1st Workshop of the ANR/DFG Project MADRAS

13–16.12.2021, Université Lyon 1

WP1 — Raphael Korbmacher and Antoine Tordeux

Bergische Universität Wuppertal

“Models for Understanding versus Models for Prediction”

Gilbert Saporta, *COMPSTAT* 2008, pp. 315–322

"Models for Understanding versus Models for Prediction"

Gilbert Saporta, *COMPSTAT* 2008, pp. 315–322

- Opposition between two modelling approaches in statistic (and elsewhere):

  1. **Model to understand :** Parsimonious representation of data to identify underlying mechanisms and parameters which may have produced it.
  2. **Model to predict :** Models whose complexity depends on the quantity and structure of the data that are assessed by its performances to predict new observations.

- **Author :** Gilbert Saporta

  University professor emeritus at the CNAM

  Research field: Applied Statistic, Statistical Computing

  Author of the French best-seller in statistic:
  Probabilités, analyse des données et statistique, Technip, 1990

Content

Models for understanding

Models (Algorithms) for prediction

Applications

Content

Models for understanding

Models (Algorithms) for prediction

Applications

Models for understanding

- Models for understanding: **Identification of underlying mechanisms**
  - $\rightarrow$ **Insights in the nature** of the phenomenon of interest
  - $\rightarrow$ **Few parameters** that should be **interpretable and that can be estimated using data**
  - $\rightarrow$ **Parsimony principle**

Models for understanding

- Models for understanding: **Identification of underlying mechanisms**
  - → **Insights in the nature** of the phenomenon of interest
  - → **Few parameters** that should be **interpretable and that can be estimated using data**
  - → **Parsimony principle**

- **Occam's razor** attributed to William of Ockham (1287–1347)

"*Among competing hypotheses, the one with the fewest assumptions should be selected*"

- Ptolemy (90–168) "*We consider it a good principle to explain the phenomena by the simplest hypothesis possible*"
- Isaac Newton (1642–1727) "*We are to admit no more causes of natural things than such as are both true and sufficient to explain their appearances*"
- Albert Einstein (1879–1955) "*Everything should be made as simple as possible, but not simpler*"

Models for understanding

- **Model :** $$y = f(x;\theta) + \varepsilon$$
  - y : **Variables to explain/predict** — *Dependent variables, regressand, output variable, ...*
  - x : **Explanatory variables** — *Independent variables, regressor, input variable, ...*
  - θ : **Parameters of the model** — *Constants to calibrate and interpret*
  - ε : **Unexplained part** — *Noise (residual) with amplitude* σ

Models for understanding

- **Model :** $$y = f(x;\theta) + \varepsilon$$
  - y : **Variables to explain/predict** — *Dependent variables, regressand, output variable, ...*
  - x : **Explanatory variables** — *Independent variables, regressor, input variable, ...*
  - θ : **Parameters of the model** — *Constants to calibrate and interpret*
  - ε : **Unexplained part** — *Noise (residual) with amplitude σ*

- Examples of parametric models : **Linear and nonlinear regression model, PLS regression** (quantitative analysis); **Logistic model, (linear) discriminant, posterior distribution** (qualitative analysis, classification)
- **Parameter calibration**: Least-squares, maximum-likelihood, Bayesian network + Confidence (credible) interval
- **Model choice:** Information criteria (Likelihood-ratio; Akaike, AIC; Bayesian, BIC, Bayes-factor) + Statistical test

Models for understanding: Limit

- **Difficulties with Big data** *Large dimension and observation number*
  - Concentration of the likelihood: Information criteria $AIC = -2ln\big(L_n(\hat{\theta})\big) + 2k$ or $BIC = -2\ln\big(L_n(\hat{\theta})\big) + \ln(n)k$ tend to select **models with minimal number of parameters**
  - **Everything is significant** ($CI = \left[\hat{\mu} \pm q\hat{\sigma}/\sqrt{n}\right] = \{\hat{\mu}\}$, $cor = 0.01$ significant, ...)

Models for understanding: Limit

- **Difficulties with Big data** *Large dimension and observation number*

  - Concentration of the likelihood: Information criteria $AIC = -2ln(L_n(\hat{\theta})) + 2k$ or $BIC = -2\ln\left(L_n(\hat{\theta})\right) + \ln(n)k$ tend to select **models with minimal number of parameters**

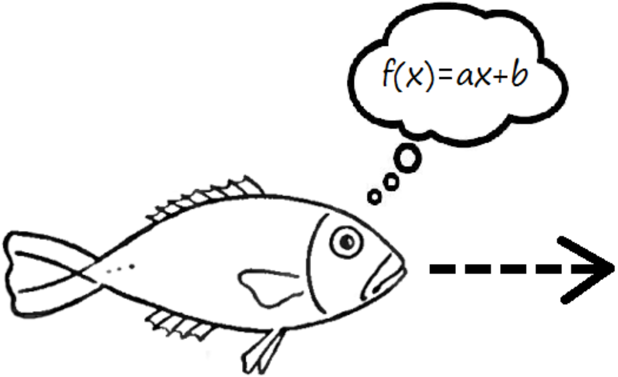
  - **Everything is significant** ($CI = [\hat{\mu} \pm q\hat{\sigma}/\sqrt{n}] = \{\hat{\mu}\}$, $cor = 0.01$ significant, ...)

- **Difficulties with complex multidimensional nonlinear relationship** *Complex system*

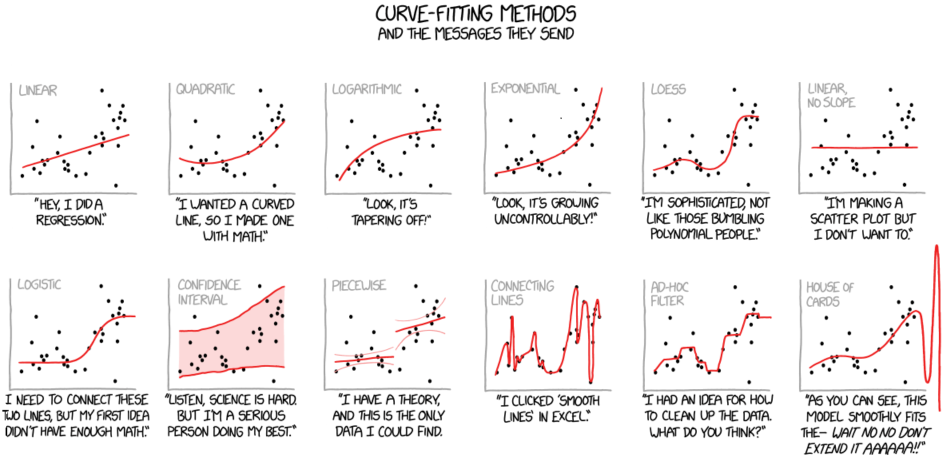
  - Correlation-based model : **Linear relationship** / Least squares: **for linear models**

  - **Modelling-bias – Limited modelling complexity**

George Box (1919–2013) : *"Essentially, all models are wrong, but some are useful"*


f(x)=ax+b

Models for understanding: Illustration II



Source: 2021/01/07/xkcd-curve-fitting-methods-and-the-messages-they-send

Content

Models for understanding

Models (Algorithms) for prediction

Applications

Models for prediction

- **Origins :** *Knowledge discovery in data bases* G. Piatetsky-Shapiro, 1980
  - A model is merely an algorithm coming more from the *data* than from a *theory*
    $\rightarrow$ No "Modelling bias"
  - Algorithm complexity itself (hyperparameter) depends on the data structure and size
  - Focus on prediction ability, i.e. capacity of making good predictions for new data

Models for prediction

- **Origins :** *Knowledge discovery in data bases* — G. Piatetsky-Shapiro, 1980
  - A model is merely an algorithm coming more from the *data* than from a *theory*

    $\rightarrow$ No "Modelling bias"
  - Algorithm complexity itself (hyperparameter) depends on the data structure and size
  - Focus on prediction ability, i.e. capacity of making good predictions for new data

- **Models for prediction : "Black-box" models** — Vladimir Vapnik, 2006
  - Same formulation $y = f_H(x;\theta) + \varepsilon$ but here f is a non-linear function depending on hyperparameters H and the dimensions of x and θ are high
  - Exemples of algorithms for prediction: Neural network, support-vector-machine, random forest – Hyperparameters: number of neurones, support vectors, decision trees.
  - Supervised learning : Training minimising a loss function (squared error, cross-entropy)
  - Black-box because the coefficients are too numerous to be interpreted and because the algorithm structure and complexity depend on the data

Can we open the black box of AI?

Davide Castelvecchi, *Nature* **538**, 20–23, 2016

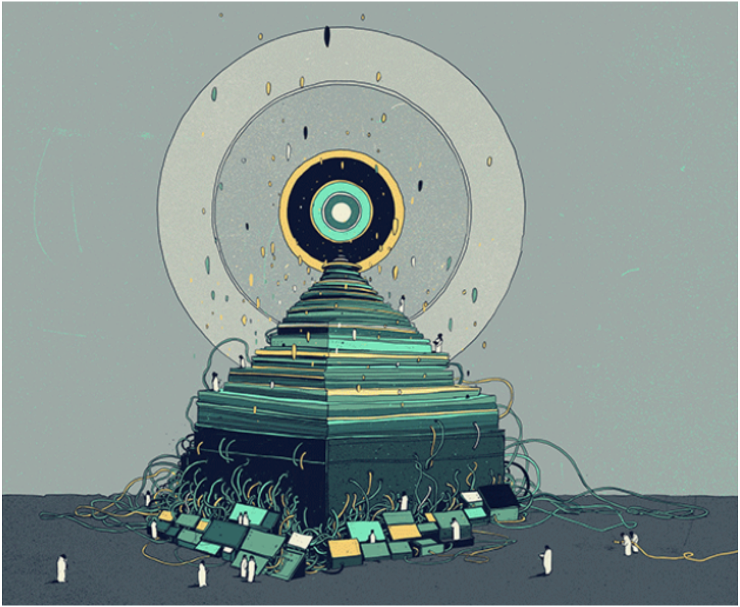


Illustration by Simon Prades

Models for prediction: Theory

- **Risk minimization**
  - L is a **loss** function, the **risk** $R = E(L)$ is the expectation of the loss
  - **Empirical risk:** $R_{emp} = \frac{1}{n} \sum_i L\big(y_i, f(x_i;\theta)\big)$

Models for prediction: Theory

- **Risk minimization**
  - L is a **loss** function, the **risk** $R = E(L)$ is the expectation of the loss
  - **Empirical risk:** $R_{emp} = \frac{1}{n}\sum_i L\big(y_i, f(x_i;\theta)\big)$

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

  with h the Vapnik–Chervonenkis dimension (i.e. the cardinality of the largest set of points that the algorithm can shatter — prediction ability)

  - No distributional assumptions are necessary (only $h \ll n$)
  - Formally Risk shared between empirical risk and a function depending on the ratio h/n (ratio h/n of interest)
  - Minimisation of the empirical risk by increasing the model complexity h
  - Increase of the complexity and prediction ability h as n increases

Models for prediction: Practice

- The VC-dimension is difficult to evaluate in practice
- **Setting the algorithm complexity:** Trade-off between quality-of-fit and training robustness
  - Too simple algorithm: precise training but weak prediction
  - Too complex: imprecise training, good prediction for training but weak for new data
  - **Bias-Variance-Dilemma** (underfitting VS overfitting)

Models for prediction: Practice

- The VC-dimension is difficult to evaluate in practice
- **Setting the algorithm complexity:** Trade-off between quality-of-fit and training robustness
  - Too simple algorithm: precise training but weak prediction
  - Too complex: imprecise training, good prediction for training but weak for new data
  - **Bias-Variance-Dilemma** (underfitting VS overfitting)

- **Empirical analysis of the algorithm complexity**
  - **Cross-validation :** (random) partition of the data in training and testing set

    Training set used to fit the models

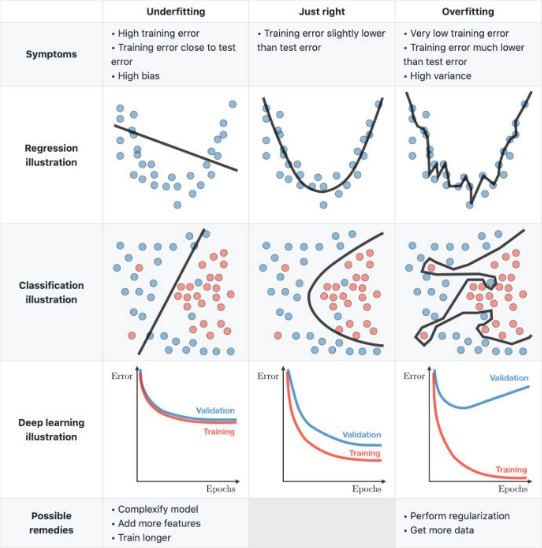
    Validation set use to estimate prediction error
  - **Bootstrap aggregating:** Repeating the operation to evaluate the precision of estimation
  - Algorithm complexity selection **by minimising the mean testing error** (cross-validation)
  - Evaluation of the estimation precision using the **empirical bootstrap distribution**

Bias-Variance-Dilemma

[1]

1. Source : elitedatascience.com/bias-variance-tradeoff

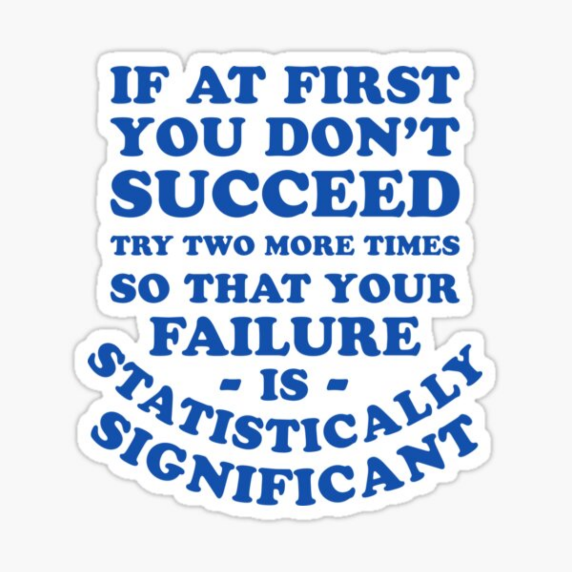
Cross-validation (underfitting VS overfitting)



	Underfitting	Just right	Overfitting
Symptoms	• High training error • Training error close to test error • High bias	• Training error slightly lower than test error	• Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			
Deep learning illustration	Error, Validation, Training, Epochs	Error, Validation, Training, Epochs	Error, Validation, Training, Epochs
Possible remedies	• Complexify model • Add more features • Train longer		• Perform regularization • Get more data

[2]

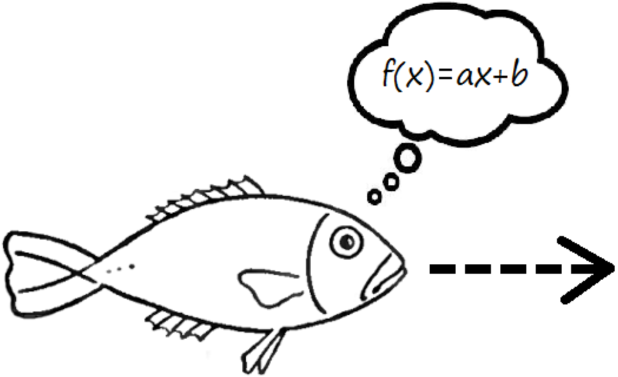
2. Source: kaggle.com/getting-started/166897



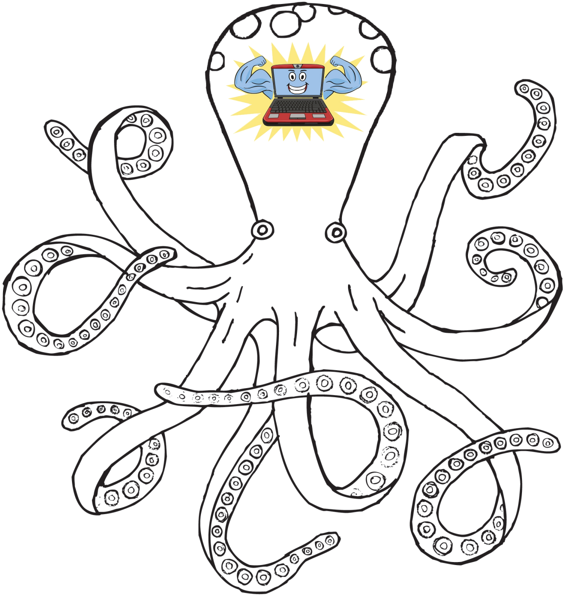
[3]

3. Popular quote

Models for understanding: Illustration



Models for prediction: Illustration



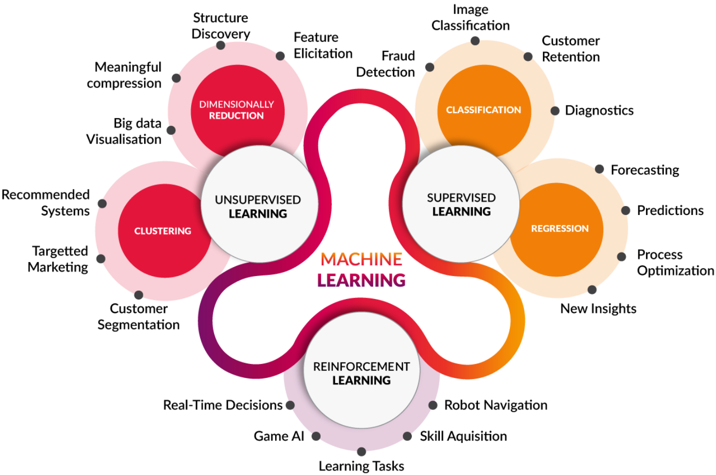
Content

Models for understanding

Models (Algorithms) for prediction

Applications

Models for prediction and machine learning: Applications in engineering



[4]

4. Source: Towards data science

Applications: Autonomous vehicles

- Driving situations are extremely varied and the driving process is poorly structured
- Defining an understandable model giving satisfying responses in any situation is not possible (especially in urban/dense situations or for mixed flow)

$\rightarrow$ Autonomous driving is a typical application field for machine learning techniques and the 'models for prediction'

Applications: Autonomous vehicles

- Driving situations are extremely varied and the driving process is poorly structured

- Defining an understandable model giving satisfying responses in any situation is not possible (especially in urban/dense situations or for mixed flow)

  $\rightarrow$ Autonomous driving is a typical application field for machine learning techniques and the 'models for prediction'

- The perception and motion planning of autonomous vehicles by machine learning actively developed since the 1990's

  - **Projects:** NAVLAB (1984), Eureka Prometheus (1985), NAHSC (1997), Cybercar (1997), Darpa Challenges (2007), Google Car (since 2010), Tesla (since 2014), PROUD (2015), DELPHI (2016), VIAC Challenge, GCDC, ...

"Simple" Neural networks
based on video analysis

Experiment:

2mn learning (120 obs only!)

→ Autonomous steering in curved roads

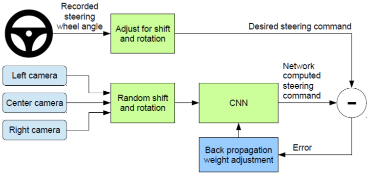
Autonomous vehicles: Example (2)

Recent work (End-to-End Deep Learning for Self-Driving Cars, Bojarski et al., 2016)

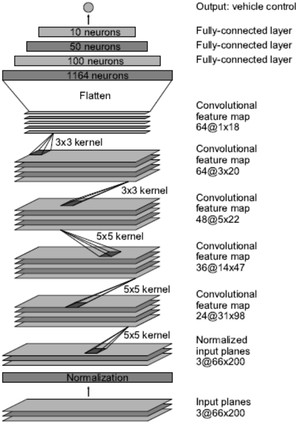
Convolutional neural networks
based on HD video analysis

DAVE-2 Project (DARPA Challenge)

Neural network: 27 M connections
and up to 250 000 parameters!



Training phase



CNN architecture

Revolution of the data science?

- Data-based approaches and machine learning techniques have clearly transformed the engineering sciences over the last 20 years
- Keywords: **Data Science, Internet of Things, Big Data, Industry 4.0, Sensor 4.0**, etc.

Revolution of the data science?

- Data-based approaches and machine learning techniques have clearly transformed the engineering sciences over the last 20 years
- Keywords: **Data Science, Internet of Things, Big Data, Industry 4.0, Sensor 4.0**, etc.

- **Revolution in the science?**
  - A. Ourmazd: Science in the age of machine learning. *Nat. Rev. Phys.* 2(7):342, 2020.
  - A. W. Senior et al.: Improved protein structure prediction using potentials from deep learning. *Nature* 577(7792): 706, 2020.
  - F. Cichos et al.: Machine learning for active matter. *Nat. Mach. Intell.* 2(2):94, 2020.
  - A. Esteva et al.: Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115, 2017.
  - K. T. Schütt et al.: Quantum-chemical insights from deep tensor neural networks. *Nature Communications* 8(1):1, 2017.
  - K. T. Butler et al.: Machine learning for molecular and materials science. *Nature* 559(7715):547, 2018.
  - K. G. Reyes and B. Maruyama: The machine learning revolution in materials? *MRS Bull.* 44(7):530, 2019.