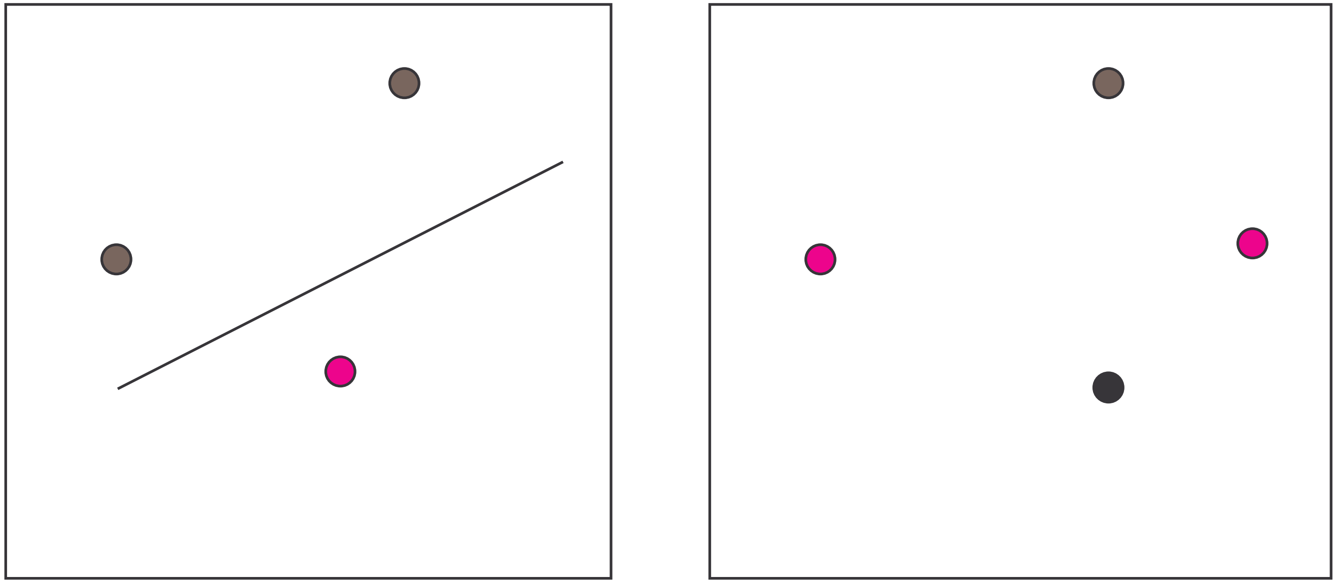
Vapnik-Cervonenkis dimension for binary supervised classification

- A measure of complexity related to the separating capacity of a family of classifiers.

- Maximum number of points which can be separated by the family of functions whatever are their labels ±1

Example

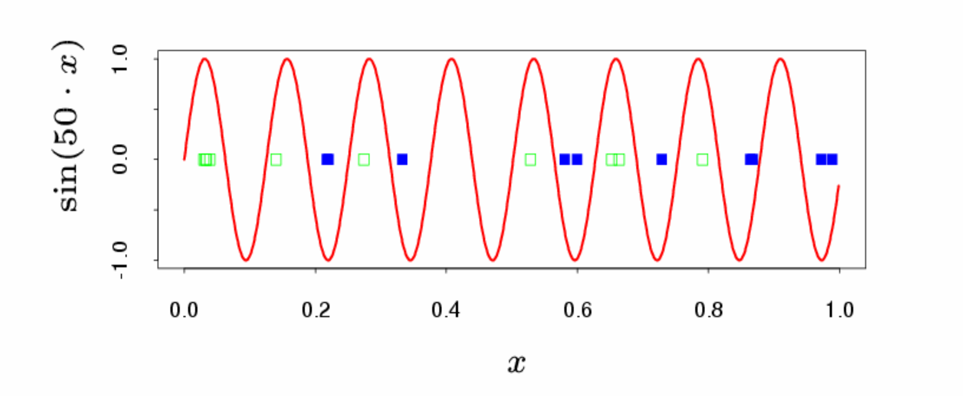
- In 2-D, the VC dimension of “free” linear classifiers is 3



- **But VC dimension is NOT equal to the number of free parameters: can be more or less**

  - The VC dimension of $f(x,w) = sign\ (sin\ (w.x)\ )$
    $c < x < 1,\ c>0,$
    with only one parameter w is infinite.

Hastie et al. 2001



5.3 Empirical risk and generalization

- Loss function $L(y;f(x,w))$
  - *Regression* $L(y;f(x,w))=(y-f(x))^2$
  - *Discrimination : misclassification rate (or cost)*
    - y *and* $\hat{y}$ *belong to* $\{-1\ ;+1\}$ $L(y;\hat{y})=\frac{1}{2}|y-\hat{y}|=\frac{1}{4}(y-\hat{y})^2$

- Risk or expected loss on new data $z = (X, y)$

$$R = E(L) = \int L(z,w)dP(z)$$

- An impossible task: minimize R on w
  - $P(z)$ unknown probability distribution
- Given n learning observations $(z_1, .. , z_n)$ sampled from $P(z)$, empirical risk minimization:

$$R_{emp} = \frac{1}{n}\sum_{i=1}^{n} L(y_i ; f(x_i ; w))$$

- Example: OLS in regression

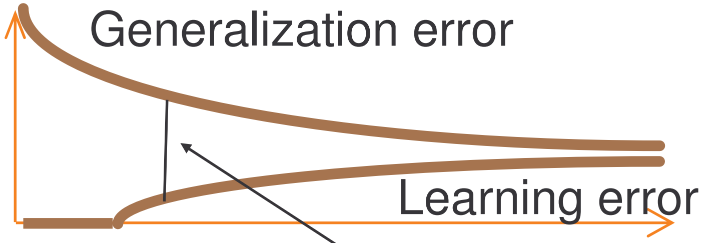
$$R_{emp} = \frac{1}{n}\sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

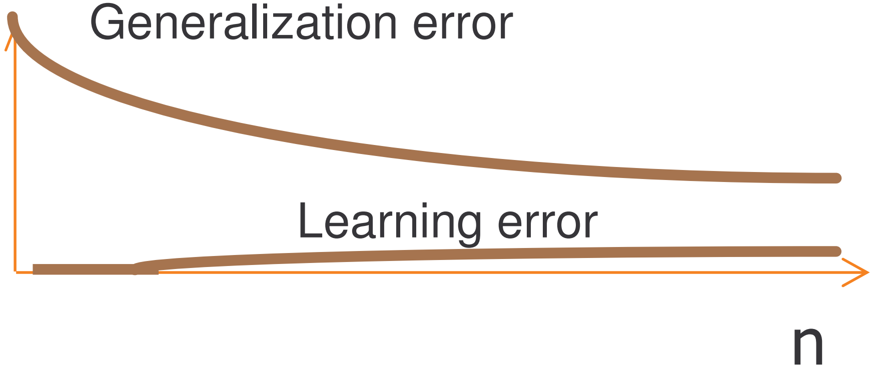
- Central problem in learning theory:
  Which is the relationship between R and empirical risk R_{emp} ? Is there convergence?

- What is the generalization capacity of this kind of model?

- Consistent learning

- Non consistent learning





h must be finite

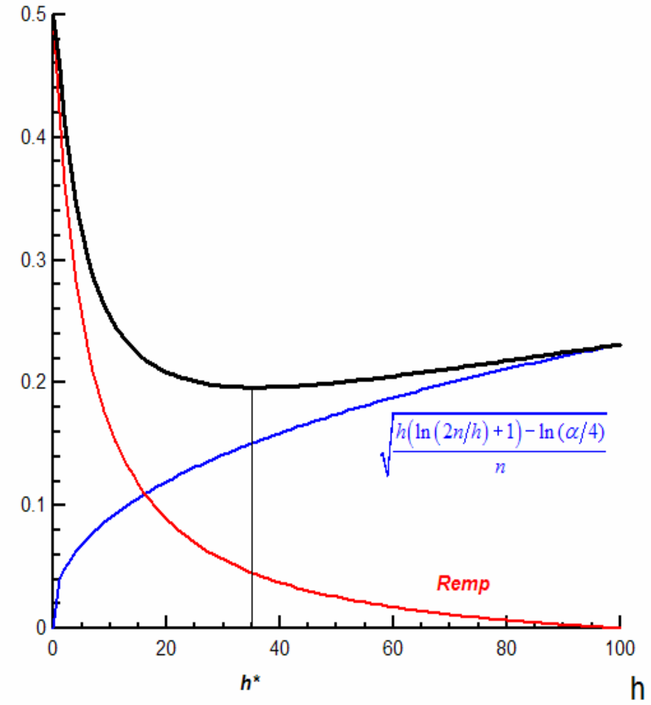
Vapnik's inequality $R < R_{\mathrm{emp}} + \sqrt{\frac{h\left(\ln\left(2n/h\right)+1\right) - \ln\left(\alpha/4\right)}{n}}$

Consequences of Vapnik's inequality

- The complexity of a family of models may increase when n increases: main difference with BIC
- Small values of h gives a small difference between R and R_{emp} . Regularization including dimension reduction techniques, provides better results in generalization than ordinary least squares.

5.4 Model choice by Structural Risk Minimization (SRM)

- For a family of embedded models with increasing VC dimension: instead of minimizing R, one minimizes the upper bound: R_{emp} + confidence interval
- For instance: polynomials of increasing degrees; ridge with decreasing C; neural networks with increasing number of nodes ot layers, etc.



Model choice by Structural Risk Minimization (SRM)

- Since it is an universal inequality, the upper bound may be too large.
- However: for any distribution , SRM provides the best solution with probability 1 (universally strong consistency) Devroye (1996) Vapnik (2006).
- SRM theory proved that the complexity differs from the number of parameters, and gives a way to handle methods where penalized likelihood is not applicable.
- Exact VC-dimension are very difficult to obtain, and in the best case, one only knows upper bounds .

about prediction criteria

- Global error rate needs a complete decision rule or classifier with a threshold
- AUC integrates all thresholds for a specific score function (but with respect to a non-uniform measure)
- Taking into account misclassification costs lead to a different loss function

5.5 Empirical model choice

- **The 3 samples procedure** (Hastie & al., 2001)
  - Learning set: estimates model parameters
  - Test : selection of the best model
  - Validation : estimates the performance for future data
- **Resample** (eg: bootstrap, 10-fold CV, …)
- **Final model : with all available data**
  - **Estimating model performance is different from estimating the model**

6 . Concluding remarks

- « There are two problems in modern science:
  - Too many people use different terminology to solve the same problem
  - Even more people use the same terminology to address completely different issues »

  (V.Cherkassky, F.Mulier, 1998)

- Two very different conceptions correspond to the same name of «model»: models of data ≠ models for prediction

- Models for understanding data correspond to the part of statistics considered as an auxiliary of science. Models for prediction belong to the other face of statistics as a decision making methodology.
- Are science and action opposed? a technique which gives good predictions improves also our knowledge. Predictive modelling belongs to empiricism (a theory of knowledge not to be confounded with pragmatism).

obrigado

References

- Borra, S. and Di Ciaccio, A.: Measuring the prediction error. A comparison of cross-validation, bootstrap and hold-out methods, in Ferreira et al. *Proceedings IASC 07*, Aveiro, Portugal
- Box, G.E.P., Draper, N.R.: *Empirical Model-Building and Response Surfaces*, p. 424, Wiley, 1987
- Cherkassky, V., Mulier, F. Learning from data , Wiley, 1998
- Devroye, L., Györfi, L., Lugosi, G. (1996) *A Probabilistic Theory of Pattern Recognition*, Springer
- Hand, D.J. (2000) Methodological issues in data mining, in J.G.Bethlehem and P.G.M. van der Heijden (editors)*, Compstat 2000 : Proceedings in Computational Statistics*, 77-85, Physica-Verlag
- Hastie, T., Tibshirani, F., Friedman J. (2001) *Elements of Statistical Learning*, Springer
- Vapnik, V. (2006) *Estimation of Dependences Based on Empirical Data*, 2nd edition, Springer

The 19th International Conference on Computational Statistics : Paris, France, August 23rd-27th 2010

http://www.compstat2010.fr
info@compstat2010.fr