Strijov V.V., Krymova E.A., Weber G.W. Evidence Optimization for Consequently Generated Models // Proceedings of the fourth global conference on power control and optimization, 2010, 1337 : 204-208.

Computers and mathematics with applications, Elsevier 2010

NONLINEAR REGRESSION MODEL GENERATION USING HYPERPARAMETERS OPTIMIZATION

Vadim Strijov, Gerhard Wilhelm Weber

Computing Center of the Russian Academy of Sciences,
Institute of Applied Mathematics of the Middle East Technical University
strijov@ccas.ru, gweber@metu.edu.tr

Abstract

The problem of the non-linear regression analysis is considered. The algorithm of the inductive model generation is described. The regression model is a superposition of given smooth functions. To estimate the model parameters, two-level Bayesian Inference technique was used. It introduces hyperparameters, which describe the distribution function of the model parameters.

Keywords: regression, coherent Bayesian inference, hyperparameters, model generation, model selection

1. Introduction

The inductive model generation algorithms invoke the problem how to estimate the importance of model elements. C. Bishop suggested a method [1–4] of evaluation for the model parameters' probability distribution function. The parameters of these functions are called hyperparameters. For each element of the model one must to estimate the probability distribution function and make a decision on whether particular element of the regression model is important or not.

The problem of the model comparison using hyperparameters was advanced after papers by D. MacKay and I. Nabney. The papers [5–8] and [9] investigate hyperparameter optimization algorithms.

In this paper an inductive model generation algorithm is described. It consists of the following steps. The data set, namely the values of several independent variables and one dependent variable are given. The set of terminal functions is given. The model parameters and hyperparameters are tuned with an optimization algorithm. For each model, the importance of superposition elements is evaluated. The importance depends on the values of the hyperparameters.

The proposed algorithm of the model generation runs iteratively in the following steps. Several best generated models are selected according to a target function. The selected models are modified and new models are generated according to generation rules.

The hyperparameter values bring the information how to modify the models to improve them.

2. Problem statement

A sample set $D=\{(\mathbf{x}_n,y_n)\}_{n=1}^{N}$ of the independent variables $\mathbf{x}\in\mathbf{R}^P$ and the corresponding depended variable $y\in\mathbf{R}$ is given. A set $G=\{g\,|\,g:\mathbf{R}\times\ldots\times\mathbf{R}\to\mathbf{R}\}$ of the smooth parametric functions $g=g(\mathbf{b},\cdot,\ldots,\cdot)$ is given. Here $\mathbf{b}$ is a vector of the parameters.

The set G inductively defines a set of superpositions $F=\{f_i\}$, where a superposition is $f_i=g_{1_i}\circ\ldots\circ g_{r_i}$ consists of no more than r functions g. The set G is called the set of primitive functions.

The superposition f defines a parametric regression model $f=f(\mathbf{w},\mathbf{x})$. The vector $\mathbf{w}=\mathbf{b}_1\vdots\mathbf{b}_2\vdots\ldots\vdots\mathbf{b}_r$, $\mathbf{w}\in\mathbf{R}^W$ is a concatenation of the parameters of the functions $g_1,\ldots,g_r$.

One must to select a model f_i from the set F so that the model f_i minimizes the given target function $p(\mathbf{w}\,|\,D,\alpha,\beta,f_i)$. This function depends on the sample set D, the model $f_i=f_i(\mathbf{w},\mathbf{x})$ and additional parameters α,β. The shape of the target function is defined by hypotheses on the sample set distribution.

The target function is defined as following. Let ν be a random variable of the regression problem $y=f_i(\mathbf{w},\mathbf{x})+\nu$. It has the normal distribution of the variance σ_ν and expectation of zero. Then, according to the maximum likelihood method, the target function is

$$p(D\,|\,\mathbf{w},\alpha,\beta,f_i)=\frac{\exp(-\beta E_D(D\,|\,\mathbf{w},f_i))}{Z_D(\beta)},$$

where $\beta=\sigma_\nu^{-2}$ and $Z_D(\beta)$ is the normalizing constant. The error function $E_D=\sum_{n=1}^{N}(f_i(\mathbf{w},\mathbf{x})-y_n)^2$ is the sum of squares of residuals.

The model parameters $\mathbf{w}^{\mathrm{MP}}$, which brings the maximum to the target function are called the most probable parameters.

3. Inductive model generation

The models are generated with the set of the primitive functions G as following. The indices *v* of the functions g_v are in the set $\mathrm{V} = \{1,...,V\}$. The mapping $\iota : \mathrm{V}^r \to \mathrm{A}$ is given. The elements $A_\iota \in \mathrm{A}$ are the every admissible combinations of K from *V*, where $K = 1,...,r$. The elements of the set $A_\iota = \{a_\iota(k)\}$ have the indexes $k = 1,...,K_\iota$. Since $a \in \mathrm{V}$, the elements $a_\iota(k)$ correspond to the functions $g_v \in G$. For each A_ι consider the set of the incidence matrices $\rho_i(A_\iota)$, $i \in \mathbf{N}$. The index i of the matrix ρ defines a unique superposition f_i of the functions g; denoted as $\rho_i = \rho_i(A_\iota)$. The number of the elements of this superposition equals K_ι. The incidence matrix $\rho_i : \{1,...,K_\iota\}^2 \to \{0,1\}$ defines the oriented graph and the superposition f_i. The superposition is called acceptable if the following conditions are held.

1. The oriented graph ρ_i is acyclic.
2. The oriented graph is one-connected, subject to $\sum_{l=1}^{K_\iota}\sum_{k=1}^{K_\iota}\rho_i(l,k) = \sum_{k=1}^{K_\iota} s(a_\iota(k)),$ where $s = s(v)\sqrt{a^2+b^2}$ is the number of arguments of the function g_v. The number of ones in the matrix ρ_i equals the overall number of arguments of the superposition f_i.
3. The number of arguments of every element of the superposition is equal to the number of arguments of the corresponded primitive function $\sum_{l=1}^{K_\iota}\rho_i(l,k) = s(a_\iota(k))$ for each $k = 1,...,K_\iota$. The number of oriented graph's vertices adjoined to the k-th node is the number $s(a_\iota(k))$ of arguments of the function g_v, where $v = a_\iota(k)$.

4. Estimation of the model hyperparameters

Consider the set of the competitive models $f_1,...,f_M$. When the data D have come, the posterior probability $P(f_i \mid D)$ of the model could be defined with the Bayes theorem

$$P(f_i \mid D) = \frac{p(D \mid f_i)P(f_i)}{\sum_{j=1}^{M} p(D \mid f_j)P(f_j)},$$

where $p(D \mid f_i)$ are predictions, which model can make about the data :

$$p(D \mid f_i) = \int p(D \mid \mathbf{w}, f_i)p(\mathbf{w} \mid f_i)d\mathbf{w}.$$

It is called the evidence of the model.

The posterior probability of the parameters $\mathbf{w}$ of the model f_i, given sample set D, equals

$$p(\mathbf{w} \mid D, f_i) = \frac{p(D \mid \mathbf{w}, f_i)p(\mathbf{w} \mid f_i)}{p(D \mid f_i)}, \qquad (*)$$

where $p(\mathbf{w} \mid f_i)$ is the prior probability of the parameters of the initial distribution. And $p(D \mid \mathbf{w}, f_i)$ is the likelihood function of the model parameters.

Introduce the regularization parameter α. It controls how well the model fits the data. The probability of the parameters given hyperparameter α equals

$$p(\mathbf{w} \mid \alpha, f_i) = \frac{\exp(-\alpha E_W(\mathbf{w} \mid f_i))}{Z_W(\alpha)},$$

where α corresponds to the inverse variance of parameters $\alpha = \sigma_W^{-2}$. And Z_W is the normalizing constant. The requirements to small parameter values suggest the normal posterior distribution with zero-mean. For the given values of the hyperparameters α and β the equation (*) for the given model f_i will be

$$p(\mathbf{w} \mid D, \alpha, \beta, f_i) = \frac{p(D \mid \mathbf{w}, \beta, f_i)p(\mathbf{w} \mid \alpha, f_i)}{p(D \mid \alpha, \beta, f_i)} =$$

$$\frac{\exp(-S(\mathbf{w} \mid f_i))}{Z_S(\alpha, \beta)}, \text{ where } S(\mathbf{w} \mid f_i) = \alpha E_W + \beta E_D$$

and Z_S is the normalizing constant. Consider an iterative algorithm to estimate the optimal values of the parameters $\mathbf{w}$ and the hyperparameters α, β, given model f_i. One must to find the values of the hyperparameters, which bring maximum to the posterior probability of the parameters and then execute the other calculations include probability of the parameters given data with fixed values of the hyperparameters.

To specify the posterior probability $p(D \mid \mathbf{w}, \alpha, \beta)$, which uses the posterior distribution of parameters, one must to approximate the error function $S(\mathbf{w})$ with the second degree Taylor series:

$$S(\mathbf{w}) \approx S(\mathbf{w}^{\mathrm{MP}}) + 2^{-1}(\mathbf{w} - \mathbf{w}^{\mathrm{MP}})A(\mathbf{w} - \mathbf{w}^{\mathrm{MP}}),$$

where the Hessian matrix $A = \nabla^2 S = \beta\nabla^2 E_D + \alpha I$. Substitute the approximate value of $S(\mathbf{w})$ into (*) and denote $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}^{\mathrm{MP}}$, obtain

$$p(\mathbf{w} \mid D, \alpha, \beta) = Z_S^{-1} \exp(-S(\mathbf{w}) - 2^{-1}\Delta\mathbf{w}^T A \Delta\mathbf{w}).$$

Evaluate the constant Z_S, which contain the hyperparameters. To estimate the hyperparameters one must to optimize the function $p(D \mid \alpha, \beta)$ subject to α and β: $\ln p(D \mid \alpha, \beta) = -E_W^{\mathrm{MP}} - 2^{-1}\sum_{j=1}^{W}(\lambda_j + \alpha)^{-1} + W2^{-1}\alpha^{-1}$. Set the last statement equal zero and transform it. The statement for evaluation α is $2\alpha E_W^{\mathrm{MP}} = W - \sum_{j=1}^{W}\alpha(\lambda_j + \alpha)^{-1}$. Denote the summand of the right part as $\gamma = \sum_{j=1}^{W}\alpha(\lambda_j + \alpha)^{-1}$. Then the optimal value of β is given by $2\beta E_D^{\mathrm{MP}} = N - \gamma$.

5. Model generation using hyperparameters

The algorithm of regression models inductive generation runs iteratively. It involves the generated models and the set of the primitive functions. The set of the measured data D and the set of the smooth functions G are given. The initial set of the competitive models $\{f_i \mid i = 1,\dots,M\}$ is given. Each model f_i in the set is a superposition of the functions $g_{ij}, j = 1,\dots,r_i \le r$. The hyperparameter α_{ij} corresponds to the element g_{ij} of the model f_i. It describes the initial probability distribution of the parameter vector $\mathbf{b}_{ij}$ of this function. The hyperparameter β_i corresponds to the model f_i. The initial values of the hyperparameter for *i*-th model are predefined according to the prior noise probability distribution function parameters. After the algorithm starts the following sequence of steps is executed. The sequence repeats the given number of iterations.

1. Minimize the error functions $S_i(\mathbf{w})$ for each model f_i with the Levenberg–Marquardt method. Estimate the parameters $\mathbf{w}_i^{\mathrm{MP}}$ of the models.

2. Define new values of the hyperparameters $\alpha_{ij}^* = (W - \gamma_i)E_W^{-1}(\mathbf{b}_{ij})$, $\beta_i^* = (N - \gamma_i)E_D^{-1}(f_i)$. They based on the initial values of the hyperparameters. Repeat the steps 1 and 2 until the parameters will be converged.

3. According to the error function values select $2^{-1}M$ best models to the further modification. Modify each model: find the element of the superposition with minimal value of the hyperparameters α_{ij}^*; replace it for the arbitrary primitive function $g \in G$.

4. Use the selected and the modified models in the next iterations.

6. Conclusion

The method of the inductive generation of the parametric regression models is described. The models are superpositions of given primitive functions. The model generation algorithm uses hyperparameters of the models. The hyperparameters are estimations of the model parameters' distribution. They show the importance of the models elements. The parameters and the hyperparameters are estimated with non-linear optimization methods.

The suggested method in intended for the regression models construction. These models of optimal complexity fit measured data and could be interpreted by experts in the field of application. As a practical example, the model on the European options volatility smile was created.

7. References

[1] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, Berlin, 2006.

[2] I. Ulusoy, C. M. Bishop, Generative versus Discriminative Methods for Object Recognition, Computer Vision and Pattern Recognition, IEEE Computer Society Conference, vol. 2, 2005, pp. 258–265.

[3] C. M. Bishop, M. E. Tipping, Bayesian Regression and Classification / J. Suykens, G. Horvath, et. al., eds., Advances in Learning Theory: Methods, Models and Applications, vol. 190, IOS Press, NATO Science Series III: Computer and Systems Sciences, 2000, pp. 267–285.

[4] C. M. Bishop, Neural networks and Machine Learning. Springer, Berlin, 1997.

[5] D. J. C. Mackay, Comparison of Approximate Methods for Handling Hyperparameters, Neural Computation, 2003, vol. 11, pp. 1035–1068.

[6] D. J. C. MacKay, Information, Inference, Learning Algorithms, Cambridge University Press, Cambridge, 2003.

[7] D. J. C. MacKay, Bayesian Interpolation / Neural Computation, vol. 4(3), 1992, pp. 415–447.

[8] D. J. C. MacKay, Choice of Basis for Laplace Approximation, Machine Learning, 1998.

[9] I. T. Nabney, NETLAB: Algorithms for Pattern Recognition, Springer, Berlin, 2004.