Title: Spike-and-slab regression

URL: https://en.wikipedia.org/wiki/Spike-and-slab_regression

PageID: 50227596

Categories: Category:Bayesian inference, Category:Bayesian statistics, Category:Machine learning

-----

Spike-and-slab regression is a type of Bayesian linear regression in which a particular hierarchical prior distribution for the regression coefficients is chosen such that only a subset of the possible regressors is retained. The technique is particularly useful when the number of possible predictors is larger than the number of observations. [1] The idea of the spike-and-slab model was originally proposed by Mitchell & Beauchamp (1988). [2] The approach was further significantly developed by Madigan & Raftery (1994) [3] and George & McCulloch (1997). [4] A recent and important contribution to this literature is Ishwaran & Rao (2005). [5]

Model description

Suppose we have P possible predictors in some model. Vector γ has a length equal to P and consists of zeros and ones. This vector indicates whether a particular variable is included in the regression or not. If no specific prior information on initial inclusion probabilities of particular variables is available, a Bernoulli prior distribution is a common default choice. [6] Conditional on a predictor being in the regression, we identify a prior distribution for the model coefficient, which corresponds to that variable (β). A common choice on that step is to use a normal prior with a mean equal to zero and a large variance calculated based on (X T X) – 1 {\displaystyle (X^{T}X)^{-1}} (where X {\displaystyle X} is a design matrix of explanatory variables of the model). [7]

A draw of γ from its prior distribution is a list of the variables included in the regression. Conditional on this set of selected variables, we take a draw from the prior distribution of the regression coefficients (if γ i = 1 then β i $\neq$ 0 and if γ i = 0 then β i = 0). $\beta\gamma$ denotes the subset of β for which γ i = 1. In the next step, we calculate a posterior probability for both inclusion and coefficients by applying a standard statistical procedure. [8] All steps of the described algorithm are repeated thousands of times using the Markov chain Monte Carlo (MCMC) technique. As a result, we obtain a posterior distribution of γ (variable inclusion in the model), β (regression coefficient values) and the corresponding prediction of y .

The model got its name (spike-and-slab) due to the shape of the two prior distributions. The "spike" is the probability of a particular coefficient in the model to be zero. The "slab" is the prior distribution for the regression coefficient values.

An advantage of Bayesian variable selection techniques is that they are able to make use of prior knowledge about the model. In the absence of such knowledge, some reasonable default values can be used; to quote Scott and Varian (2013): "For the analyst who prefers simplicity at the cost of some reasonable assumptions, useful prior information can be reduced to an expected model size, an expected R 2 , and a sample size ν determining the weight given to the guess at R 2 ." [6] Some researchers suggest the following default values: R 2 = 0.5, ν = 0.01, and π = 0.5 (parameter of a prior Bernoulli distribution). [6]

See also

Bayesian model averaging

Bayesian structural time series

Lasso

References

Further reading

Congdon, Peter D. (2020). "Regression Techniques using Hierarchical Priors". Bayesian Hierarchical Models (2nd ed.). Boca Raton: CRC Press. pp. 253– 315. ISBN 978-1-03-217715-1 .