

Машины на опорных векторах (SVM-регрессия)

В.В.Вьюгин

ИППИ РАН



SVM-регрессия (SVR)

Основное предположение:

$(\bar{x}_i, y_i)$ – i.i.d согласно P , где $\bar{x}_i \in \mathcal{R}^n$, $y_i \in \mathcal{R}$.

Ошибки обобщения будут оцениваться через P .

Конкретный вид распределения P не используется (все оценки равномерны по P)



ε -нечувствительные функции потерь

Используются функции потерь – линейная

$$L^\varepsilon(\bar{x}, y, f) = |y - f(\bar{x})|_\varepsilon = \max\{0, |y - f(\bar{x})| - \varepsilon\}$$

и квадратичная

$$L_2^\varepsilon(\bar{x}, y, f) = |y - f(\bar{x})|_\varepsilon^2.$$



Прямая задача минимизации в случае линейной функции потерь

$$|\bar{w}|^2 + C \sum_{i=1}^l (\xi_i + \hat{\xi}_i) \rightarrow \min$$

при условиях $((\bar{w} \cdot \bar{x}_i) + b) - y_i \leq \varepsilon + \xi_i, \quad i = 1, \dots, l,$
 $y_i - ((\bar{w} \cdot \bar{x}_i) + b) \leq \varepsilon + \hat{\xi}_i, \quad i = 1, \dots, l,$
 $\xi_i, \hat{\xi}_i \geq 0, \quad i = 1, \dots, l.$



Лагранжиан прямой задачи имеет вид

$$\begin{aligned} L(\bar{w}, b, \bar{\xi}, \hat{\xi}, \bar{\alpha}, \hat{\alpha}) = & |\bar{w}|^2 + C \sum_{i=1}^l (\xi_i + \hat{\xi}_i) + \\ & + \sum_{i=1}^l \alpha_i ((\bar{w} \cdot \bar{x}_i) + b - y_i - \varepsilon - \xi_i) + \\ & + \sum_{i=1}^l \hat{\alpha}_i (y_i - (\bar{w} \cdot \bar{x}_i) - b - \varepsilon - \hat{\xi}_i). \end{aligned}$$



Двойственная задача

$$\sum_{i=1}^l y_i (\hat{\alpha}_i - \alpha_i) - \varepsilon \sum_{i=1}^l (\hat{\alpha}_i + \alpha_i) - \\ - \frac{1}{2} \sum_{i,j=1}^l (\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)(\bar{x}_i \cdot \bar{x}_j) \rightarrow \max$$

$$\text{при условиях } \sum_{i=1}^l (\hat{\alpha}_i - \alpha_i) = 0,$$

$$\hat{\alpha}_i \geq 0, \alpha_i \geq 0,$$

$$0 \leq \alpha_i, \hat{\alpha}_i \leq C, \quad i = 1, \dots, l.$$



Условия Каруша–Куна–Таккера

$$\alpha_i((\bar{w} \cdot \bar{x}_i) + b - y_i - \varepsilon - \xi_i) = 0, \quad i = 1, \dots, l,$$

$$\hat{\alpha}_i(y_i - (\bar{w} \cdot \bar{x}_i) - b - \varepsilon - \hat{\xi}_i) = 0, \quad i = 1, \dots, l,$$

$$(\alpha_i - C)\xi_i = 0, \quad (\hat{\alpha}_i - C)\hat{\xi}_i = 0,$$

$$\xi_i \hat{\xi}_i = 0, \quad \alpha_i \hat{\alpha}_i = 0, \quad i = 1, \dots, l.$$



Опорные векторы – это $\bar{x}_i$, для которых $\alpha_i > 0$ или $\hat{\alpha} > 0$.

Если y_i находится вне слоя размера ε вокруг оптимальной гиперплоскости (гиперповерхности), то $\alpha_i = C$ или $\hat{\alpha}_i = C$ (для линейной нормы).

$0 < \alpha_i < C$ может быть только для векторов со значениями y_i на границе слоя.

Векторы $\bar{x}_i$, у которых значения y_i расположены внутри слоя, заведомо не являются опорными; для них $\alpha_i = 0$ и $\hat{\alpha}_i = 0$.



Весовой вектор – линейная комбинация опорных векторов

$$\bar{w} = \frac{1}{2} \sum_{i=1}^l (\hat{\alpha}_i - \alpha_i) \bar{x}_i.$$

Выполнено $\alpha_i \hat{\alpha}_i = 0$.

Функция регрессии

$$f(\bar{x}) = \sum_{i=1}^l \alpha_i (\bar{x}_i \cdot \bar{x}) + b^*.$$



Для пространства с ядром

$$\sum_{i=1}^l y_i(\hat{\alpha}_i - \alpha_i) - \varepsilon \sum_{i=1}^l (\hat{\alpha}_i + \alpha_i) - \\ - \frac{1}{2} \sum_{i,j=1}^l (\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j) K(\bar{x}_i, \bar{x}_j) \rightarrow \max$$

$$\text{при условиях } \sum_{i=1}^l (\hat{\alpha}_i - \alpha_i) = 0,$$

$$\hat{\alpha}_i \geq 0, \alpha_i \geq 0,$$

$$0 \leq \alpha_i, \hat{\alpha}_i \leq C, \quad i = 1, \dots, l.$$

$$\text{Функция регрессии} - f(\bar{x}) = \sum_{i=1}^l \alpha_i K(\bar{x}_i, \bar{x}) + b^*.$$



Прямая задача минимизации в случае квадратичной функции потерь

$$|\bar{w}|^2 + C \sum_{i=1}^l (\xi_i^2 + \hat{\xi}_i^2) \rightarrow \min$$

при условиях $((\bar{w} \cdot \bar{x}_i) + b) - y_i \leq \varepsilon + \xi_i, \quad i = 1, \dots, l,$
 $y_i - ((\bar{w} \cdot \bar{x}_i) + b) \leq \varepsilon + \hat{\xi}_i, \quad i = 1, \dots, l,$
 $\xi_i, \hat{\xi}_i \geq 0, \quad i = 1, \dots, l.$



Гребневая регрессия – частный случай SVR

При $\varepsilon = 0$ в случае квадратичной функции потерь получаем задачу гребневой регрессии

$$|\bar{w}|^2 + C \sum_{i=1}^l ((\bar{w} \cdot \bar{x}_i) + b - y_i)^2 \rightarrow \min$$



Лагранжиан прямой задачи имеет вид

$$\begin{aligned} L(\bar{w}, b, \bar{\xi}, \hat{\xi}, \bar{\alpha}, \hat{\alpha}) = & |\bar{w}|^2 + C \sum_{i=1}^l (\xi_i^2 + \hat{\xi}_i^2) + \\ & + \sum_{i=1}^l \alpha_i ((\bar{w} \cdot \bar{x}_i) + b - y_i - \varepsilon - \xi_i) + \\ & + \sum_{i=1}^l \hat{\alpha}_i (y_i - (\bar{w} \cdot \bar{x}_i) - b - \varepsilon - \hat{\xi}_i). \end{aligned}$$



Двойственная задача

$$\sum_{i=1}^l y_i(\hat{\alpha}_i - \alpha_i) - \varepsilon \sum_{i=1}^l (\hat{\alpha}_i + \alpha_i) - \\ - \frac{1}{2} \sum_{i,j=1}^l (\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)((\bar{x}_i \cdot \bar{x}_j) + \frac{1}{C} \delta_{ij}) \rightarrow \max$$

$$\text{при условиях } \sum_{i=1}^l (\hat{\alpha}_i - \alpha_i) = 0,$$

$$\hat{\alpha}_i \geq 0, \alpha_i \geq 0, \quad i = 1, \dots, l,$$

где $\delta_{ij} = 1$ тогда и только тогда, когда $i = j$.



Условия Каруша–Куна–Таккера

$$\alpha_i((\bar{w} \cdot \bar{x}_i) + b - y_i - \varepsilon - \xi_i) = 0, \quad i = 1, \dots, l,$$

$$\hat{\alpha}_i(y_i - (\bar{w} \cdot \bar{x}_i) - b - \varepsilon - \hat{\xi}_i) = 0, \quad i = 1, \dots, l,$$

$$\xi_i \hat{\xi}_i = 0, \quad \alpha_i \hat{\alpha}_i = 0, \quad i = 1, \dots, l.$$



Вероятность ошибки SVM-регрессии

Пусть $\gamma, \varepsilon > 0$. Распределение P сконцентрировано в шаре радиуса R с центром в начале координат. Тогда с P^l -вероятностью $1 - \delta$ на выборке $S = ((\bar{x}_1, y_1), \dots, (\bar{x}_l, y_l))$ произвольная гипотеза $f \in \mathcal{L}$ имеет оценку вероятности ошибки обобщения:

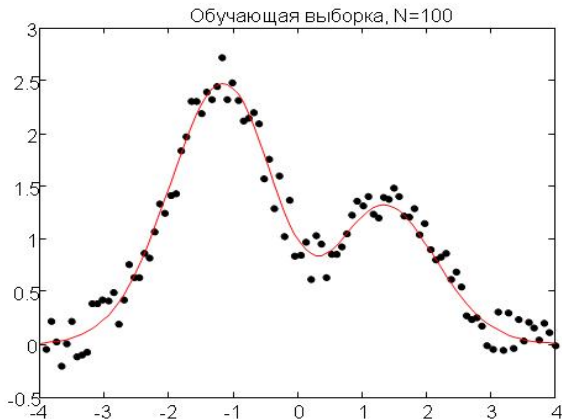
$$\begin{aligned} & P\{|f(\bar{x}) - y| > \gamma + \varepsilon\} \leq \\ & \leq \frac{c}{l} \left(\frac{|\bar{w}|^2 R^2 + |\bar{\xi}|_\varepsilon^2}{\gamma^2} \log^2 l + \log \frac{1}{\delta} \right), \end{aligned}$$

где $|\bar{\xi}|_\varepsilon^2 = \xi_1^2 + \dots, \xi_l^2$, $\xi_i = \max\{0, |y_i - f(\bar{x}_i)| - \varepsilon\}$, $i = 1, \dots, l$.



Эксперименты с SVM-регрессией



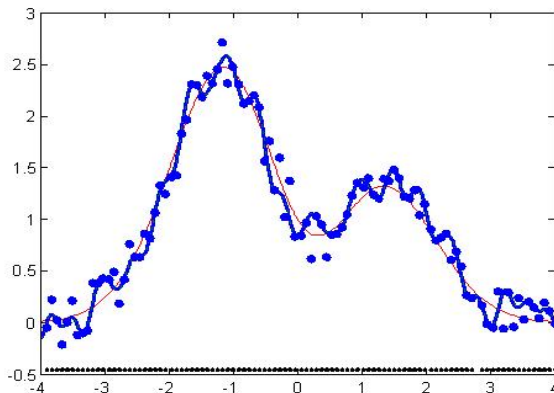


Обучающая выборка $N = 100$

$$y(x) = (1 - x + 2x^2)\exp(-x^2/2) + 0.2\xi$$



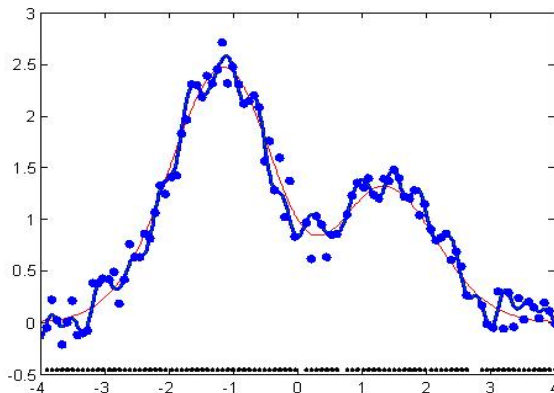
$\epsilon = 0.001$ #SV=99/100 MAE=0.106 RMSE=0.126



Тестовая выборка $N = 1000$
 $\epsilon = 0.001, SV = 99/100, er_1 = 0.106, er_2 = 0.126$



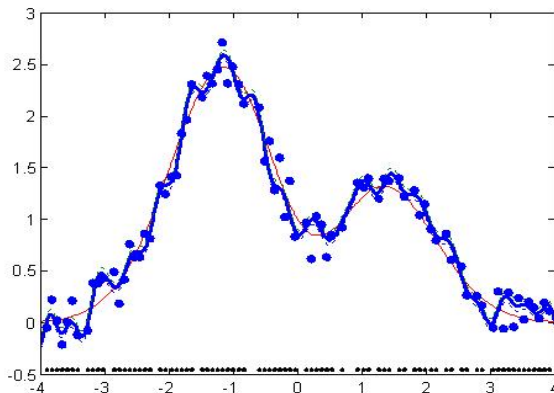
$\epsilon = 0.005$ #SV=96/100 MAE=0.105 RMSE=0.124



Тестовая выборка $N = 1000$
 $\epsilon = 0.005, SV = 96/100, er_1 = 0.105, er_2 = 0.124$



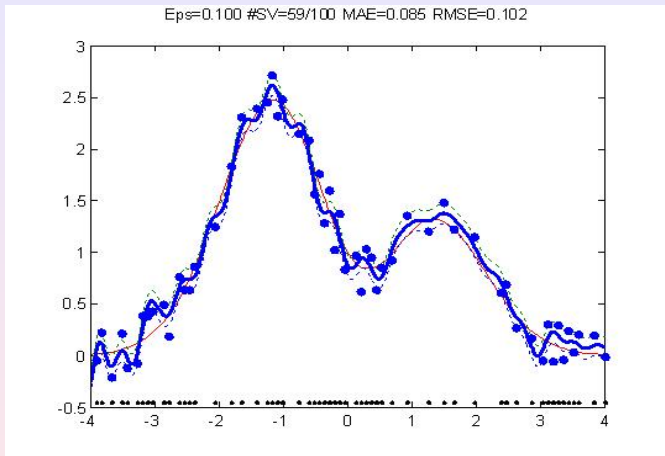
$\epsilon = 0.050$ #SV=82/100 MAE=0.097 RMSE=0.112



Тестовая выборка $N = 1000$

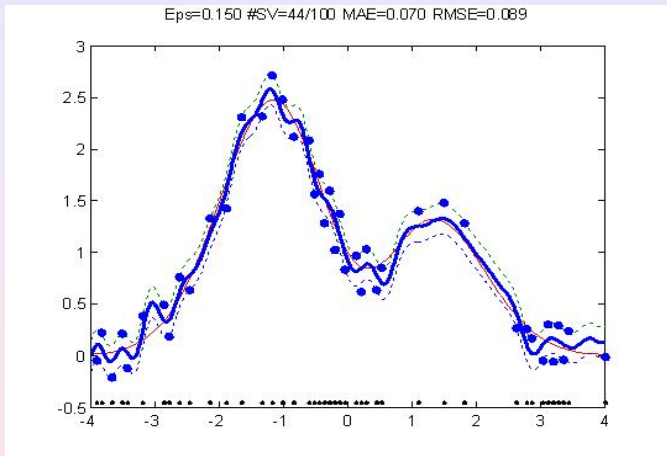
$\epsilon = 0.050, SV = 82/100, er_1 = 0.097, er_2 = 0.112$





Тестовая выборка $N = 1000$
 $\epsilon = 0.100, SV = 59/100, er_1 = 0.085, er_2 = 0.102$

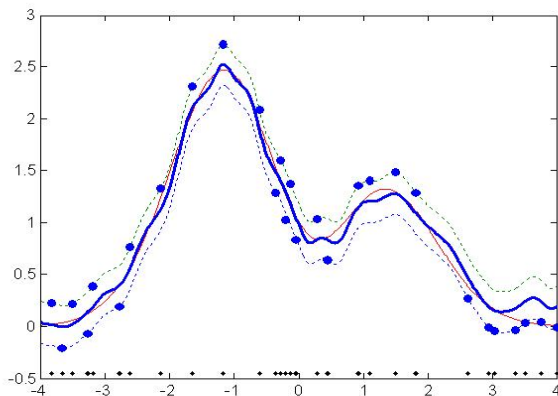




Тестовая выборка $N = 1000$
 $\epsilon = 0.150, SV = 44/100, er_1 = 0.070, er_2 = 0.089$

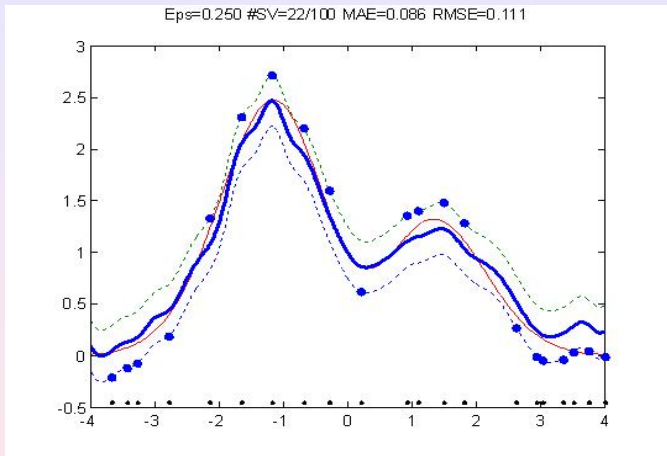


Eps=0.200 #SV=30/100 MAE=0.059 RMSE=0.081



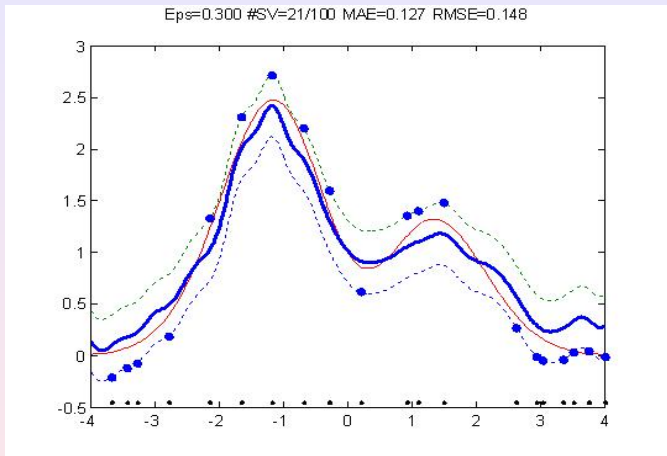
Тестовая выборка $N = 1000$
 $\varepsilon = 0.200, SV = 30/100, er_1 = 0.059, er_2 = 0.081$





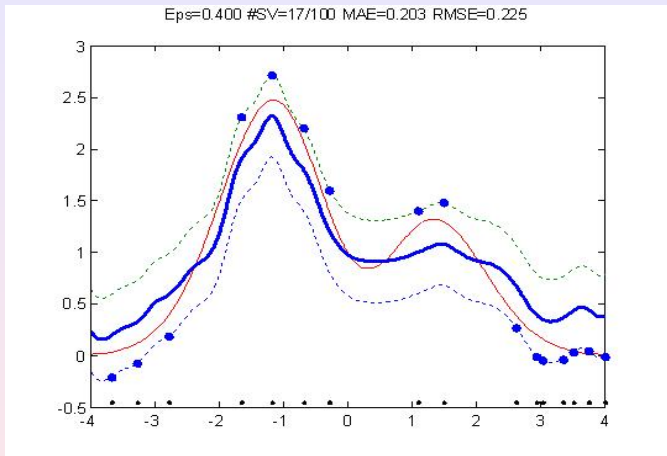
Тестовая выборка $N = 1000$
 $\epsilon = 0.250, SV = 22/100, er_1 = 0.086, er_2 = 0.111$





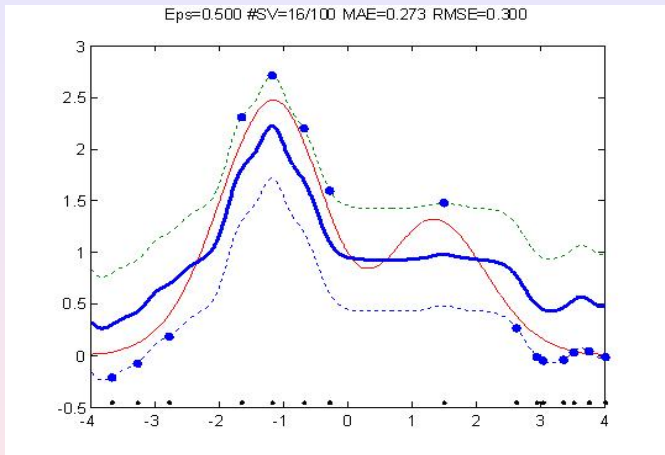
Тестовая выборка $N = 1000$
 $\epsilon = 0.300, SV = 21/100, er_1 = 0.127, er_2 = 0.148$





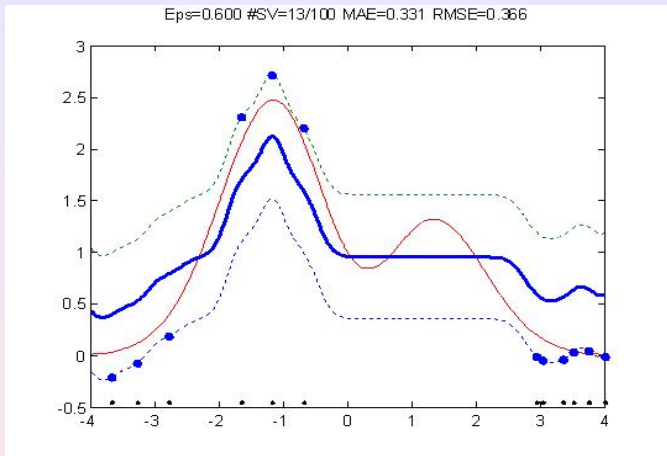
Тестовая выборка $N = 1000$
 $\epsilon = 0.400, SV = 17/100, er_1 = 0.203, er_2 = 0.225$





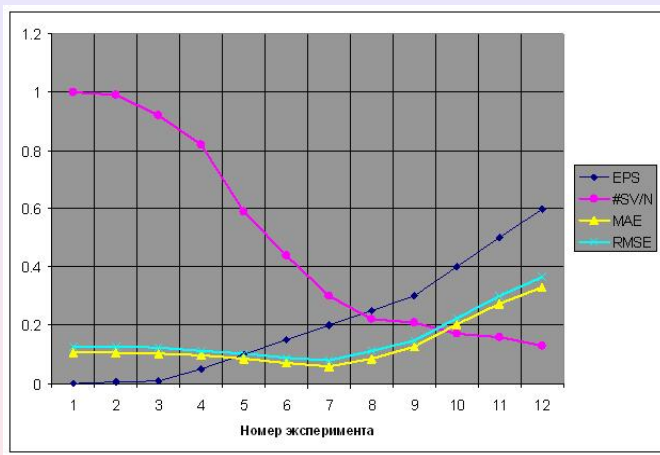
Тестовая выборка $N = 1000$
 $\epsilon = 0.500, SV = 16/100, er_1 = 0.273, er_2 = 0.300$





Тестовая выборка $N = 1000$
 $\epsilon = 0.600, SV = 13/100, er_1 = 0.331, er_2 = 0.366$





Значения ϵ , доля SV, ошибки - абс., квадр.



Прямая задача минимизации с переменным ε в случае линейной функции потерь

$$|\bar{w}|^2 + C \left(\sum_{i=1}^l (\xi_i + \hat{\xi}_i) + v\varepsilon \right) \rightarrow \min$$

при условиях $((\bar{w} \cdot \bar{x}_i) + b) - y_i \leq \varepsilon + \xi_i, \quad i = 1, \dots, l,$

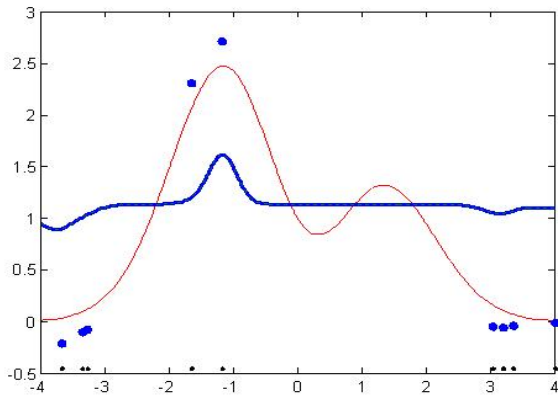
$y_i - ((\bar{w} \cdot \bar{x}_i) + b) \leq \varepsilon + \hat{\xi}_i, \quad i = 1, \dots, l,$

$\xi_i, \hat{\xi}_i \geq 0, \varepsilon > 0, \quad i = 1, \dots, l,$

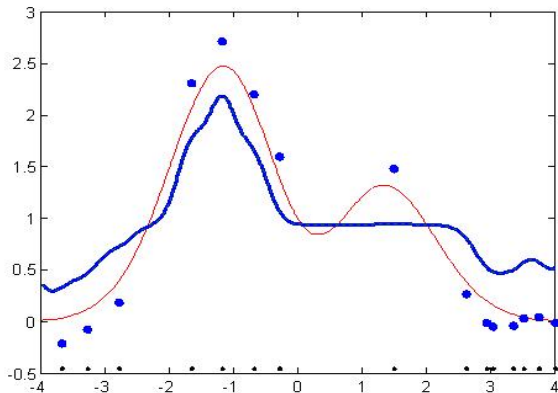
C и v – параметры.



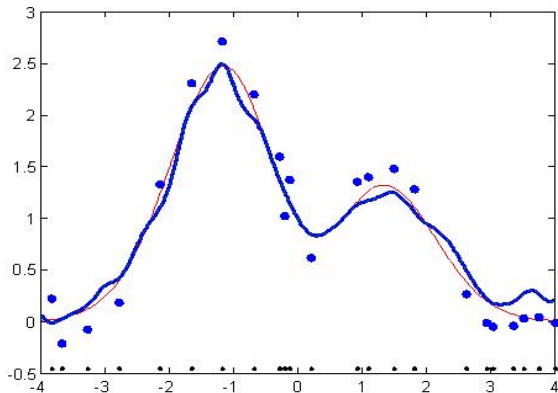
$C=10.0$ $\text{Nu}=0.001$ $\text{\#SV}=10/100$ $\text{MAE}=0.582$ $\text{RMSE}=0.683$



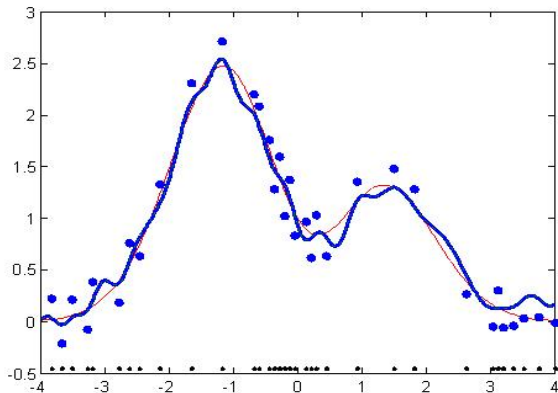
$C=10.0$ $Nu=0.005$ $\#SV=16/100$ $MAE=0.296$ $RMSE=0.326$



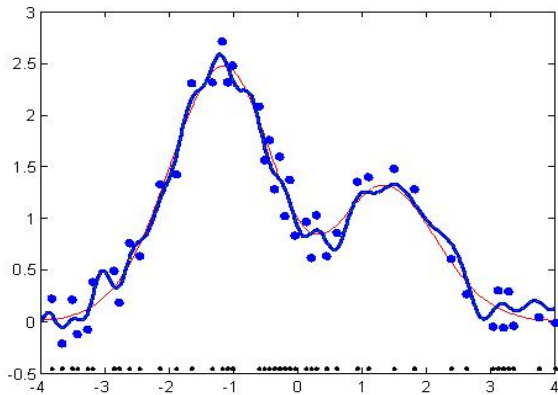
$C=10.0$ $Nu=0.010$ $\#SV=24/100$ $MAE=0.070$ $RMSE=0.097$



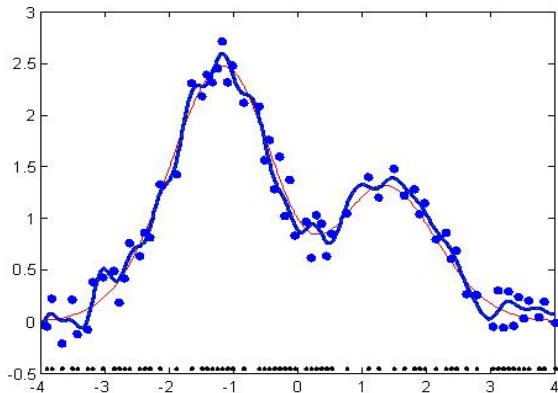
$C=10.0$ $Nu=0.100$ $\#SV=35/100$ $MAE=0.067$ $RMSE=0.088$



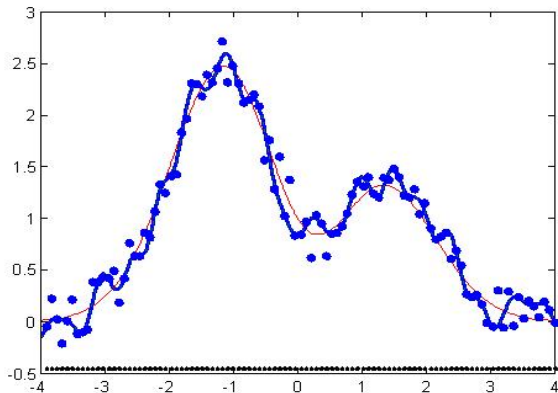
$C=10.0$ $Nu=0.200$ $\#SV=44/100$ $MAE=0.077$ $RMSE=0.096$

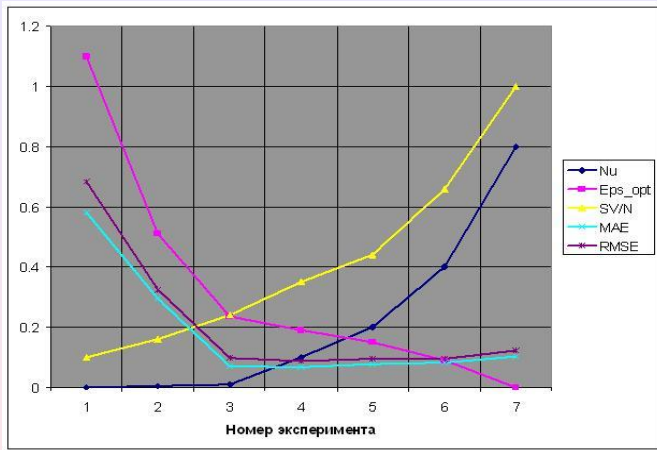


$C=10.0$ $Nu=0.400$ $\#SV=66/100$ $MAE=0.082$ $RMSE=0.096$



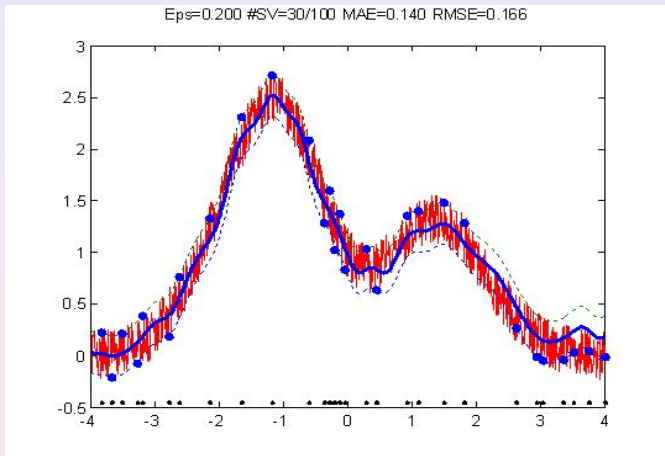
$C=10.0$ $Nu=0.800$ $\#SV=100/100$ $MAE=0.103$ $RMSE=0.122$





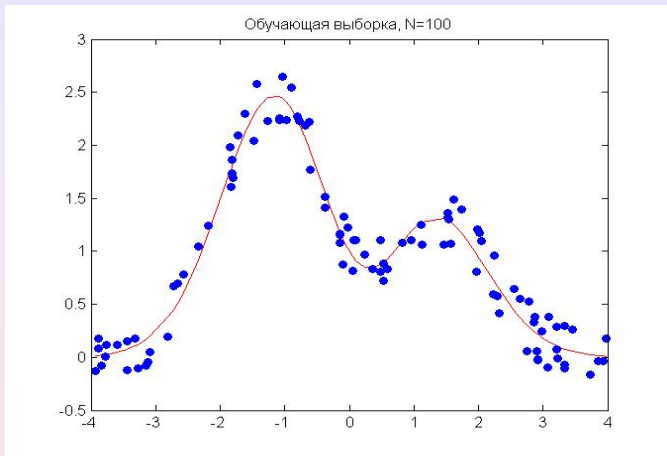
Другие примеры





Функция с большой соболевской нормой: $N = 1000$
 $\epsilon = 0.200$, $SV = 30/100$, $er_1 = 0.140$, $er_2 = 0.166$

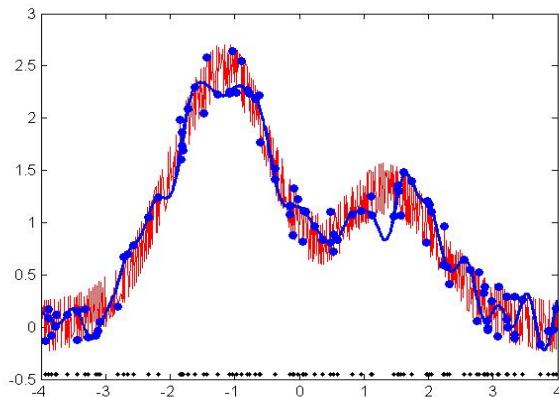




Гладкая функция с колебаниями: $N = 1000$
$$y(x) = (1 - x + 2x^2)\exp(-x^2/2) + 0.2\sin(200x)$$

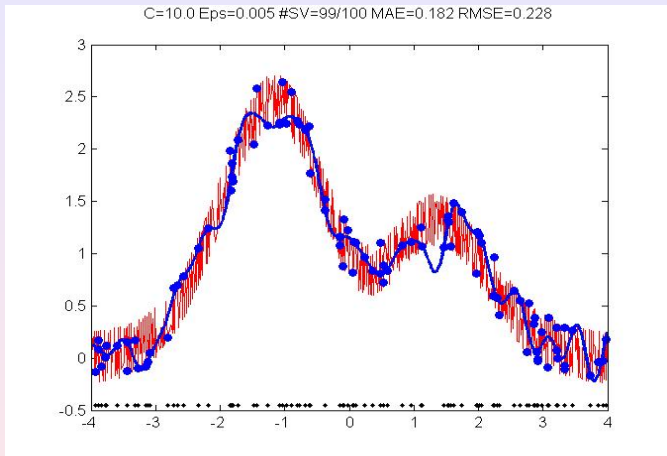


$C=10.0$ Eps=0.001 #SV=100/100 MAE=0.182 RMSE=0.228



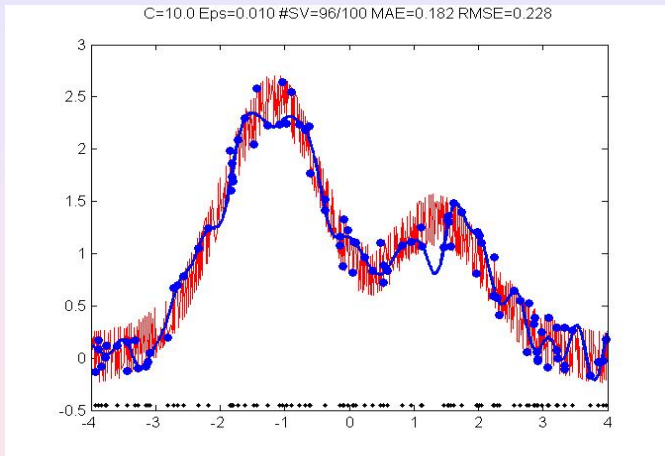
Гладкая функция с колебаниями: $N = 1000$
 $C = 10, \varepsilon = 0.001, SV = 100/100, er_1 = 0.182, er_2 = 0.228$





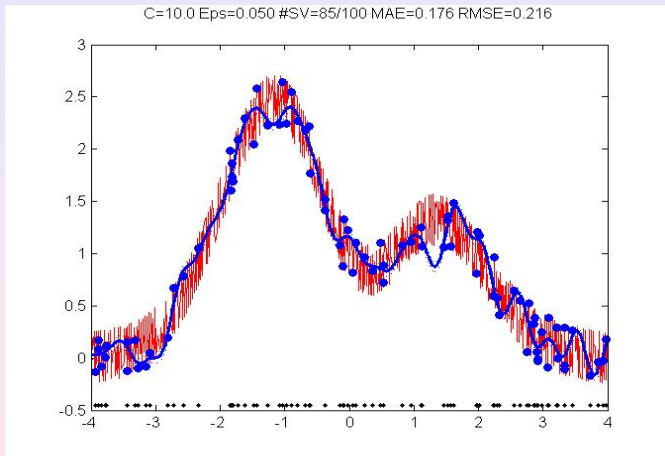
Гладкая функция с колебаниями: $N = 1000$
 $C = 10, \varepsilon = 0.005, SV = 99/100, er_1 = 0.182, er_2 = 0.228$





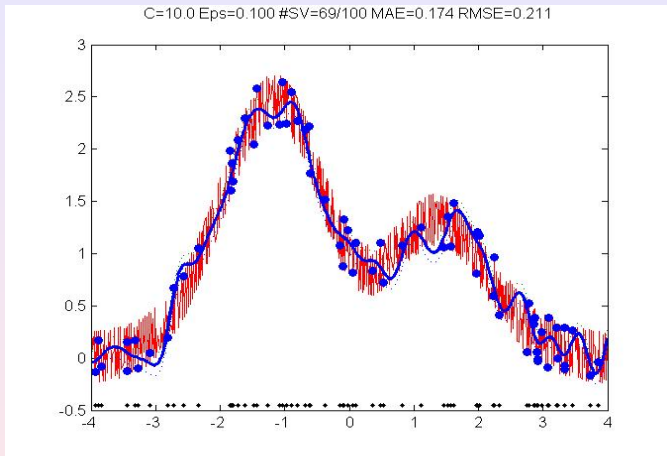
Гладкая функция с колебаниями: $N = 1000$
 $C = 10, \varepsilon = 0.010, SV = 96/100, er_1 = 0.182, er_2 = 0.228$





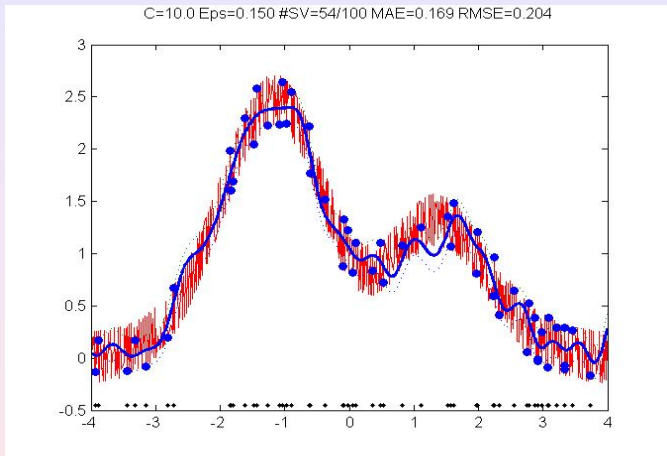
Гладкая функция с колебаниями: $N = 1000$
 $C = 10, \varepsilon = 0.050, SV = 85/100, er_1 = 0.176, er_2 = 0.216$





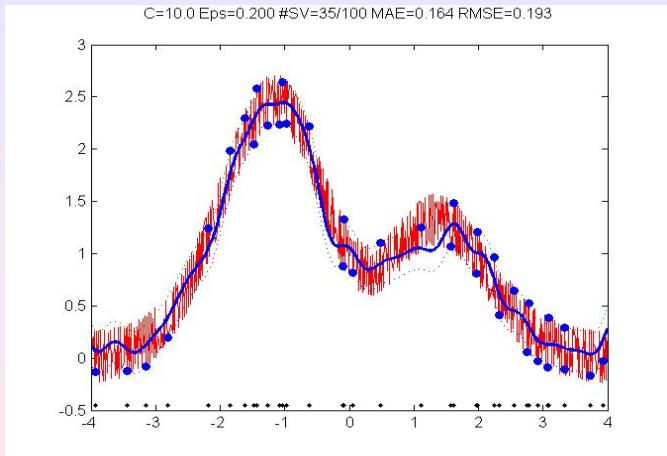
Гладкая функция с колебаниями: $N = 1000$
 $C = 10, \varepsilon = 0.100, SV = 69/100, er_1 = 0.174, er_2 = 0.211$





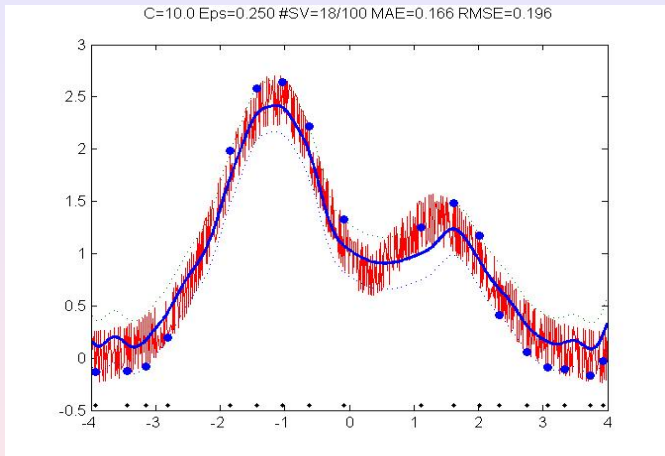
Гладкая функция с колебаниями: $N = 1000$
 $C = 10, \varepsilon = 0.150, SV = 54/100, er_1 = 0.169, er_2 = 0.204$





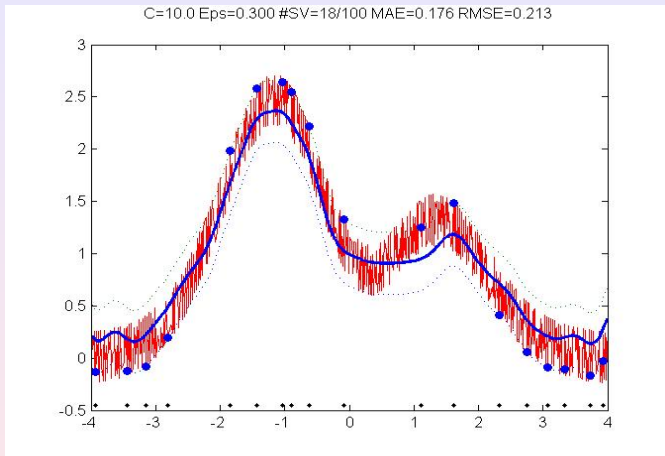
Гладкая функция с колебаниями: $N = 1000$
 $C = 10, \varepsilon = 0.200, SV = 35/100, er_1 = 0.164, er_2 = 0.193$





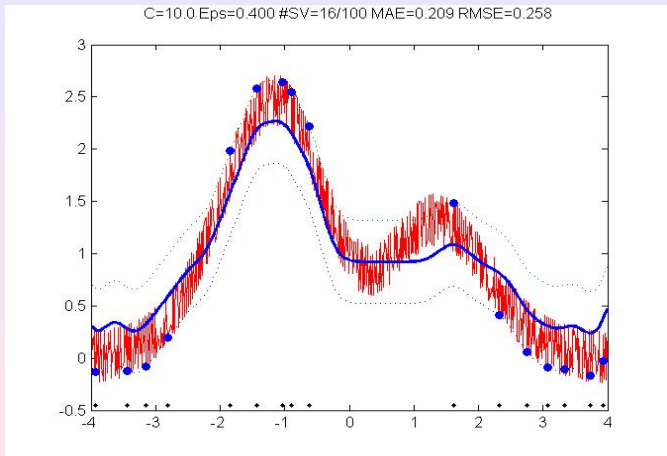
Гладкая функция с колебаниями: $N = 1000$
 $C = 10, \varepsilon = 0.250, SV = 18/100, er_1 = 0.166, er_2 = 0.196$





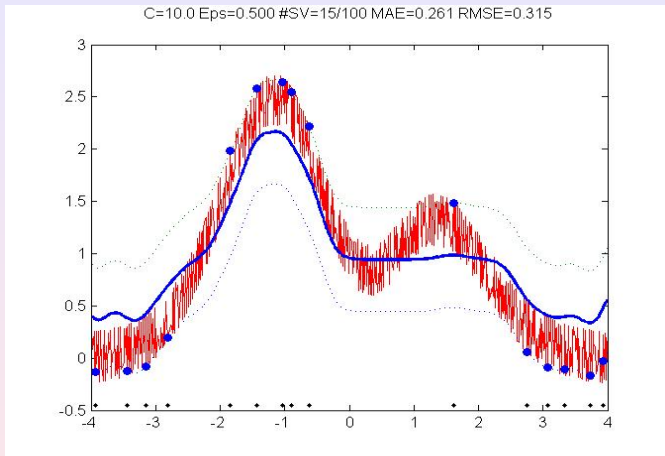
Гладкая функция с колебаниями: $N = 1000$
 $C = 10, \varepsilon = 0.300, SV = 18/100, er_1 = 0.176, er_2 = 0.213$





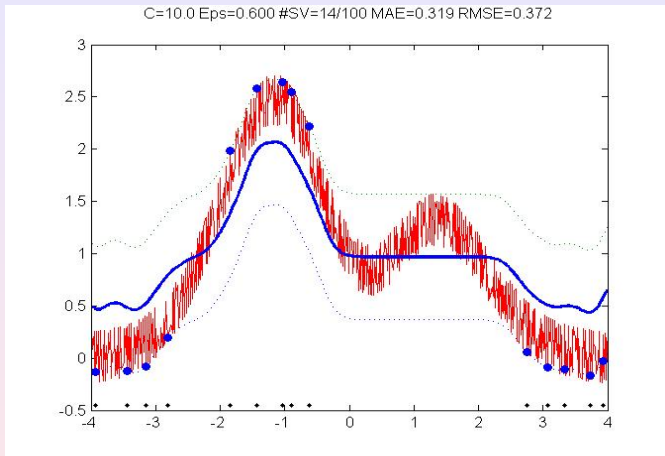
Гладкая функция с колебаниями: $N = 1000$
 $C = 10, \varepsilon = 0.400, SV = 16/100, er_1 = 0.209, er_2 = 0.258$





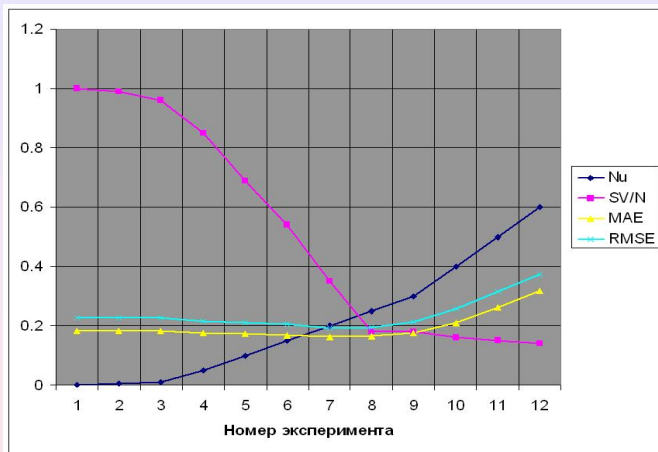
Гладкая функция с колебаниями: $N = 1000$
 $C = 10, \varepsilon = 0.500, SV = 15/100, er_1 = 0.261, er_2 = 0.315$





Гладкая функция с колебаниями: $N = 1000$
 $C = 10, \varepsilon = 0.600, SV = 14/100, er_1 = 0.319, er_2 = 0.372$





Гладкая функция с колебаниями: $N = 1000$
Сравнительный анализ всех экспериментов

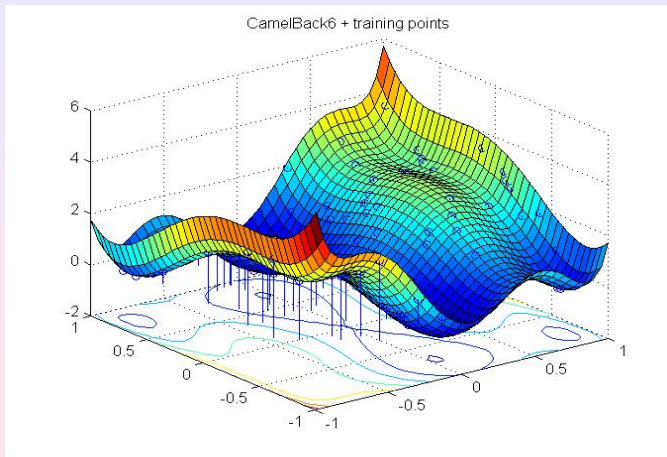


Восстановление двумерной поверхности.

$$f_{\text{Sixh}}(x_1, x_2) = (4 - 2.1x_1^2 + x_1^{4/3}) \cdot x_1^2 + x_1x_2 + (-4 + 4x_2^2) \cdot x_2^2$$

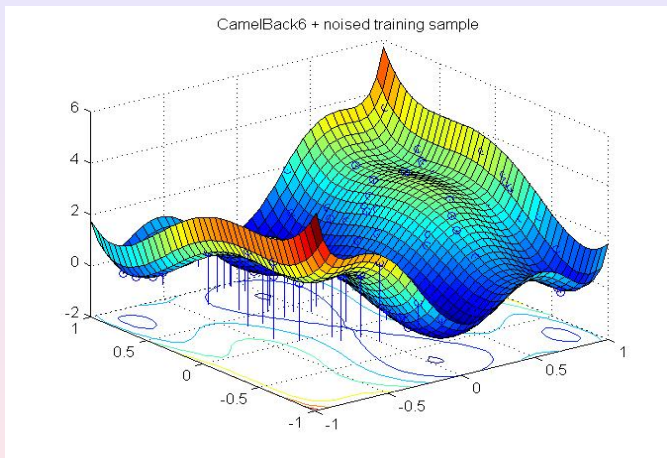
$$-3 \leq x_1 \leq 3, -2 \leq x_2 \leq 2$$





Поверхность с обучающей выборкой





Поверхность с зашумленной обучающей выборкой



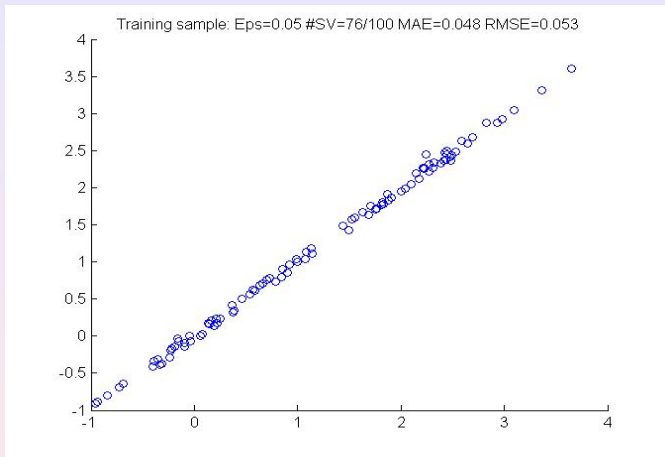


Диаграмма рассеяния на обучающей выборке. $\varepsilon = 0.05$

По оси X – предсказанные значения

По оси Y – истинные значения



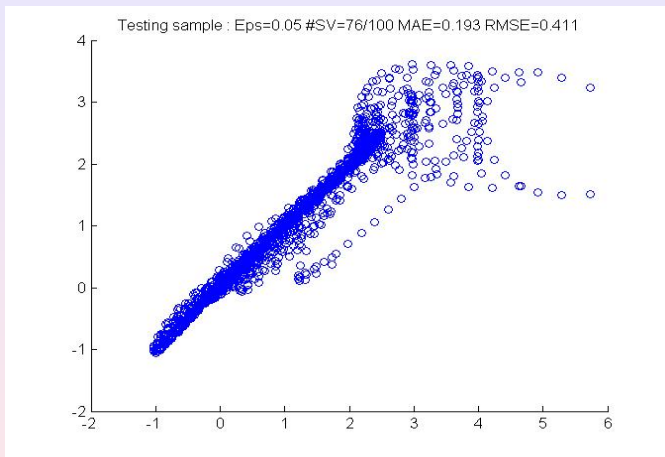
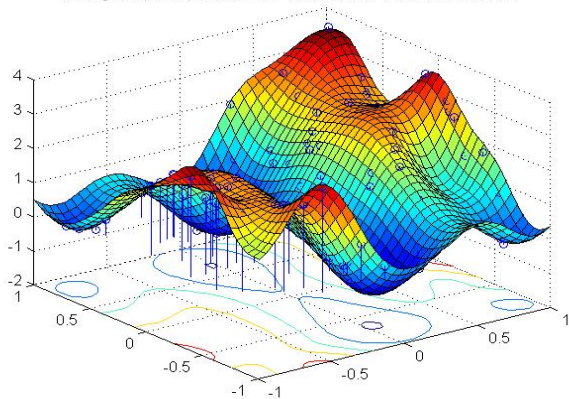


Диаграмма рассеяния на тестовой выборке. $\varepsilon = 0.05$

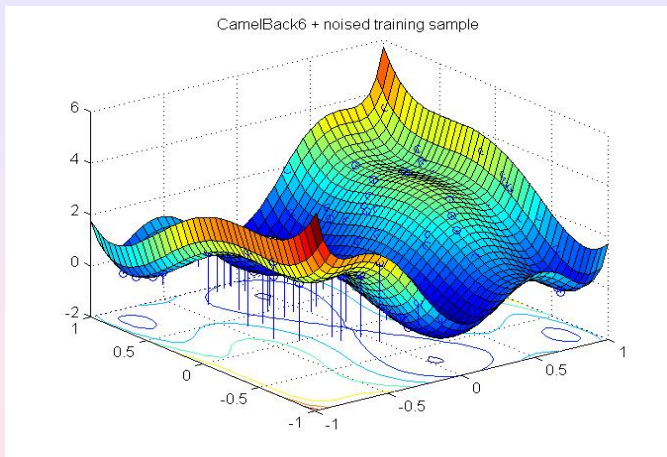


Testing sample : Eps=0.05 #SV=76/100 MAE=0.193 RMSE=0.411



Восстановленная поверхность. $\varepsilon = 0.05$





Исходная поверхность



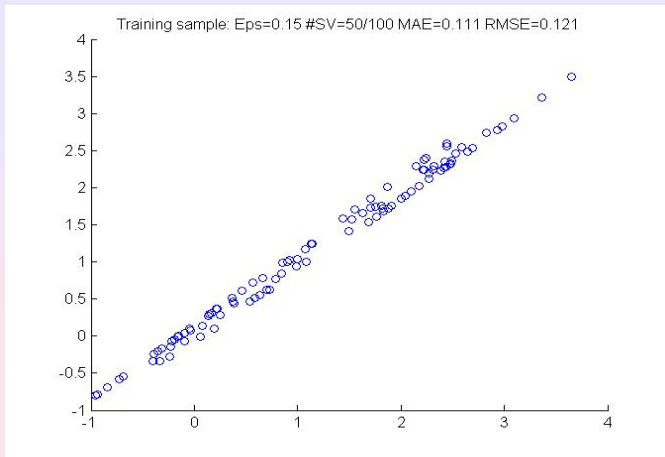


Диаграмма рассеяния на обучающей выборке. $\epsilon = 0.15$



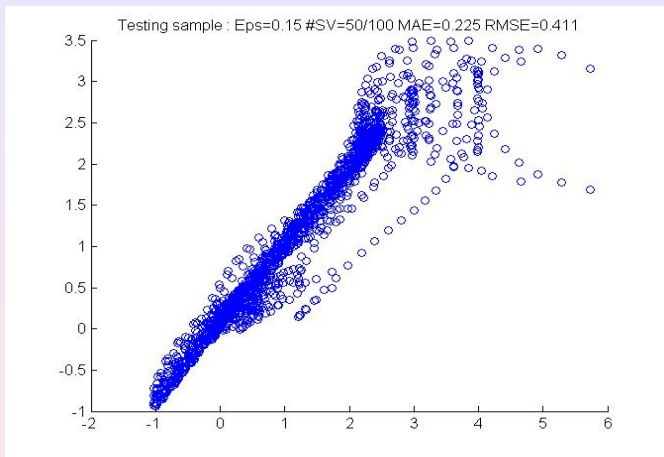
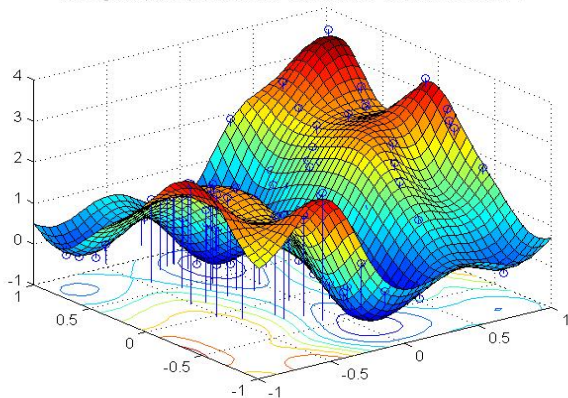


Диаграмма рассеяния на тестовой выборке. $\varepsilon = 0.15$



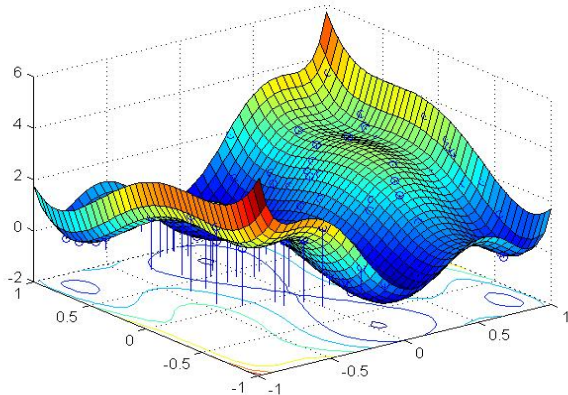
Testing sample : Eps=0.15 #SV=50/100 MAE=0.225 RMSE=0.411



Восстановленная поверхность. $\varepsilon = 0.15$



CamelBack6 + noised training sample



Исходная поверхность



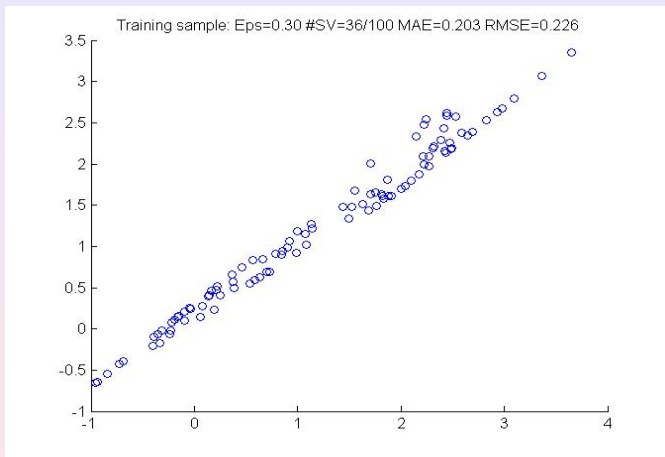


Диаграмма рассеяния на обучающей выборке. $\epsilon = 0.30$



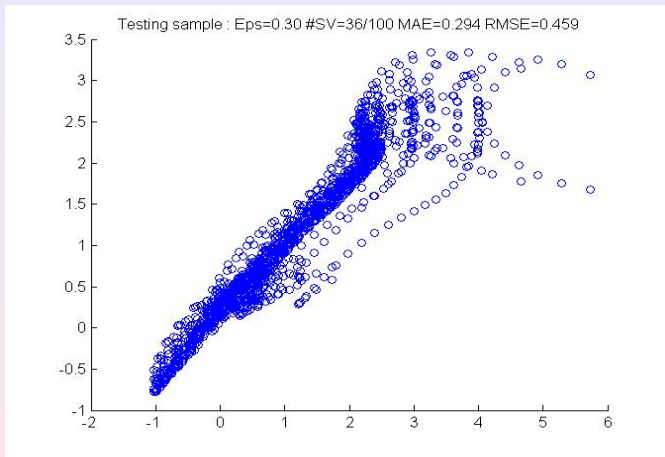
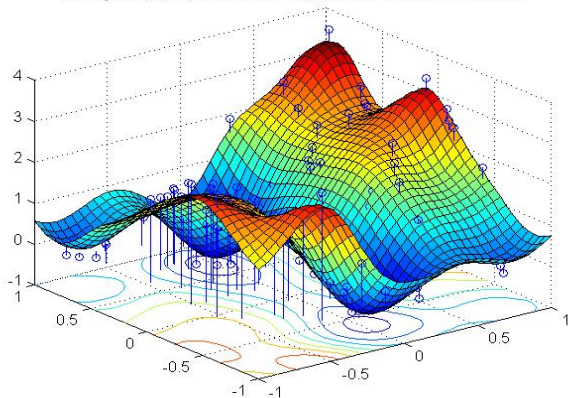


Диаграмма рассеяния на тестовой выборке. $\varepsilon = 0.30$



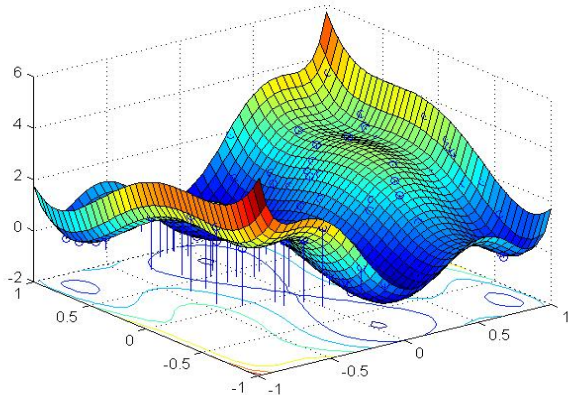
Testing sample : Eps=0.30 #SV=36/100 MAE=0.294 RMSE=0.459



Восстановленная поверхность. $\varepsilon = 0.30$



CamelBack6 + noised training sample



Исходная поверхность



Некоторые SVM-ресурсы

Литература и программы SVM

www.support-vector.net

Программы SVM

<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>

База данных для машинного обучения находится по адресу

<http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/>

Данные для решения задач классификации и регрессии.

<http://archive.ics.uci.edu/ml/>

Данные в формате MATLAB (рукописные цифры – база USPS)

<http://www.cs.toronto.edu/~roweis/data.html>

