Regularized Approach to Linear Fitting

Jaromír Kukal, Dana Majerová, Luboš Hainc

ICT Prague, Department of Computing and Control Engineering

Abstract. *The paper is devoted to solving the traditional task of linear regression. It is useful to remove the absolute term from the regression formula or to specify the kernel inputs. The solution is obtained by regularization or using the pseudo–inverse matrices. Four approaches to given tasks are defined and compared. All the algorithms are realized in Matlab system.*

Keywords: *linear regression, ill-posed problem, regularization, bias elimination, kernel, pseudo–inversion, Matlab.*

1 Introduction

The linear regression is a standard tool for the statistical data processing. But the data are sometime ill–posed because of poor planning of experiments or over–fitting effect. It occurs in case of linear neural network learning when the number of training patterns is less than the number of network inputs. The problem of reducing the number of input is also interested in large class of AI tasks. The input signals have various degrees of accessibility and the neural network pruning can be based on complete or partial regularization. Four approaches to this general problem are studied in this paper. All the algebraic calculations were done in the Matlab system. The features of given results are demonstrated on a provocative over–fitting example of two–point parabola fitting.

2 Four Approaches to Linear Data Fitting

Let $m, n \in \mathbf{N}$ be the number of samples and the number of variables. Let

$$\mathbf{X} = \{x_{ij}\}_{1\leq i\leq m, 1\leq j\leq n}$$

$$\overrightarrow{y} = \{y_i\}_{1\leq i\leq m}$$

be the input data matrix and the output data vector. There are four approaches how to solve the data fitting task in sense of least sum of squares technique.

2.1 Classic LSQ Approach

The model also includes the bias a_0 and has the form

$$y_i = a_0 + \sum_{j=1}^{n} a_j x_{ij}.$$

Let $\mathbf{P} = (\overrightarrow{\mathbf{1}}|\mathbf{X})$ be expanded input matrix. Then we solve $\mathbf{P}\overrightarrow{a} = \overrightarrow{y}$ using residuum vector $\overrightarrow{r} = \mathbf{P}\overrightarrow{a} - \overrightarrow{y}$. Finding minimum value of the residuum vector norm $\|\overrightarrow{r}\|$ we obtain classic but danger solution

$$\overrightarrow{a} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\overrightarrow{y}.$$

The solution isn't defined when $det(\mathbf{P}'\mathbf{P}) = 0$.

2.2 Regularized Approach

Using the same model we find the minimum value of the norm $\|\vec{a}\|$ with respect to the constrain $\|\vec{r}\| \leq C$. The regularized solution

$$\vec{a}_\lambda = (\mathbf{P}'\mathbf{P} + \lambda\mathbf{I})^{-1}\mathbf{P}'\vec{y}$$

is defined for all real $\lambda > 0$. The regularized approach has two limit cases $\vec{a}_{0+} = \mathbf{P}^+\vec{y}$ and $\vec{a}_{+\infty} = \vec{0}$ where the upper symbol + denotes the Moore–Penrose pseudo–inversion.

2.3 Bias-less Approach

The main disadvantage of regularization is the convergence of vector $\vec{a}$ to zero vector for large parameter λ. It results the convergence of a_0 to zero, too. In large scale of application there is useful when a_0 converges to the mean value of output y_i. Let

$$\mu_j = \frac{1}{m}\sum_{i=1}^{m} x_{ij},$$

$$\nu = \frac{1}{m}\sum_{i=1}^{m} y_i.$$

The bias–less model has the form

$$y_i - \nu = \sum_{j=1}^{n} b_j(x_{ij} - \mu_j).$$

Defining $q_{ij} = x_{ij} - \mu_j, z_i = y_i - \nu$ we solve $\mathbf{Q}\vec{b} = \vec{z}$ using the residuum vector $\vec{s} = \mathbf{Q}\vec{b} - \vec{z}$. Finding minimum value of the norm $\|\vec{b}\|$ with respect to the constrain $\|\vec{s}\| \leq D$ the another regularized solution is obtained as

$$\vec{b}_\lambda = (\mathbf{Q}'\mathbf{Q} + \lambda\mathbf{I})^{-1}\mathbf{Q}'\vec{z}$$

for all real $\lambda > 0$. There are also two limit cases $\vec{b}_{0+} = \mathbf{Q}^+\vec{z}$ and $\vec{b}_{+\infty} = \vec{0}$. The bias can be additionally recalculated as $b_\lambda^0 = \mu - \vec{\mu} \cdot \vec{b}_\lambda$. Then the bias value comes to the average output value for large values of λ.

2.4 Kernel weight minimization

Let $n_1 \leq n$ be the number of kernel input variables and $n_2 = n + 1 - n_1$ be the number of free input variables. Then the matrix $\mathbf{P}$ can be rearranged in columns and then split to the kernel and free parts as $(\mathbf{F}|\mathbf{G})$. The split model has the form

$$y_i = \sum_{j=1}^{n_1} u_j f_{ij} + \sum_{j=1}^{n_2} v_j g_{ij}.$$

It is necessary to regularize the task $\mathbf{F}\vec{u} + \mathbf{G}\vec{v} = \vec{y}$ using the residuum vector $\vec{w} = \mathbf{F}\vec{u} + \mathbf{G}\vec{v} - \vec{y}$. Finding the minimum of kernel weight norm $\|\vec{u}\|$ with respect to the constrain $\|\vec{w}\| \leq E$ the adequate regularized solution is obtained as

$$\vec{u}_\lambda = \mathbf{H}_\lambda(\vec{y} - \mathbf{G}\vec{v}_\lambda),$$

$$\vec{v}_\lambda = (\mathbf{G}'\mathbf{T}_\lambda\mathbf{G})^+\mathbf{G}'\mathbf{T}_\lambda\vec{y}$$

where $\mathbf{H}_\lambda = (\mathbf{F}'\mathbf{F} + \lambda\mathbf{I})^{-1}\mathbf{F}$ and $\mathbf{T}_\lambda = \mathbf{I} - \mathbf{F}\mathbf{H}_\lambda$ for all real $\lambda > 0$. The limit values for large λ are $\vec{u}_{+\infty} = \vec{0}$ and $\vec{v}_{+\infty} = \mathbf{G}^+\vec{y}$. It means the kernel weights are pruned while the free weights are fitted.

3 Matlab Realization

Four functions were realized in the Matlab system. Using the previous notation they are called as

```
[a]=ClassicLSQ(X,y);
[a]=RegularizedLSQ(X,y,lambda);
[b,b0]=BiasLessLSQ(X,y,lambda);
[u,v]=KernelLSQ(F,G,y,lambda);
```

The functions can be called with `lambda=0`, `lambda=+inf` or without parameter lambda. The four case discussion is realized inside the functions. The functions return zero vectors in case of failure. The first two functions were built only for the demonstration of their poor properties while the remaining two functions are efficient tools for the ill–posed linear regression.

4 Illustrative Example

There is difficult to understand the sense of various approaches to fitting task from their formal description. The four techniques were compared on naive example of fitting parabola using only two points $(-1;+1)$ and $(+1;+2)$. It is typical rank deficient problem with two inputs $x_1 = x, x_2 = x^2$, three parameters a_0, a_1, a_2 but only two equations

$$a_0 - a_1 + a_2 = 1,$$
$$a_0 + a_1 + a_2 = 2.$$

This system of equations has infinite number of solutions. One of them $(a_0, a_1, a_2) = (3/2, 1/2, 0)$ is intuitively preferred and leads to line fitting. Because of various complexity of input signals there are two possibilities how to set the task kernel. The kernel can consist of two inputs $x_1 = x, x_2 = x^2$ or only one input $x_1 = x^2$ in the second case. The input matrices are defined in Matlab notation as

```
X=[-1 +1;+1 +1]; y=[1;2];
F1=[-1 +1;+1 +1]; G1=[+1;+1];
F2=[+1;+1]; G2=[+1 -1;+1 +1];
```

The first classic LSQ approach gives no result because of $det(\mathbf{P}'\mathbf{P}) = 0$.

The second regularized approach gives various results depending on regularization parameter λ. The general solution is

$$(a_0, a_1, a_2) = \left(\frac{3}{\lambda+4}, \frac{1}{\lambda+2}, \frac{3}{\lambda+4}\right).$$

Then

$$y = \frac{3(1+x^2)}{\lambda+4} + \frac{x}{\lambda+2}.$$

This result is against intuition because of $a_0 \leq 3/4 < 3/2$ and $a_2 > 0$.

The third bias–less approach produces

$$(b^0, b_1, b_2) = \left(\frac{3}{2}, \frac{1}{\lambda + 2}, 0\right).$$

Then

$$y = \frac{3}{2} + \frac{x}{\lambda + 2}.$$

This result comes to the intuitive linear solution for $\lambda \to 0+$. When λ is large the fitting comes to constant model $y = 3/2$.

The fourth approach with linear and quadratic column in the kernel gives the result

$$(u_1, u_2, v_1) = \left(\frac{1}{\lambda + 2}, 0, \frac{3}{2}\right)$$

which is the same fitting as in the bias–less approach. Using only the quadratic column in the kernel we have $(u_1, v_1, v_2) = (0, 3/2, 1/2)$ and then

$$y = \frac{3}{2} + \frac{x}{2}$$

for all $\lambda > 0$.

5 Conclusions

Bias elimination or kernel specification is a good preprocessing for finding the regularized solution of ill-posed problems. The new Matlab functions BiasLessLSQ and KernelLSQ are recommended for the linear fitting stabilization. The zero value of parameter lambda corresponding to the condition $\lambda \to 0+$ is recommended for the first experiments with any data set. From the pedagogical point of view, refusing of regularization principle often gives no results and using regularization principle in general form gives poor results. Our study brings the ways how to solve this sad dilemma.

References

[1] *Matlab 5.3.* The MathWorks, Inc. 1999.

[2] Press W.H., Tuekolsky S.A., Vetterling W.T., Flannery B.P. *Numerical Recipes in C, The Art of Scientific Computing.* Cambridge University Press. Cambridge. 1992.

Contact

Jaromír Kukal, Dana Majerová, Luboš Hainc

INSTITUTE OF CHEMICAL TECHNOLOGY, Prague
Department of Computing and Control Engineering
Technická 5, 166 28 Prague 6 Dejvice
Phone: 420-2-2435 4212, fax: 420-2-2435 5053
E-mails: {Jaromir.Kukal, Dana.Majerova, Lubos.Hainc}@vscht.cz
WWW addresses: http://www.vscht.cz/ (university),
http://uprt.vscht.cz/ (department),
http://phobos.vscht.cz/ (research group)