Model Selection

Martin Sewell
Department of Computer Science
University College London

April 2007 (updated August 2008)

Abstract

An introduction to model selection.

Science is the systematic study of the universe—through observation and experiment—in the pursuit of knowledge that allows us to *generalize*. Although considered bad form in the current climate of political correctness (Lind 1998, 2004; Ellis 2004; Browne 2006; Sewell 2007b), the ability to generalize is a distilled version of what science is all about. Given some data, there will always be an infinite number of models or hypotheses that fit the data equally well and without making further assumptions there is no reason to prefer one model or hypothesis over another. Therefore, one is forced to make assumptions that provide an *inductive bias*.

Model selection is the task of choosing a model with the correct inductive bias, which in practice means selecting parameters in an attempt to create a model of optimal complexity for the given (finite) data. For a good book on model selection, see Burnham and Anderson (2002). Many methods of model selection employ some form of *parsimony*: that is, if they fit the data equally well, they prefer a simpler model (see Zellner, Keuzenkamp and McAleer (2001)). For example, *Occam's razor* advises us that when competing theories have equal predictive power, one should choose the theory that introduces the fewest assumptions. For more details on Occam's razor, see Hoffmann, Minkin and Carpenter (1997) and the references therein. *Bayesians* use *probability* to choose among hypotheses, $P(\text{hypothesis}|\text{data}, \text{background information})$ (Howson and Urbach 1989). *Popperians* choose among hypotheses that are equally consistent with the observations by preferring those which are more *falsifiable* (Popper 1934, 1959). *Likelihoodists* understand the plausibility of a hypothesis in terms of evidential support and they consider $P(\text{data}|\text{hypothesis})$ (Edwards 1992). *Minimum description length* (MDL) (Rissanen 1978) is a technique from algorithmic information theory which dictates that the best hypothesis for a given set of data is the one that leads to the largest compression of the data. One seeks to minimize the sum of the length, in bits, of an effective description of the model and the length, in bits, of an effective description of the data when encoded with the help of the model. *Classical Neyman–Pearson hypothesis testing* considers

P(data|null hypothesis) (the method is flawed, see Atkins and Jarrett (1979); Minka (1998); Gabor (2004); Sewell (2008b)). The *Akaike information criterion* (AIC) (Akaike 1973) proposes that one should trade off the complexity of the model with its goodness of fit to the sample data. The model with the lowest AIC should be preferred. $AIC = -2 \log L + 2k$, where $\log L$ is the maximum log-likelihood and k is the number of parameters.

A taxonomy of model selection:

Empirical

- Adjusted R^2 (Wherry 1931)
- Bootstrap (Efron 1979)
- Cross-validation (Stone 1974; Geisser 1975)
  - Generalized cross-validation (GCV) (Craven and Wahba 1979)
  - K-fold cross-validation
  - Leave-one-out cross-validation
- Jacknife [1]
- Linear regression
- Shibata's model selector (sms) (Shibata 1981)
- Signal-to-noise ratio
- Test set validation

Theoretical

- Akaike information criterion (AIC)
  - AIC (Akaike 1973)
  - AICc (Hurvich and Tsai 1989)
  - QAIC (Lebreton, *et al.* 1992)
  - QAICc (Lebreton, *et al.* 1992)
  - AICW (Wilks 1995)
- CAT (Parzen 1974, 1977)
- Mallows' Cp (Cp) (Mallows 1973)
- Deviance information criterion (DIC) (Spiegelhalter, *et al.* 2002)
- FIC (Wei 1992)
- Final prediction error (FPE) (Akaike 1969)

[1] Richard von Mises was the first to conceive and apply the jackknife.

- FPEα (Bhansali and Downham 1977)
- FPEC (de Luna 1998)
- FPER (Larsen and Hansen 1994)
- GM (Geweke and Meese 1981)
- Generalized prediction error (GPE) (Moody 1991, 1992)
- Hannan and Quinn Criterion (HQ) (Hannan and Quinn 1979)
- KIC (Cavanaugh 1999)
- KICc (Cavanaugh 2004)
- Minimum description length (MDL) (Rissanen 1978)
- Minimum message length (MML) (Wallace and Boulton 1968)
- Predicted squared error (PSE) (Barron 1984)
- PRESS (Allen 1974)
- Schwarz criterion (also Schwarz information criterion (SIC) or Bayesian information criterion (BIC) or Schwarz-Bayesian information criterion) (Schwarz 1978)
- Structural risk minimization (SRM) (Vapnik and Chervonenkis 1974)
- TIC (Takeuchi's information criterion) (Takeuchi 1976)
- VC-dimension (Vapnik and Chervonekis 1968; Vapnik and Chervonenkis 1971; Vapnik 1979)

I am in the process of writing a paper on 'Overfitting and Data Mining' (i.e. model selection) as I have been invited to speak at the *Automated Trading 2008* conference (Sewell 2008a).

Ensemble methods seek to combine models in an optimal way, so are related to model selection, see Sewell (2007a).

References

AKAIKE, Hirotugu, 1969. Fitting Autoregressive Models for Prediction. *Annals of The Institute of Statistical Mathematics*, **21**(2), 243–247.

AKAIKE, Hirotogu, 1973. Information Theory and an Extension of the Maximum Likelihood Principle. *In*: B. N. PETROV and F. CSAKI, eds. *Second International Symposium on Information Theory*. Budapest: Akademiai Kiado, pp. 267–281.

ALLEN, David M., 1974. The Relationship Between Variable Selection and Data Agumentation and a Method for Prediction. *Technometrics*, **16**(1), 125–127.

ATKINS, Liz, and David JARRETT, 1979. The significance of 'significance tests'. *In*: John IRVINE, Ian MILES, and Jeff EVANS, eds. *Demystifying Social Statistics*. London: Pluto Press, Chapter 8, pp. 87–109.

BARRON, Andrew R., 1984. Predicted Squared Error: A Criterion for Automatic Model Selection. *In*: Stanley J. FARLOW, ed. *Self-Organizing Methods in Modeling: GMDH Type Algorithms*, Volume 54. New York: Marcel Dekker, Chapter 4, pp. 87–103.

BHANSALI, R. J., and D. Y. DOWNHAM, 1977. Some Properties of the Order of an Autoregressive Model Selected by a Generalization of Akaike's EPF Criterion. *Biometrika*, **64**(3), 547–551.

BROWNE, Anthony, 2006. *The Retreat of Reason: Political Correctness and the Corruption of Public Debate in Modern Britain*. Second ed. London: Civitas.

BURNHAM, Kenneth P., and David R. ANDERSON, 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Second ed. New York: Springer-Verlag.

CAVANAUGH, Joseph E., 1999. A Large-Sample Model Selection Criterion Based on Kullback's Symmetric Divergence. *Statistics & Probability Letters*, **42**(4), 333–343.

CAVANAUGH, Joseph E., 2004. Criteria for Linear Model Selection Based on Kullback's Symmetric Divergence. *Australian & New Zealand Journal of Statistics*, **46**(2), 257–274.

CRAVEN, Peter, and Grace WAHBA, 1979. Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation. *Numerische Mathematik*, **31**(4), 377–403.

de Luna, Xavier, 1998. An Improvement of Akaike's PFE Criterion to Reduce its Variability. *Journal of Time Series Analysis*, **19**(4), 457–471.

EDWARDS, A. W. F., 1992. *Likelihood*. Baltimore, MD: The Johns Hopkins University Press. Originally published Cambridge: Cambridge University Press, 1972.

EFRON, B., 1979. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, **7**(1), 1–26.

ELLIS, Frank, 2004. *Political Correctness and the Theoretical Struggle: From Lenin and Mao to Marcuse and Foucault*. Auckland: Maxim Institute.

GABOR, George, 2004. Classical Statistics: Smoke and Mirrors. Department of Mathematics and Statistics, Dalhousie University, Halifax, NS.

GEISSER, Seymour, 1975. The Predictive Sample Reuse Method with Applications. *Journal of the American Statistical Association*, **70**(350), 320–328.

GEWEKE, John, and Richard MEESE, 1981. Estimating Regression Models of Finite but Unknown Order. *International Economic Review*, **22**(1), 55–70.

HANNAN, E. J., and B. G. QUINN, 1979. The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society. Series B (Methodological)*, **41**(2), 190–195.

HOFFMANN, Roald, Vladimir I. MINKIN, and Barry K. CARPENTER, 1997. Ockham's Razor and Chemistry. *HYLE International Journal for Philosophy of Chemistry*, **3**, 3–28.

HOWSON, Colin, and Peter URBACH, 1989. *Scientific Reasoning: The Bayesian Approach*. Chicago: Open Court.

HURVICH, Clifford M., and Chih-Ling TSAI, 1989. Regression and Time Series Model Selection in Small Samples. *Biometrika*, **76**(2), 297–307.

LARSEN, Jan, and Lars Kai HANSEN, 1994. Generalization Performance of Regularized Neural Network Models. *In*: John VLONTZOS, Jenq-Neng HWANG, and Elizabeth WILSON, eds. *Proceedings of the 1994 IEEE Workshop on Neural Networks for Signal Processing IV*. Piscataway, NJ: IEEE, pp. 42–51.

LEBRETON, Jean-Dominique, *et al.*, 1992. Modeling Survival and Testing Biological Hypotheses Using Marked Animals: A Unified Approach with Case Studies. *Ecological Monographs*, **62**(1), 67–118.

LIND, Bill, 1998. The origins of political correctness. `http://www.academia.org/lectures/lind1.html`. Address to 13th Accuracy in Academia conference, George Washington University, 10 July 1998.

LIND, William S., 2004. *"Political Correctness:" A Short History of an Ideology*. Alexandria, VA: Free Congress Foundation.

MALLOWS, C. L., 1973. Some Comments on C_p. *Technometrics*, **15**(4), 661–675.

MINKA, Thomas P., 1998. Pathologies of Orthodox Statistics. MIT Media Lab note, revised 2001.

MOODY, John E., 1991. Note on Generalization, Regularization and Architecture Selection in Nonlinear Learning Systems. *In*: B. H. JUANG, S. Y. KUNG, and C. A. KAMM, eds. *Neural Networks for Signal Processing: Proceedings of the 1991 IEEE Workshop*. Los Alamitos: IEEE, pp. 1–10.

MOODY, J. E., 1992. The *Effective* Number of Parameters: An Analysis of Generalization and Regularization in Nonlinear Learning Systems. *In*: John E. MOODY, Steve J. HANSON, and Richard P. LIPPMANN, eds. *Advances in Neural Information Processing Systems 4*. San Mateo, CA: Morgan Kaufmann, pp. 847–854.

PARZEN, Emanuel, 1974. Some Recent Advances in Time Series Modeling. *IEEE Transactions on Automatic Control*, **AC-19**(6), 723–730.

PARZEN, E., 1977. Multiple Time Series: Determining the Order of Approximating Autoregressive Schemes. *In*: P. R. KRISHNAIAH, ed. *Multivariate Analysis–IV*. Amsterdam: North Holland, pp. 283–295.

POPPER, Karl, 1934. *Logik der Forschung*. Tübingen: J. C. B. Mohr.

POPPER, Karl, 1959. *The Logic of Scientific Discovery*. London: Hutchinson.

RISSANEN, J., 1978. Modeling by Shortest Data Description. *Automatica*, **14**(5), 465–471.

SCHWARZ, Gideon, 1978. Estimating the Dimension of a Model. *The Annals of Statistics*, **6**(2), 461–464.

SEWELL, Martin, 2007a. Ensemble Learning. Department of Computer Science, University College London, London.

SEWELL, Martin, 2007b. Political correctness. `http://pc.martinsewell.com/`.

SEWELL, Martin, 2008a. Overfitting and data mining. Talk to be presented at the Automated Trading 2008 conference in London on 15 October 2008.

SEWELL, Martin, 2008b. Statistical inference (and what is wrong with classical statistics). *In*: Ted HARDING, Martin SEWELL, and Ray THOMAS, eds. *The Social Construction of Statistics*. London: Pluto Press. Provisional.

SHIBATA, Ritei, 1981. An Optimal Selection of Regression Variables. *Biometrika*, **68**(1), 45–54.

SPIEGELHALTER, David J., *et al.*, 2002. Bayesian Measures of Model Complexity and Fit. *Statistical Methodology*, **64**(4), 583–639.

STONE, M., 1974. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological)*, **36**(2), 111–147.

TAKEUCHI, K., 1976. Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematical Sciences)*, **153**, 12–18. In Japanese.

VAPNIK, V., 1979. *Estimation of Dependences Based on Empirical Data [in Russian]*. Moscow: Nauka.

VAPNIK, V. N., and A. Ja. CHERVONEKIS, 1968. On the uniform convergence of relative frequencies of events to their probabilities. *Doklady Akademii Nauk USSR*, **181**(4).

VAPNIK, V. N., and A. Ya. CHERVONENKIS, 1971. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and Its Applications*, **16**(2), 264–280. Translated by B. Seckler.

VAPNIK, V. N., and A. Ya. CHERVONENKIS, 1974. *Teoriya Raspoznavaniya Obrazov: Statisticheskie Problemy Obucheniya. (Russian) [Theory of Pattern Recognition: Statistical Problems of Learning]*. Moscow: Nauka.

WALLACE, C. S., and D. M. BOULTON, 1968. An Information Measure for Classification. *Computer Journal*, **11**(2), 185–194.

WEI, C. Z., 1992. On Predictive Least Squares Principles. *The Annals of Statistics*, **20**(1), 1–42.

WHERRY, R. J., 1931. A New Formula for Predicting the Shrinkage of the Coefficient of Multiple Correlation. *The Annals of Mathematical Statistics*, **2**(4), 440–457.

WILKS, Daniel S., 1995. *Statistical Methods in the Atmospheric Sciences: An Introduction*. Volume 59 of *International Geophysics*. San Diego, CA: Academic Press.

ZELLNER, Arnold, Hugo A. KEUZENKAMP, and Michael MCALEER, eds., 2001. *Simplicity, Inference and Modelling: Keeping it Sophisticatedly Simple*. Cambridge: Cambridge University Press.

Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset. Model selection is a process that can be applied both across different types of models (e.g. logistic regression, SVM, KNN, etc.) and across models of the same type configured with different model hyperparameters (e.g. different kernels in an SVM). 3. Model selection and evaluationÂ¶. 3.1. Cross-validation: evaluating estimator performance. 3.1.1. Computing cross-validated metrics.Â 3.1.4. Cross validation and model selection. 3.2. Tuning the hyper-parameters of an estimator. 3.2.1. Exhaustive Grid Search. Model Selection Strategies. Backward elimination begins with the largest model and eliminates variables one-by-one until we are satisfied that all remaining variables are important to the model. Forward selection starts with no variables included in the model, then it adds in variables according to their importance until no other important variables are found. There is no guarantee that backward elimination and forward selection will arrive at the same final model. Model Selection TutorialÂ¶. In this tutorial, we are going to look at scores for a variety of Scikit-Learn models and compare them using visual diagnostic tools from Yellowbrick in order to select the best model for our data. The Model Selection TripleÂ¶. Discussions of machine learning are frequently characterized by a singular focus on model selection. Be it logistic regression, random forests, Bayesian methods, or artificial neural networks, machine learning practitioners are often quick to express their preference.