Постановка задач и выбор моделей в машинном обучении

Вадим Викторович Стрижов

Московский физико-технический институт

Осенний семестр 2019

Purpose

The aim of the study is to suggest a method to forecast a structure of a regression model superposition, which approximates a data set in terms of some quality function.

The problem

Algorithms of model selection are computationally complex due to the large number of models.

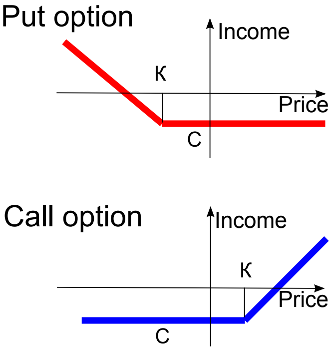
Solution

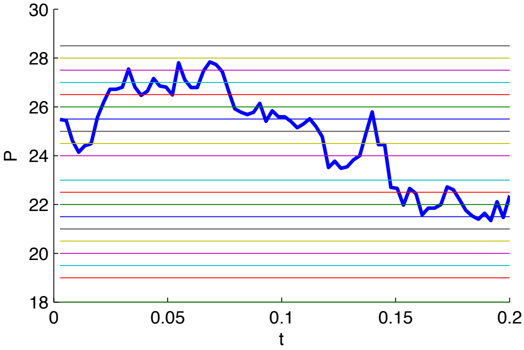
We suggest to build an algorithm of forecasting a model structure based on previously selected models.

Options are financial instruments that convey the right, but not the obligation, to engage in a future transaction on some underlying security.

$$C_t = F(\sigma, S, r, K, t),$$

C_t — option price,
σ — volatility,
S — underlying price,
r — risk-free rate,
K — strike price,
t — time to expiration.

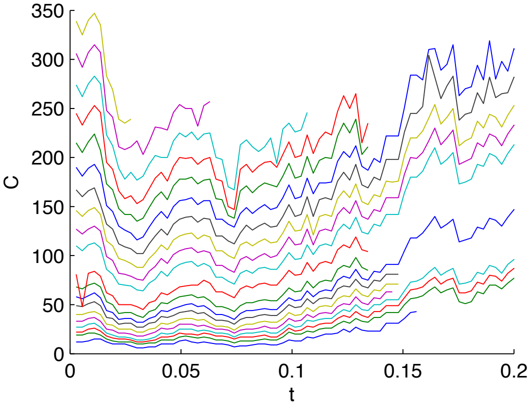




t is the time to expiration,

S is the underlying price.

Horizontal lines correspond to different strike prices K.



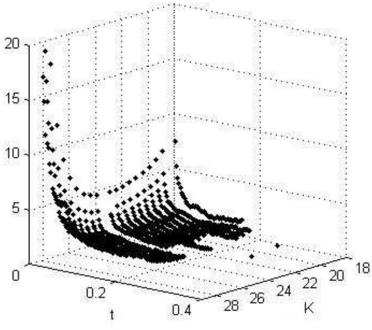
t is the time to expiration,
C is the call option historical price.

1. The regression data is a set
$$\{(\mathbf{x}_n, \sigma_n)\}_{n=1}^{N}, \text{ where } x_n = (t_n, K_n).$$
2. A set of the primitive functions is given $G = \{g_1, \dots, g_v\}$.
3. Superpositions of primitives g define parametric regression models
$$f = f(\mathbf{w}, \mathbf{x}), \qquad f \in F.$$
4. The problem is to select a model f that minimises SSE
$$E_D = \sum_{n=1}^{N} (f(\mathbf{w}, \mathbf{x}_n) - y_n)^2.$$

The implied volatility of an option is the argument minimum of the difference between historical price of the option and its fair price.

$$\sigma^{\text{imp}} = \arg\min_{\sigma}(C_{\text{hist}} - C(\sigma, S, r, K, t)).$$

- σ^{imp} is the dependent variable,
- K and t are the independent variables in the regression model.



t is the time to expiration,
K is the strike price and z-axis σ^{imp} is implied volatility.

The model is given by experts of the Russian Trade System

$$\sigma = \sigma(\mathbf{w}) = w_1 + w_2(1 - \exp(-w_3 x^2)) + \frac{w_4 \arctan(w_5 x)}{w_5},$$

$$\text{where} \quad x = \frac{\log(K) - \log(C(t))}{\sqrt{t}}.$$

Volatility surface modeling rules of thumb [Duglish, 2006]

- The volatility depends only on the moneyness

$$\frac{d\sigma}{dP} = \frac{\partial \sigma}{\partial C(P)} \frac{dC(P)}{dP}.$$

- The volatility depends on the time as an inverse square

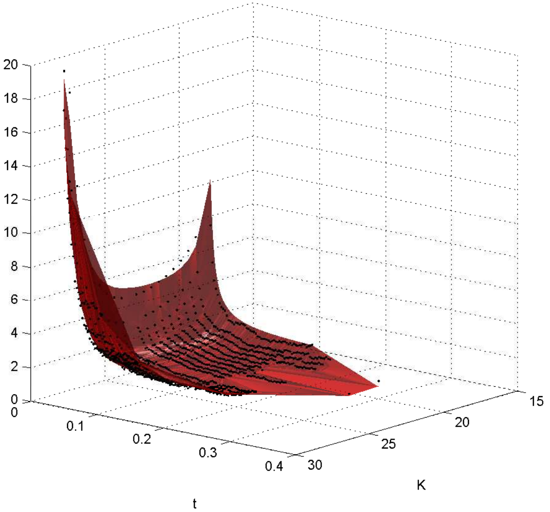
$$\sigma = \Phi\left(\frac{ln(K/F)}{\sqrt{t}}\right).$$

- We use a data set of the quarterly options for SPX for the beginning of 2008.
- The initial model is given by the RTS experts:

$$\sigma = \sigma(\mathbf{w}) = w_1 + w_2(1 - \exp(-w_3 x^2)) + \frac{w_4 \arctan(w_5 x)}{w_5},$$

where $x = \frac{\ln K - \ln C(t)}{\sqrt{t}}$.

$$\sigma = (w_1 K + w_2)\mathcal{N}\left(\frac{\ln K}{\sqrt{t}}, w_3\right) + w_5 \arctan \frac{\ln K - w_6 K^2 - w_7 K - w_8}{\sqrt{t}}$$



During the computational experiment:

- 10 runs of the algorithm were made,
- more than 22000 models generated.

The 20 best models satisfy the expert requirements:

- inverse-square dependence on time to expiration,
- Most part has polynomial and exponent dependence on strike,
- mean error is 1.18%, max error is 15.1%.

So the models are interpretable and adequate as well.

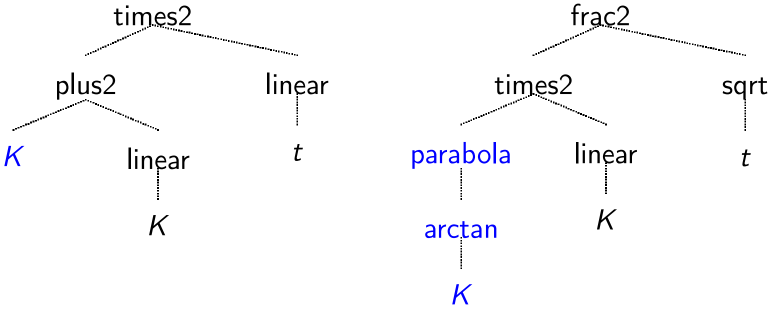
Function	Description	Parametres
	$g(\mathbf{w}, x_1, x_2)$	
plus2	$y = x_1 + x_2$	–
times2	$y = x_1 x_2$	–
frac2	$y = x_1 / x_2$	–
	$g(\mathbf{w}, x)$	
inv	$y = 1/x$	–
add	$y = x + a$	a
normalpdf	$y = \frac{\lambda}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\xi)^2}{2\sigma^2}\right) + a$	λ, σ, ξ, a
linear	$y = ax + b$	a, b
parabolic	$y = ax^2 + bx + c$	a, b, c
sqrt	$y = \sqrt{x}$	–
arctan	$y = \arctan(ax)$	a

The model generation algorithm contains three main steps.
Iterations begin:

1. Optimize parameters of every model from the competitive set $f_1, \dots, f_M$:

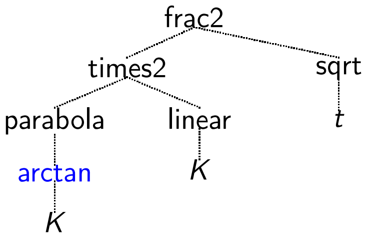
$$\mathbf{w}^{\text{MP}} = \arg\min_{\mathbf{w}} E_D(\mathbf{w}|D, f_i).$$

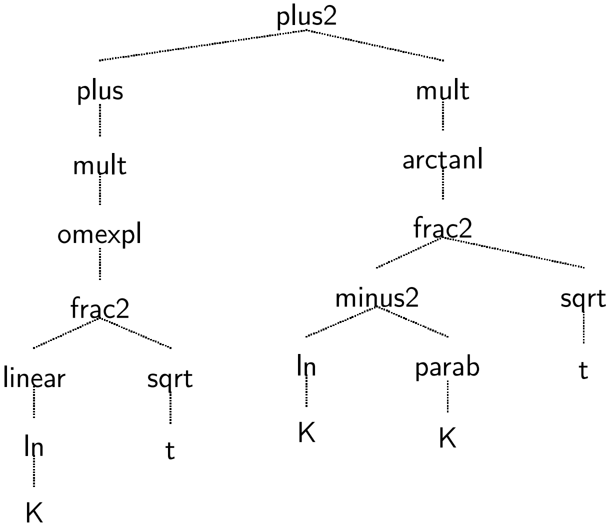
2. Make the element exchange:
   1. select a pair of indexes $i, j \in \{1, \ldots, M\}$ randomly,
   2. select the elements g_{ik} and g_{jl} in the models f_i and f_j,
   3. new models f_i' и f_j' are created by this exchange



Element modification

3. Make the modification of the generated models $\{f_i'\}$.
   1. select an element g_{ik} from the set of the elements of f_i' randomly,
   2. select an element g_s from the elements of G,
   3. change g_{ik} to g_s, if the numbers of arguments coincide.




plus2
plus
mult
mult
arctanl
omexpl
frac2
frac2
minus2
sqrt
linear
sqrt
ln
parab
t
ln
t
K
K
K

Let us consider

- a set $\mathfrak{D} = \{(\mathbf{D}_k, f_k)\}$;
- $\mathbf{D}_k = (\underset{m\times n}{\mathbf{X}}, \underset{m\times 1}{\mathbf{y}})$;
- $f_k \in \mathcal{F}$ is a model that approximates $\mathbf{D}_k$;
- $\mathcal{G}$ is a set of primitive functions;
- $\mathcal{F}$ is a set of superpositions of primitive functions $g \in \mathcal{G}$:

$$\mathcal{F} = \{f_s \mid \mathbf{f}_s : (\hat{\mathbf{w}}_k, \mathbf{X}) \mapsto \mathbf{y}, s \in \mathbb{N}\}.$$

One must

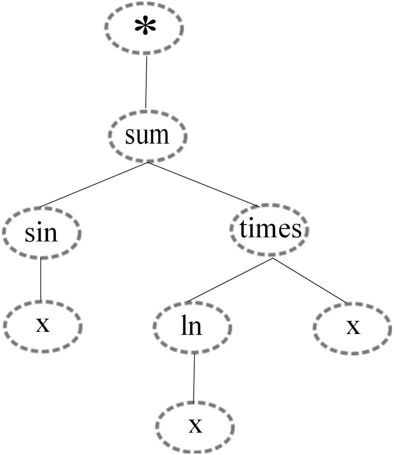
to find an algorithm $a : \mathbf{D}_k \mapsto f_s$.

For a set of superpositions $\mathcal{F}$ we need to find an index $\hat{s}$, such that the function $f_{\hat{s}}$ will bring the minimal value of the error function S among all $f \in \mathcal{F}$:

$$\hat{s} = \arg \min_{s \in \{1,\ldots,|\mathcal{F}|\}} S(f_s \mid \hat{\mathbf{w}}_k, \mathbf{D}_k),$$

where $\hat{\mathbf{w}}_k$ is a vector of optimal parameters of the model f_s for each $f \in \mathcal{F}$ given $\mathbf{D}$:

$$\hat{\mathbf{w}}_k = \arg \min_{\mathbf{w} \in \mathbb{W}_s} S(\mathbf{w} \mid f_s, \mathbf{D}_k).$$

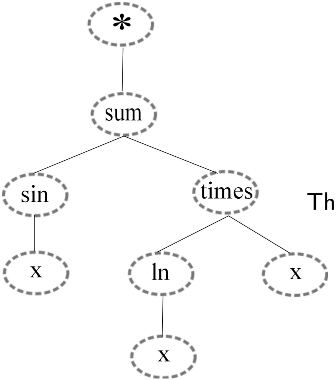


$$f = \sin(x) + (\ln x)x;$$

The tree Γ_f

1. The root is denoted $*$;
2. $V_i \mapsto g_r$;
3. $\mathrm{val}(V_j) = v(g_{r(i)})$;
4. $\mathrm{dom}(g_{r(i)}) \supset \mathrm{cod}(g_{r(j)})$;
5. the arguments g_r are ordered;
6. x_i are the leaves of Γ_f.

1. We denote the root of the tree Γ_f by a special symbol " * ". The has only one child node;
2. each non-root node V_i of the tree Γ_f has a corresponding elementary functions from the set $\mathcal{G}$;
3. the number of children nodes V_j of some node V_i is equal to the number of arguments of a corresponding function g_r: $v = v(g_r)$;
4. the domain of a function corresponding to the node V_j contains the codomain of a function of it's parent node V_i: $\mathrm{dom}(g_{r(i)}) \supset \mathrm{cod}(g_{r(j)})$;
5. the order of the children nodes of a node V_i relates to the order of the arguments of the corresponding function $g_r, r = r(i)$;
6. the leaves of the tree Γ_f relate to the free variables x_i.



$f = \sin(x) + (\ln x)x$

The matrix Z_f of links the tree Γ_f

	sum	times	ln	sin	x
$*$	1	0	0	0	0
sum	0	1	1	0	0
times	0	0	0	1	1
ln	0	0	0	0	1
sin	0	0	0	0	1

The matrix P_f of link probabilities of the tree Γ_f

	sum	times	ln	sin	x
$*$	0.7	0.1	0.1	0.1	0.2
sum	0.2	0.7	0.8	0.1	0.2
times	0.1	0.3	0	0.8	0.8
ln	0.2	0.1	0.3	0.1	0.9
sin	0.1	0.2	0.1	0	0.8

$a : \mathbf{D}_k \mapsto f_s.$

The goal is:

to find a matrix P_s of link probabilities;
to find $Z_{f_s} = \arg \max\limits_{Z \in \mathcal{M}} \sum_{i,j} P_{ij} \times Z_{i,j}$,

where $\mathcal{M}$ is a set of matrices, each one encoding a superposition from $\mathcal{F}$.

Let K be the maximal acceptable complexity value.

- Claim the node i of the tree open: $i = 1$.
- While the number of ones in the matrix does not exceed K repeat:
  1. chose $c_j = \max\limits_{j=1,\dots,l} P_{ij}$ for all open nodes i;
  2. overbuild the matrix Z_f: $j^* = \arg\max\limits_j c_j, \quad Z_f(i, j^*) = 1$;
  3. add the node j^* to the list of open nodes if $(i, j^*) \in P'$;
- if the number of ones is larger than K, associate open nodes to free variables: $k^* = \arg\max\limits_k P''_{ik}, \quad (i, k^*) = 1$ for all open nodes i.

The aim of the experiment

Is to verify the suggested procedure of forecasting a superposition.

The data set $\mathfrak{D} = \{(\mathbf{D}_s, f_s)\}$ and the algorithm a:$\mathbf{D} \mapsto \hat{\Gamma}$ are given.
The Leave-One-Out procedure is executed in the following way:

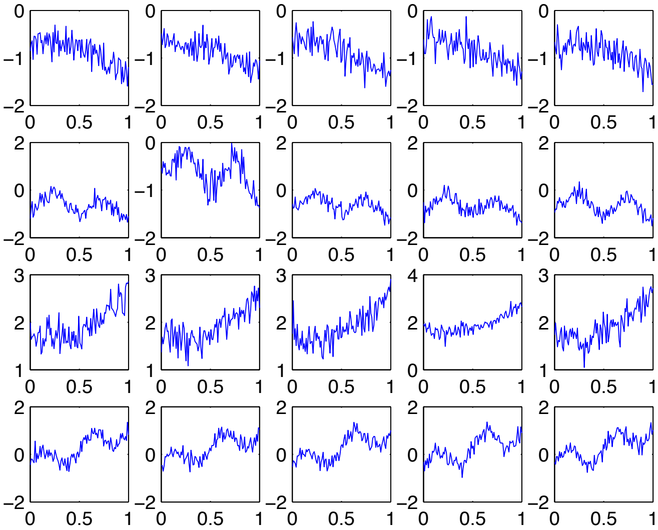
1. The parameters a are optimized on the basis of a learning sample $\mathfrak{D} \setminus \{\mathbf{D}_k\}$.
2. The tree $\hat{\Gamma}_k = a(\mathbf{D}_k)$ is constructed.
3. Based on $\hat{\Gamma}_k$ a model $\hat{f}_s$ is designed.
4. The parameters $\hat{\mathbf{w}}_k$ of the model $\hat{f}_s$ are optimized.
5. The error function value $S(\hat{\mathbf{w}}_k, \hat{f}_s, f_s) = \| \mathbf{y} - f(\mathbf{w}_k, \mathbf{X}) \|_2$ is computed.

Statistical hypothesis (or generating a synthetic data set)

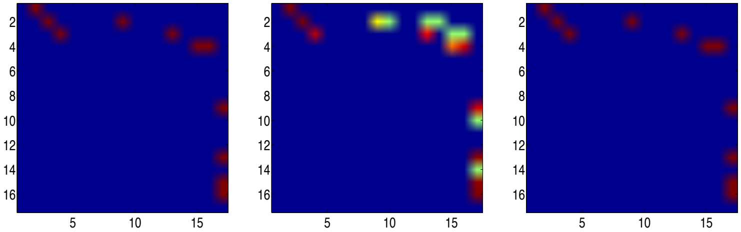
1. Fixate the the model f_s from a set $\mathcal{F}$ and the parameters $\mathbf{w}_s \in \mathbb{W}_s$;
2. initialize the matrix $\mathbf{X}$;
3. compute $f(\mathbf{w}_s, \mathbf{X})$;
4. fixate τ_f, $|\tau_f| < \epsilon$;
5. compute $\mathbf{y} = f(\mathbf{w}_s, \mathbf{X}) + \tau_f$;
6. repeat r times for each model $f \in \mathcal{F}$

Thus we obtain a data set: pairs of the sets $\mathbf{D} = (\underset{m\times n}{\mathbf{X}}, \underset{m\times 1}{\mathbf{y}})$ with the corresponding models f.

The data set

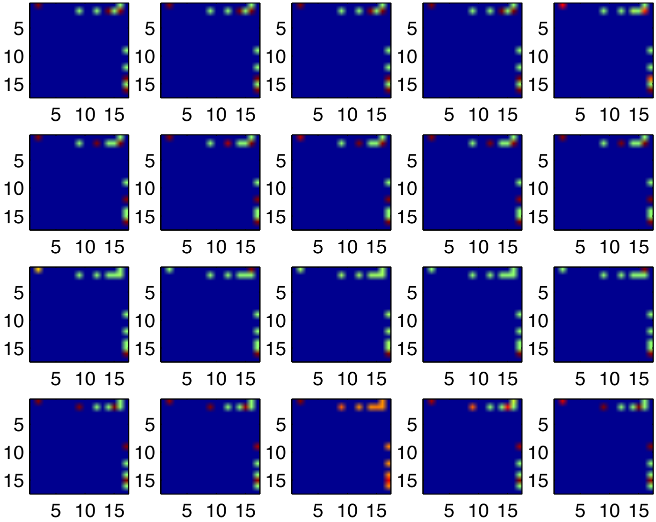


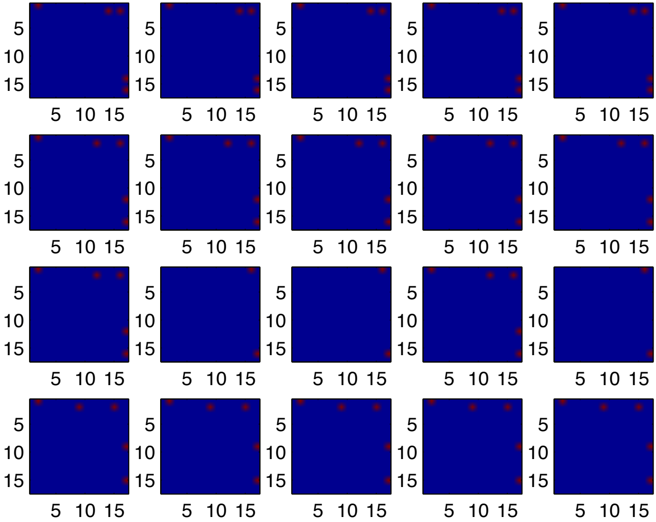
The original and the forecasted superposition



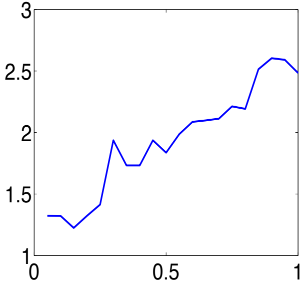
$$f = w_1 \cos(\alpha_1 x + \alpha_2) + w_2 x + w_3 \ln(\alpha_3 x + \alpha_4).$$

The obtained matrices of probabilities P_f

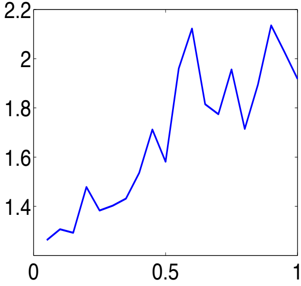



5
10
15
5 10 15

The dependance of the error function on the noise level and on the model parameters



The error function S depends on the noise ϵ



The error function S depends on the parameter bias $\delta\mathbf{w}$

- A problem of forecasting the structure of superposition was stated and solved.
- We suggest a description of allowable superpositions that satisfies the necessary restrictions.
- We propose an algorithm that constructs an allowable superposition using a matrix of probabilities of forecast.
- We have designed an algorithm of forecasting the structure of a regression model. Implemented on synthetic data, the algorithm performs adequately.