Statistics Café

A seminar series on machine learning

Objectives

- To predict or to explain?
- What is machine learning?
- Popular/powerful approaches
  - Classification and decision trees
  - Artificial neural networks
  - Understanding the architecture/theory
  - Strengths and weaknesses → choosing a method
  - Applications in ecological research

Disclaimer

- We're no experts in this field!
- Interactive seminar series with the objective to teach ourselves
- External experts to extend the self-taught basics

Schedule

24.10. (SH)	Statistical modeling: the two cultures Interpretation vs Prediction and the role of algorithmic models
07.11. (SH?)	Tree-based methods: Regression and classification trees
21.11. (CFD)	Tree-based ensemble methods: bagging, random forests, boosting
05.12. (CS)	Introduction to artifical neural networks (ANNs): theory, application, examples
19.12.	From linear regression to ANNs without hidden layers: feed-forward and backpropagation
16.01.	Convolutional neural networks: theory, application, examples
01.02. (Friday!)	Deep learning (external: Dr Pan Kessel, machine learning group, TU Berlin)
13.02.	Wrap-up: What have we learned? The role of predictive modelling/machine learning/algorithmic models in ecology and evolution.

Statistical modelling

The two cultures

Leo Breiman, 2001

Goals in Science

Breiman, 2001

Goals in Science

- Describing

  Estimating population size, occupancy probability, etc.

Dynamic ecology blog; McGill, 2018

Goals in Science

- Describing

  Estimating population size, occupancy probability, etc.

- Understanding

  Causal relationships: drivers of species distribution, mechanisms of diversification, etc.

Dynamic ecology blog; McGill, 2018

Goals in Science

- Describing

  Estimating population size, occupancy probability, etc.

- Understanding

  Causal relationships: drivers of species distribution, mechanisms of diversification, etc.

- Predicting

  Population size in 10 years from now, predict species distribution in inaccessible area

Dynamic ecology blog; McGill, 2018

Goals in Science

- Which goal do you pursue?
  Describing, Understanding, Predicting
- What's your experimental design?
  Experiment, Observation
- Which analysis tools do you use?
  t-test, ANOVA, GLM, GLMM, GAM, random forest, neural networks, ...?

To explain

- As Breiman puts it: “The data modelling culture”
- Nature = stochastic model
    - Linear regression
    - Logistic regression
    - …
- Assumption: We know Nature’s structure
- Used to test hypotheses
- Simple, interpretable picture of the relationship between x and y

To explain

- ‘explaining’
  - Following the ‘gold standard’ of science: highly controlled experimental designs
  - Likely to know Nature
  - Inferring causality: drivers of changes in y

- ‘explanatory modelling’
  - Field observations of y and x
  - Unlikely to know Nature → assumptions
  - Inferring correlates of y

Shmueli, 2010

Limitations of interpretation

- ‘explaining’
    - Learning about small, contained parts of Nature
    - Predictions might still be bad
- ‘explanatory modelling’
    - Infer correlation rather than causality
    - Moderate predictive power
    - Hypothesis testing can be flawed due to unjustified assumptions
    - Problems arise mostly when modelling complex systems (i.e. many predictor, interactions) ⟶ multiplicity of good models
    - (Block-) cross-validated predictive accuracy as ‘new’ standard measure of fit (Stone, 1974; Roberts et al. 2017)

Still usefull?

Breiman, 2001; Shmueli, 2010

The *Rashomon* Effect

- Japanese movie

  Four people, from different vantage points, witness the death of another person. All report the same facts, but their story of what happened differ.

- Translation:
  - Different realisations of Nature (story of what happens, i.e. $f(x)$)
  - Similar error rates/goodness of fit (same facts)

Breiman, 2001

The *Rashomon* Effect

- Example:

  Subset selection in linear regression: 30 variables

  140,000 five-variable subsets in competition

  Many five-variable subsets with RSS within 1.0% of the lowest RSS

  Conclusion 1: $y = 2.1 + 3.8x_1 - 0.6x_8 + 83.2x_{12} - 2.1x_{17} + 3.2x_{22}$

  Conclusion 2: $y = -8.9 + 4.6x_5 + 0.01x_6 + 12.0x_{15} + 17.5x_{21} + 0.2x_{22}$

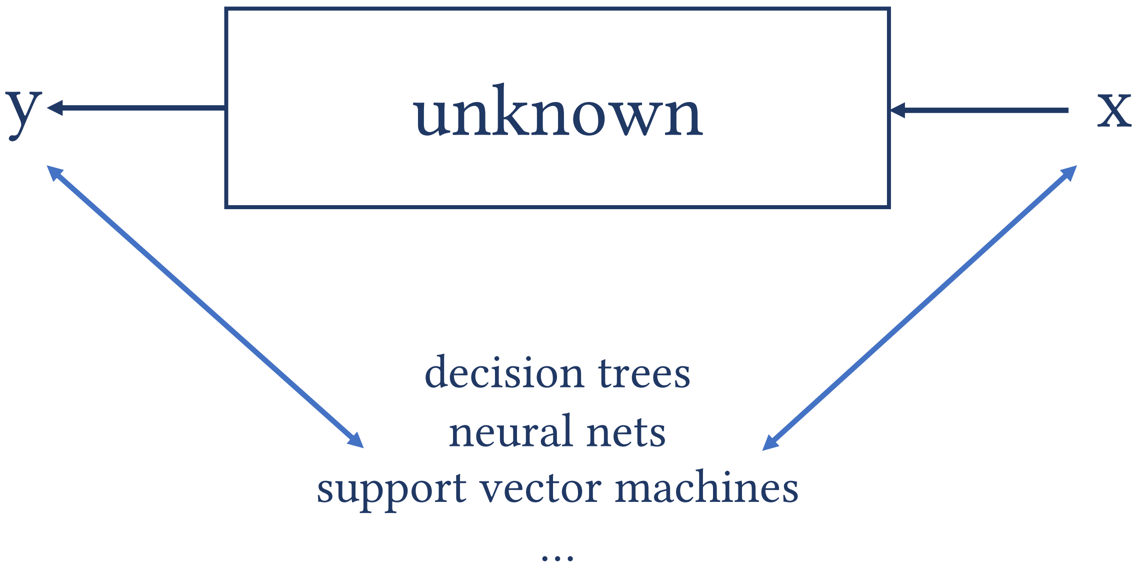
  Conclusion 3: $y = -76.7 + 9.3x_2 + 22.0x_7 - 13.2x_8 + 3.4x_{11} + 7.2x_{28}$

- See Breiman, 1996 for 'instability' in algorithmic models

Breiman, 2001

To predict

- As Breiman puts it: “The algorithmic modelling culture”
- Algorithmic models, (machine learning,) artificial intelligence
- Nature = complex and unknown (black box)
- Goal: finding f(x) (e.g Vapnik 1998; Breiman, 2000)



Breiman, 2001

Predictive modelling in science

- Often considered ‘unscientific’ (see Berk, 2008)
- Not really part of the scientific method
- Rather used in applications (Shmueli, 2010)
- But:
  - Akaike: “The predictive point of view is a prototypical point of view to explain the basic activity of statistical analysis” (in Findley & Parzen, 1998)
  - Deming: “The only useful function of a statistician is to make predictions” (in Wallis, 1980)
- With new large datasets (e.g. ICARUS): Is it possible to ‘control’ the data?
  - More and more problems stop ‘looking like nails’ (Breiman, 2001)

Discussion

- To explain vs to predict?

  Can we use predictions to increase our understanding of a system?
- How useful is explanatory modelling?
- Interpretability of simple stochastic models
- What about model averaging?
- Simplicity vs accuracy
  - It seems that: the more complex the more accurate
  - Should we change our goal from ‘interpretability’ to ‘accurate information’?
- Applications/strengths of algorithmic models in ecology

References

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Breiman, L. (2000). *Some infinity theory for predictor ensembles*. Technical Report 579, Statistics Dept. UCB.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical science*, *16*(3), 199-231.

Findley, D. F., & Parzen, E. (1998). A conversation with Hirotugu Akaike. In *Selected Papers of Hirotugu Akaike* (pp. 3-16). Springer, New York, NY.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., ... & Warton, D. I. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913-929.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44-47.

Shmueli, G. (2010). To explain or to predict? *Statistical science*, *25*(3), 289-310.

Vapnik, V. (1998). *Statistical learning theory. 1998* (Vol. 3). Wiley, New York.

Wallis, W. A. (1980). The statistical research group, 1942–1945. *Journal of the American Statistical Association*, *75*(370), 320-330.

McGill, B. (2014). Are you in science to understand, describe or predict? *Dynamic Ecology Blog*, https://dynamicecology.wordpress.com/2018/03/14/are-you-in-science-to-understand-describe-or-predict/