The search of a parametric regression model in an inductive-generated set*

Strijov, V.
Computing Center of the Russian Academy of Sciences
e-mail: strijov@ccas.ru

The algorithm of the search for the optimal parametric regression model in a model set is described. The model set is a set of the superpositions of given smooth functions. To estimate the probability of the parameters two-level Bayesian Inference technique was used. To illustrate the approach a problem of modelling a pressure in a spray chamber of a combustion engine is described.

1. Introduction

The model search is the iterative "generation-selection" algorithm. The model generation rules and the target function (the model selection criterion) are defined. The series of the competitive models are generated. Each model is a superposition of the elements of a given set of smooth parametric functions. After the model generation each element of superpositions accept its importance hyperparameter. The model parameters and the hyperparameters are tuned in turn. The target function for each model is evaluated. The best generated models are chosen to be modified. The hyperparameter values bring the information how to modify the models to improve them.

The problem of the search for the optimal regression model has a long history, though it remains one of the most actual problems in the field of the pattern recognition. A. G. Ivakhnenko in 1968 created the Group Method of Data Handling, GMDH [1]. According to this method, a model of the optimal complexity is searched in the series of the generated models. For example, to generate the models as the polynomials, Fourier series and the others functional superpositions were used. A. G. Ivakhnenko and his successors created many model generation algorithms and suggested model quality estimation methods.

To generate models one must decide either each element of the model important or not. C. Bishop suggested in [2] the Bayesian regression method. It based on the evaluation the probability distribution function for the model parameters. To do that he introduced the hyperparameters — parameters for probability distribution function of a model parameters. For each element of the model one must to estimate the Gaussian probability distribution function and make a decision either particular element of the regression model important or not.

To modify the models LeCun suggested [3] an optimal brain damage method. To improve a model one must prune the less important elements of the model in case if the approximation quality does not fall significantly. When the elements of the model is pruned one could estimate an impact of those elements to the target function.

*This project is supported by the GFBR, grant № 04-01-00401.

The problem of the model comparison and selection was advanced after papers [4–6] by D. MacKay. He suggests to use two-level Bayesian inference for model selection instead of an information criterion, for example, Akaike Information Criterion. On the first level one computes the probability distribution functions for the parameters of each model from a given set. On the second level the model evidence is computed. According to this method the Occam rule is the following: the probability of the complex model is less than the probability of the simple model, if the values of the target function for these models are the same.

Let us pose the problem on the search of the optimal regression model as following. A sample set $\{\mathbf{x}_1, ..., \mathbf{x}_N | \mathbf{x} \in \mathbb{R}^M\}$ of the independent variables and a set $\{y_1, ..., y_N | y \in \mathbb{R}\}$ of the corresponding depended variables are given. Denote by D the data set $\{(\mathbf{x}_i, y_i)\}$.

A set $G = \{g | g : \mathbb{R} \times ... \times \mathbb{R} \longrightarrow \mathbb{R}\}$ of the smooth parametric functions $g = g(\mathbf{b}, \cdot, \cdot, ..., \cdot)$ are given. The first variable of g is the row vector of parameters $\mathbf{b}$, the following are real-valued variables considered as the independent variables. An arbitrary superposition specifies a parametric repression model $f = f(\mathbf{w}, \mathbf{x})$. Let it includes no more than r functions g. It depends on the independent variables $\mathbf{x}$ and on the parameter vector $\mathbf{w}$. The vector $\mathbf{w} \in \mathbb{R}^W$ consists of the attached parameter vectors of the functions $g_1, ..., g_r$, that is, $\mathbf{w} = \mathbf{b}_1 \vdots \mathbf{b}_2 \vdots ... \vdots \mathbf{b}_r$, where $\vdots$ is the sign of the vector attachment. Denote by $\Phi = \{f_i\}$ the set of the superpositions, which are inductively-generated by elements of the set G.

One must find the model f_i, which brings the maximum to the probability function $p(\mathbf{w}|D, \alpha, \beta, f_i)$. This function will be defined later. It includes the model $f_i = f_i(\mathbf{w}, \mathbf{x})$ and its additional parameters α and β.

2. The regression model choice and target function

The general approach to the comparison of the non-linear models is the following. Consider a set of competitive models $f_1, ..., f_M$. Denote by $P(f_i)$ the prior probability of the model f_i. When the data D have come, the posterior probability $P(f_i|D)$ of the model could be defined with the Bayes theorem,

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

where $P(D|f_i)$ are predictions, which model can make about the data. It is called the evidence of the model. The denominator of the fraction brings satisfaction to the condition $\sum_{i=1}^{M} P(f_i|D) = 1$.

The probabilities of the models f_1 and f_2 the given data, could be compared as

$$\frac{P(f_1|D)}{P(f_2|D)} = \frac{P(f_1)p(D|f_1)}{P(f_2)p(D|f_2)}. \tag{1}$$

The left part $p(D|f_1)/p(D|f_2)$ is the ratio between the evidence of the models. The ratio $P(f_1)/P(f_2)$ is the prior preference between the models. If there is no reason to make different prior probabilities, one could compare the models using $p(D|f_i)$. In the parameter space the evidence $p(D|f_i)$ is

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

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

$$p(\mathbf{w}|D, f_i) = \frac{p(D|\mathbf{w}, f_i)p(\mathbf{w}|f_i)}{p(D|f_i)}, \tag{2}$$

where $p(\mathbf{w}|f_i)$ the prior probability of the parameters, $p(D|\mathbf{w}, f_i)$ is the likelihood function of the model parameters. Denominator $p(D|f_i)$ is needed to satisfy the condition $\int p(\mathbf{w}|D, f_i)d\mathbf{w} = 1$. It is specified by the integral $\int p(\mathbf{w}'|D, f_i)p(\mathbf{w}'|f_i)d\mathbf{w}'$. Equations (2) and (1) are called the first and the second level of the Bayesian inference.

Denote by ν the random variable of the regression model $y = f_i(\mathbf{b}, \mathbf{x}) + \nu$ with additive gaussian noise of variation σ_ν and of zero expectation. Then the likelihood function is

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

where $\beta = \frac{1}{\sigma_\nu^2}$. The denominator $Z_D(\beta)$ is specified by

$$Z_D(\beta) = \left(\frac{2\pi}{\beta}\right)^{\frac{N}{2}}. \tag{3}$$

The weighed error function in the data space is

$$\beta E_D = \frac{\beta}{2}\sum_{n=1}^{N}(f_i(\mathbf{x}_n) - y_n)^2. \tag{4}$$

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

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

where α corresponds variance variance of parameters, $\alpha = \sigma_{\mathbf{w}}^{-2}$ and the normalizing constant Z_W is

$$Z_W(\alpha) = \left(\frac{2\pi}{\alpha}\right)^{\frac{W}{2}}. \tag{5}$$

The requirements to small parameter values [7] suppose the gaussian posterior distribution with zero-mean:

$$p(\mathbf{w}) = \frac{1}{Z_W}\exp\left(-\frac{\alpha}{2}||\mathbf{w}||^2\right).$$

Since the variables α and β are the parameters of distributions of the model parameters further they will be called hyperparameters. Eliminate the normalizing constant Z_W since it does not depend on the parameters $\mathbf{w}$ and evaluate the logarithm, then

$$\alpha E_W = \frac{\alpha}{2}||\mathbf{w}||^2. \tag{6}$$

This error function regularizes the parameters imposing a fine for the large values of the parameters.

For given values of the hyperparameters α and β the equation (2) for the given model f_i will be

$$p(\mathbf{w}|D,\alpha,\beta) = \frac{p(D|\mathbf{w},\beta)p(\mathbf{w}|\alpha)}{p(D|\alpha,\beta)}.$$

Rewrite the error function as $S(\mathbf{w}) = \alpha E_W + \beta E_D$, obtain

$$p(\mathbf{w}|D,\alpha,\beta,f_i) = \frac{\exp(-S(\mathbf{w}|f_i))}{Z_S(\alpha,\beta)}, \tag{7}$$

where Z_S is normalizing factor.

3. Model parameters evaluation

To evaluate the optimal values of the parameters $\mathbf{w}$ and hyperparameters α,β for the given model f_i one must integrate them. The posterior probability is

$$p(\mathbf{w}|D) = \iint p(\mathbf{w},\alpha,\beta|D)d\alpha d\beta = \iint p(\mathbf{w}|\alpha,\beta,D)p(\alpha,\beta|D)d\alpha d\beta. \tag{8}$$

The computational complexity of such kind of integration is very large. However the integral could be simplified if suitable values of the parameters will be assigned. The approximation of the integral is based on the following. The posterior probability of the hyperparameters $p(\alpha,\beta|D)$ has a definite peak around most probably values of the hyperparameters $\alpha^{\text{MP}},\beta^{\text{MP}}$. This approximation is known as the Laplace approximation [8]. Under such assumption the integral (8) is simplified to

$$p(\mathbf{w}|D) \approx p(\mathbf{w}|\alpha^{\text{MP}},\beta^{\text{MP}},D)\iint p(\alpha,\beta|D)d\alpha d\beta \approx p(\mathbf{w}|\alpha^{\text{MP}},\beta^{\text{MP}},D).$$

One must find the values of the hyperparameters, which bring maximum to the posterior probability of the parameters and then execute the others calculations include $p(\mathbf{w}|D)$ with fixed values of the hyperparameters.

Tho specify the posterior probability $p(\mathbf{w}|\alpha,\beta,D)$ one must approximate error function $S(\mathbf{w})$ with the second degree Taylor series:

$$S(\mathbf{w}) \approx S(\mathbf{w}^{\text{MP}}) + \frac{1}{2}(\mathbf{w}-\mathbf{w}^{\text{MP}})^T A(\mathbf{w}-\mathbf{w}^{\text{MP}}). \tag{9}$$

In (9) there is no first degree term, since it is supposed that $\mathbf{w}^{\text{MP}}$ defines the local minimum of the error function:

$$\frac{\partial S(\mathbf{w}^{\text{MP}})}{\partial w_\xi} = 0$$

for each value of ξ. The matrix A is the Hessian matrix. It depends on the error function:

$$A = \nabla^2 S(\mathbf{w}^{\text{MP}}) = \beta\nabla^2 E_D(\mathbf{w}^{\text{MP}}) + \alpha I.$$

Denote by H the first term of the right part of the equation, then $A = H + \alpha I$.

Substitute the approximate value of $S(\mathbf{w})$ into (7) and denote $\Delta\mathbf{w} = \mathbf{w} - \mathbf{w}^{\text{MP}}$. Then

$$p(\mathbf{w}|\alpha,\beta,D) = \frac{1}{\hat{Z}_S}\exp\left(-S(\mathbf{w}^{\text{MP}}) - \frac{1}{2}\Delta\mathbf{w}^T A\Delta\mathbf{w}\right).$$

Evaluate the constant $\hat{Z}_S$ which is necessary for the Laplace approximation:

$$\hat{Z}_S = \exp(-S(\mathbf{w}^{\mathrm{MP}}))(2\pi)^{\frac{W}{2}}(\det A)^{-\frac{1}{2}}. \tag{10}$$

To maximize the function $p(D|\alpha,\beta)$ one has to vary the values of the hyperparameters α and β. To do that one has to integrate the data distribution function over the parameter space $\mathbf{w}$:

$$p(D|\alpha,\beta) = \int p(D|\mathbf{w},\alpha,\beta)p(\mathbf{w}|\alpha,\beta)d\mathbf{w} = \int p(D|\mathbf{w},\alpha,\beta)p(\mathbf{w}|\alpha)d\mathbf{w}, \tag{11}$$

where the second integral is fare because the model parameters does not depend on the noise. To simplify the computations assume that $p(\alpha,\beta)$ are distributed uniformly.

Using (4), (6), write (11) as

$$p(D|\alpha,\beta) = \frac{1}{Z_D(\beta)}\frac{1}{Z_D(\alpha)}\int \exp(-S(\mathbf{w})))d\mathbf{w}.$$

From (3), (5), (10) and the previous statement it comes

$$\ln p(D|\alpha,\beta) = -\alpha E_W^{\mathrm{MP}} - \beta E_D^{\mathrm{MP}} - \frac{1}{2}\ln|A| + \frac{W}{2}\ln\alpha + \frac{N}{2}\ln\beta - \frac{N}{2}\ln(2\pi). \tag{12}$$

To optimize this statement with variable α, one has to evaluate the derivative

$$\frac{d}{d\alpha}\ln|A| = \frac{d}{d\alpha}\ln\left(\prod_{j=1}^{W}\lambda_j + \alpha\right) = \frac{d}{d\alpha}\sum_{j=1}^{W}\ln(\lambda_j+\alpha) = \sum_{j=1}^{W}\frac{1}{\lambda_j+\alpha} = \mathrm{tr}(A^{-1}). \tag{13}$$

In this statement $\lambda_1, ..., \lambda_W$ are eigenvalues of the matrix H. Since the error function is not a quadratic function of the parameters (unlikely the linear of RBF regression) it is impossible to evaluate the optimal value of α directly. The Hessian matrix is not a constant, it depends on the parameters $\mathbf{w}$. Since we accept $A = H + \alpha I$ for the vector $\mathbf{w}^{\mathrm{MP}}$, which depends on α, then the eigenvalues of H indirectly depend on α. Thus the statement (13) does not involve the parameters of the model.

Using this approximation the derivative (12) of α equals

$$\ln p(D|\alpha,\beta) = -E_W^{\mathrm{MP}} - \frac{1}{2}\sum_{j=1}^{W}\frac{1}{\lambda_j+\alpha} + \frac{W}{2\alpha}.$$

Set the last statement equal zero and transform it. So one has the statement for evaluation α

$$2\alpha E_W^{\mathrm{MP}} = W - \sum_{j=1}^{W}\frac{\alpha}{\lambda_j+\alpha}. \tag{14}$$

Denote the deduction of the right part as γ then

$$\gamma = \sum_{j=1}^{W}\frac{\alpha}{\lambda_j+\alpha}.$$

Those components λ_j for which $\alpha \ll \lambda_j$ impacts γ more than the components for which $0 < \lambda_j \ll \alpha$. Thus γ could be interpreted of the measure of the number of well-defined parameters of the model.

To evaluate hyperparameter β consider the optimization problem (12). Denote μ_j the eigenvalues of the matrix $\nabla^2 E_D$. Since $H = \beta\nabla^2 E_D$, then $\lambda_j = \beta\mu_j$ and

$$\frac{d\lambda_j}{d\beta} = \mu_j = \frac{\lambda_j}{\beta}.$$

So,

$$\frac{d}{d\beta}\ln|A| = \frac{d}{d\beta}\sum_{j=1}^{W}\ln(\lambda_j + \alpha) = \frac{1}{\beta}\sum_{j=1}^{W}\frac{\lambda_j}{\lambda_j + \alpha}.$$

Evaluate the derivative as in the previous case of α, one defines the optimal value β as

$$2\beta E_D^{\text{MP}} = N - \sum_{j=1}^{W}\frac{\lambda_j}{\lambda_j + \alpha} = N - \gamma. \tag{15}$$

Further the optimal values of the hyperparameters α and β evaluation and usage is described.

4. The algorithm of the search for the optimal model

The search for the optimal model is executed iteratively. Before it starts, the set of the measured data D and the set G of the basic functions g are given. The initial set of the competitive models $F_0 = \{f_1, ..., f_M | f \in \Phi\}$ are given. Each model f_i in the set is a superposition of the functions $\{g_{ij}\}_{j=1}^{r_i}$. The hyperparameter α_{ij} corresponds the element g_{ij} of the model f_i. The hyperparameter β_i corresponds to the model f_i. The initial values of the hyperparameters are predefined. After the algorithm starts the following sequence of steps is executed.

1. Evaluate the model parameters $\mathbf{w}_i^{\text{MP}}$. To do that the method of the scaled conjugate gradients [9] minimizes the error function $S_i(\mathbf{w})$ for each model $f_i, i = 1, ..., M$.

2. Using (14) and (15) evaluate the hyperparameters α_{ij}^{new} and β_i^{new}. The hyperparameter β_i of the function f_i is evaluated in the data space. It eqials

$$\beta_i^{\text{new}} = \frac{N - \gamma_i}{E_D(f_i)}.$$

The hyperparameter α_{ij} is evaluated in the parameter space for each function g_{ij} in the superposition f_i. It equals

$$\alpha_{ij}^{\text{new}} = \frac{W - \gamma_i}{E_W(\mathbf{b}_{ij})}.$$

The evaluation of the parameters and the hyperparameters are repeated until the local minimum of $S_i(\mathbf{w})$ will be found.

3. Generate the successor models $f_1', ..., f_M'$ using the following rules. Repeat for each index i. Chose a function g_{ij} with the minimal value of α_{ij} in the model f_i. Chose a random function $g_{\xi k}$ in the random model f_ξ. Generate the successor f_i' by replacing g_{ij} for $g_{\xi k}$.

4. Each model f_i' could be modified. The probability η of the modification is given. For each model f_i replace the function g_{ij} with the minimal value of α_{ij} by the random function from the set G.

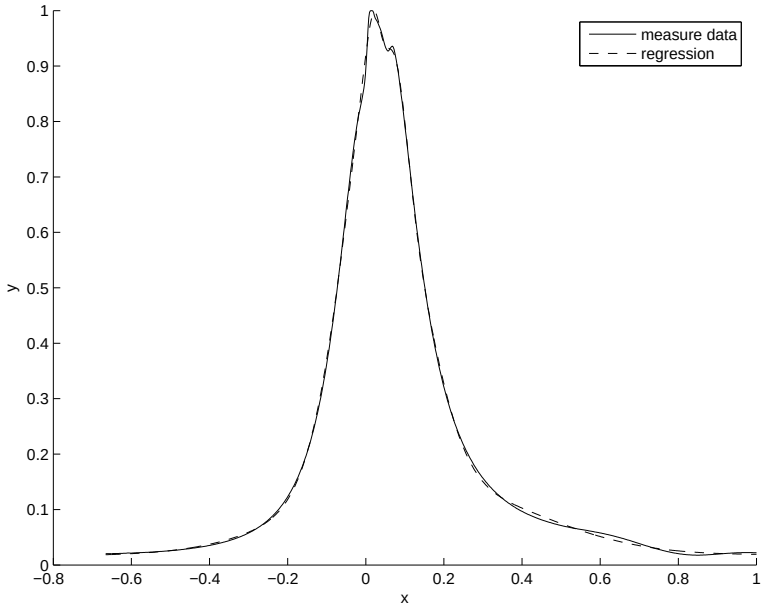


Fig. 1. The source sample and the sample approximated with the model № 2

5. Join the successor and predecessor models in one set and select M best models according to the error function S. These models will be used in the further iterations.

The algorithm stops either when the target function reach a given value or when the sequence repeats a given number of times.

5. The numerical experiment

The proposed algorithm was used to solve the following application problem. The pressure in the combusting camera of a diesel engine was measured during series of the pressure cycles. The results of the measurement was represented as a sample set. The source date are represented at the fig. 1 by the solid line. The dashed line shows the recovered data. The x-axis shows the independent variable, the y-axis shows the dependent variable. The data contain 4000 samples. To verify the obtained model 118 pressure curves were used.

Experts made the set G which were used to generate models. The functions from the set are listed in the table 1. The set F_0 of the initial models was also given by the experts.

Several thousands of the models were generated during the numerical experiment. Three models were selected. They are shown in the table 2. The quality of the models were evaluated with the error functions ρ_1, ρ_2. The model complexity is the number of parameters in the vector $\mathbf{w}$. The values of the errors are obtained by the average for 118 pressure curves. The error function ρ_1 is mean squared error

$$\rho_1 = \sqrt{\frac{1}{N}\sum_{i=1}^{n}\left(\frac{y_i - f(\mathbf{x}_i)}{\max(y_i)}\right)^2},$$

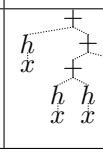
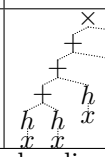
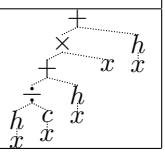
the error ρ_2 is maximal relative error

$$\rho_2 = \max_{i=1,\ldots,N}\frac{|y_i - f(\mathbf{x}_i)|}{\max(y_i)}.$$

№	Function	Description	Parameters
Function of two variables, $g(\mathbf{b}, x_1, x_2)$			
1	plus	$y = x_1 + x_2$	–
2	times	$y = x_1 x_2$	–
3	divide	$y = x_1 / x_2$	–
Function of one variable, $g(\mathbf{b}, x_1)$			
4	multiply	$y = ax$	a
5	add	$y = x + a$	a
6	gaussian	$y = \frac{\lambda}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\xi)^2}{2\sigma^2}\right) + a$	λ, σ, ξ, a
7	linear	$y = ax + b$	a, b
8	parabolic	$y = ax^2 + bx + c$	a, b, c
9	cubic	$y = ax^3 + bx^2 + cx + d$	a, b, c, d
10	logsig	$y = \frac{\lambda}{1+\exp(-\sigma(x-\xi))} + a$	λ, σ, ξ, a

T a b l e 1. The set G of the basic functions

The row "Description" of the table 2 shown the tree-like model structure. Let the second model be an example. This model is the superposition of eight functions $f_2 = g_1(g_2(g_3(g_4(g_5(x), g_6(x)), g_7(x)), x), g_8(x))$. The functions $g_1 = \times(\varnothing, \cdot, \cdot)$ and $g_2, ..., g_4 = +(\varnothing, \cdot, \cdot)$ are addition ane multiplication. They have the empty parameter vector as the first variable. The functions $g_5, ..., g_7 = h(\mathbf{b}_i, \cdot)$, $i = 1, ..., 3$ are the gaussian function; and $g_8 = l(\mathbf{b}_4, \cdot)$ is the linear function. The function $h = \frac{\lambda_i}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(x-\xi_i)^2}{2\sigma_i^2}\right)$ have the parameters vector $\mathbf{b}_i = \langle \lambda_i, \mu_i, \sigma_i, a_i \rangle$, and the function $l = (ax + b)$ has the parameter vector $\mathbf{b}_4 = \langle a, b \rangle$.

Model	1	2	3
Error ρ_1	0.0034	0.0037	0.0035
Error ρ_2	0.0421	0.0325	0.00338
Num. of parameters	16	16	16
Description			

Legend: h — gaussian, c — cubic, l — linear,
+ — plus, × — times,÷ — divide

T a b l e 2. Description of the selected models

The model f_2 could be represented as $f(\mathbf{w}, \mathbf{x}) = l(\mathbf{b}_4, x)^{-1} \times \left(x + \sum_{i=1}^{3} h(\mathbf{b}_i, x)\right)$, where $\mathbf{x} = x$ и $\mathbf{w} = \mathbf{b}_1 \vdots \mathbf{b}_2 \vdots \mathbf{b}_3 \vdots \mathbf{b}_4$. The full representation of the model is

$$y = (ax + b)^{-1}\left(x + \sum_{i=1}^{3} \frac{\lambda_i}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(x-\xi_i)^2}{2\sigma_i^2}\right) + a_i\right).$$

The model f_2 was used by experts for their applications: forecasting and analysis of the oxygen concentration at the exhaust manifold of the diesel engine.

Conclusions

The universal regression models (for example the neural networks or the radial basis functions) often have a big number of parameters and turn out overtrained. To construct simple and precise

models the problem of the search of a parametric regression model in an inductive-generated set was posed.

To choose a model of the optimal complexity from the set of the competitive models the two-level Bayesian inference was used. Since the inference integrals are very complex to be computed the approximation procedures were developed.

The algorithm for model generation and selection was suggested. It uses the hyperparameters which correspond the elements of the models. These hyperparameters show the importance of the elements. The algorithm iteratively creates successor models. The models are modified according to the importance criterion and selected according to the target function. The complexity of the models are restricted automatically during the model comparison.

The algorithm was tested on the problem of the pressure curve approximation. The pressure was measures in the combusting camera of the diesel engine. The obtained optimal model is up to the industry standards.

References

[1] *Malada, H.R., Ivakhnenko, A. G.* Inductive Learning Algorithms for Complex Systems Modeling. CRC Press, 1994.

[2] *Bishop, C.M., Tipping, M.E.* Bayesian regression and classification // Suykens, J., Horvath, G. et. al., eds. Advances in Learning Theory: Methods, Models and Applications, Volume 190. IOS Press, NATO Science Series III: Computer and Systems Sciences, 2000. P 267–285.

[3] *LeCun, Y., Denker, J. S., and Solla, S. A.* Optimal brain damage // Touretzky, D.S., ed. Advances in Neural Information Processing Systems. Morgan Kaufmann, San Mateo, CA, 1990. P. 598–605.

[4] *MacKay, D.* Information, inference, learning algorithms. Cambridge University Press, 2003.

[5] *MacKay, D.* Hyperparameters: optimise or integrate out? // Heidberger, G., ed. Maximum entropy and Bayesian Methods. Santa Barbara, Dordrecht: Kluwer, 1993.

[6] *MacKay, D.* Bayesian interpolation // Neural Computation 4(3), 1992. P. 415–447.

[7] *Nabney, I.T.* NETLAB: Algorithms for pattern recognition. Springer, 2004. P. 330.

[8] *MacKay, D.* Choice of basis for Laplace approximation // Machine Learning, vol. 33(1), 1998.

[9] *Branch, M.A., Coleman, T.F., Li, Y.* A Subspace, Interior, and Conjugate Gradient Method for Large-Scale Bound-Constrained Minimization Problems // SIAM Journal on Scientific Computing, vol. 21(1), 1999. P. 1–23.