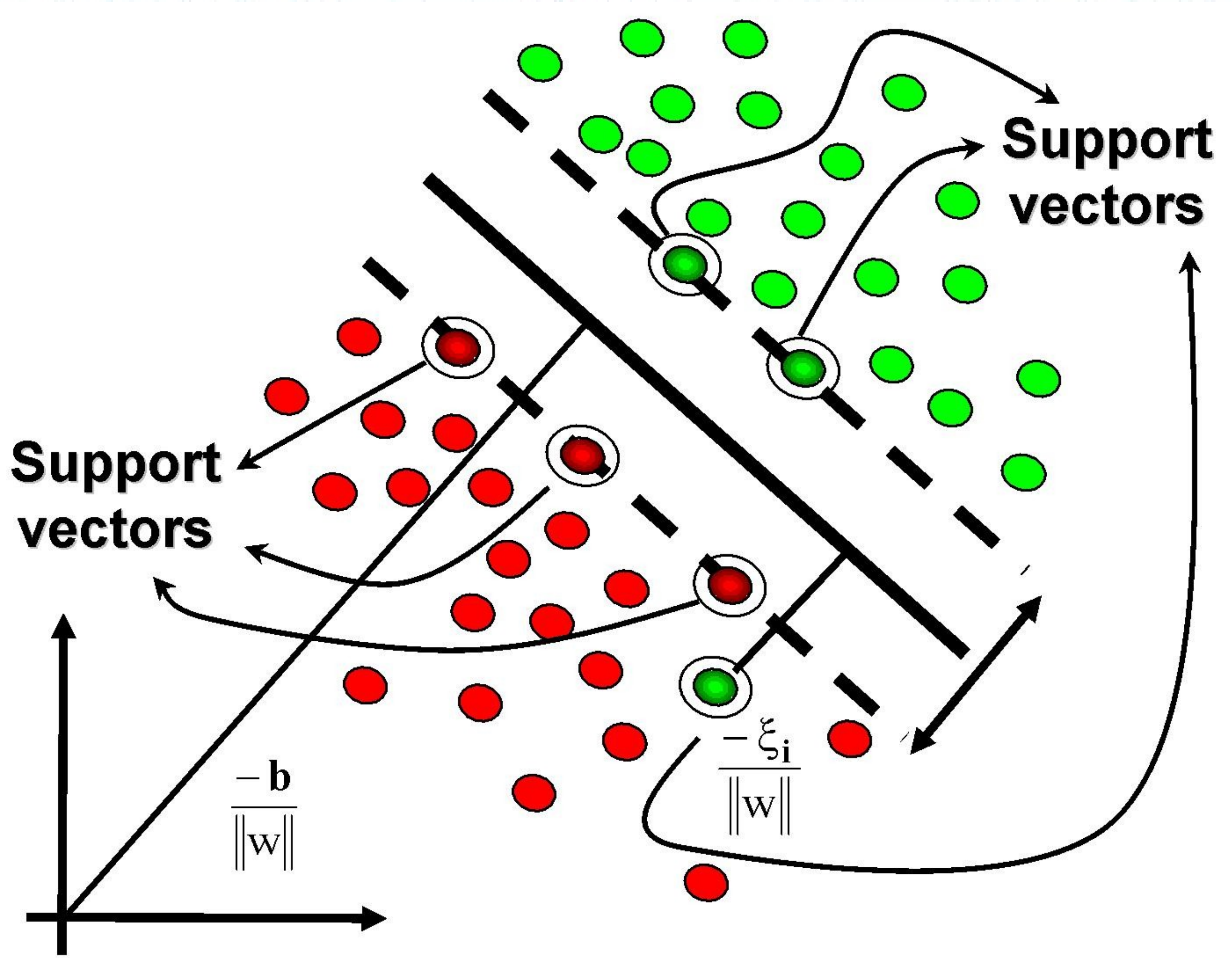
Машинное обучение
Метод опорных векторов (SVM)



https://yandexdataschool.ru/edu-process/courses/machine-learning

Содержание лекции

- Случаи линейно разделимой и неразделимой выборок
- Двойственная задача
- Типы объектов
- Нелинейное обобщение SVM
- SVM-регрессия
- L_1 регуляризация

Самая широкая разделяющая полоса

- Рассмотрим линейный классификатор:

$$a(x, w) = \operatorname{sign}\big(\langle w, x\rangle - w_0\big)$$

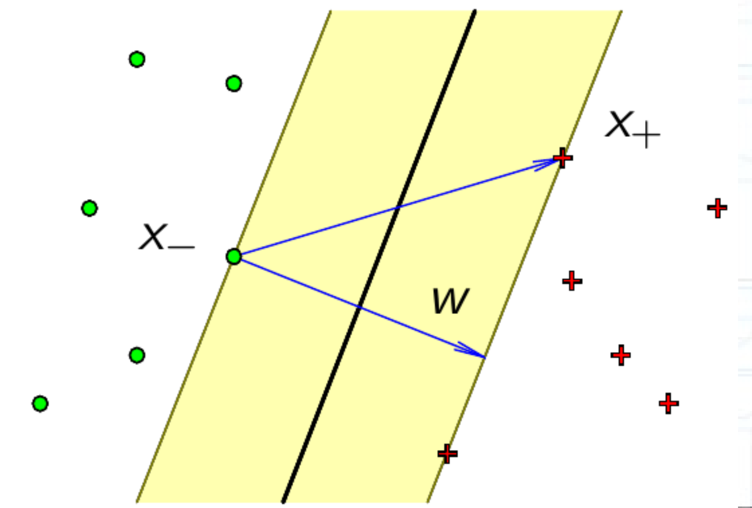
- Допустим, что обучающая выборка линейно разделима:

$$\exists w, w_0: \quad M_i(w, w_0) = y_i\big(\langle w, x_i\rangle - w_0\big) > 0, \quad i = 1, \dots, \ell$$

- w и w0 определены с точностью до множителя $\Rightarrow$ нормируем $\min\limits_{i=1,\dots,\ell} M_i(w, w_0) = 1$

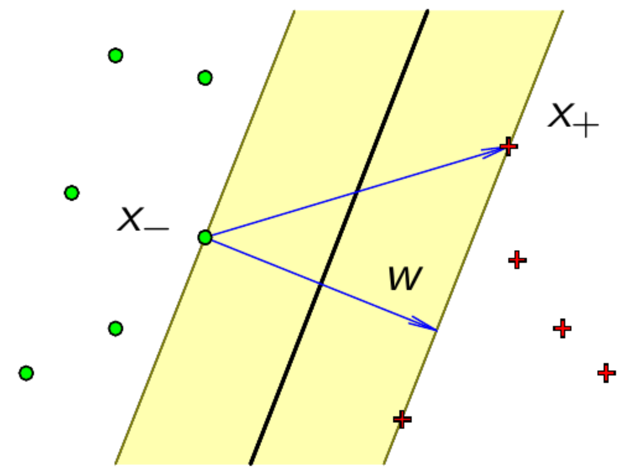
- Ширина полосы:

$$\frac{\langle x_+ - x_-, w\rangle}{\|w\|} = \frac{2}{\|w\|} \to \max$$



Метод опорных векторов для линейно разделимой выборки

$$\begin{cases} \frac{1}{2}\|w\|^2 \to \min\limits_{w,w_0}; \\ M_i(w,w_0) \geqslant 1, \quad i = 1,\dots,\ell \end{cases}$$



Что делать, если выборка не разделима гиперплоскостью?

Случай линейно неразделимой выборки

$$\begin{cases} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}\xi_i \to \min\limits_{w,w_0,\xi}; \\ M_i(w,w_0) \geqslant 1-\xi_i, \quad i=1,\dots,\ell; \\ \xi_i \geqslant 0, \quad i=1,\dots,\ell. \end{cases}$$

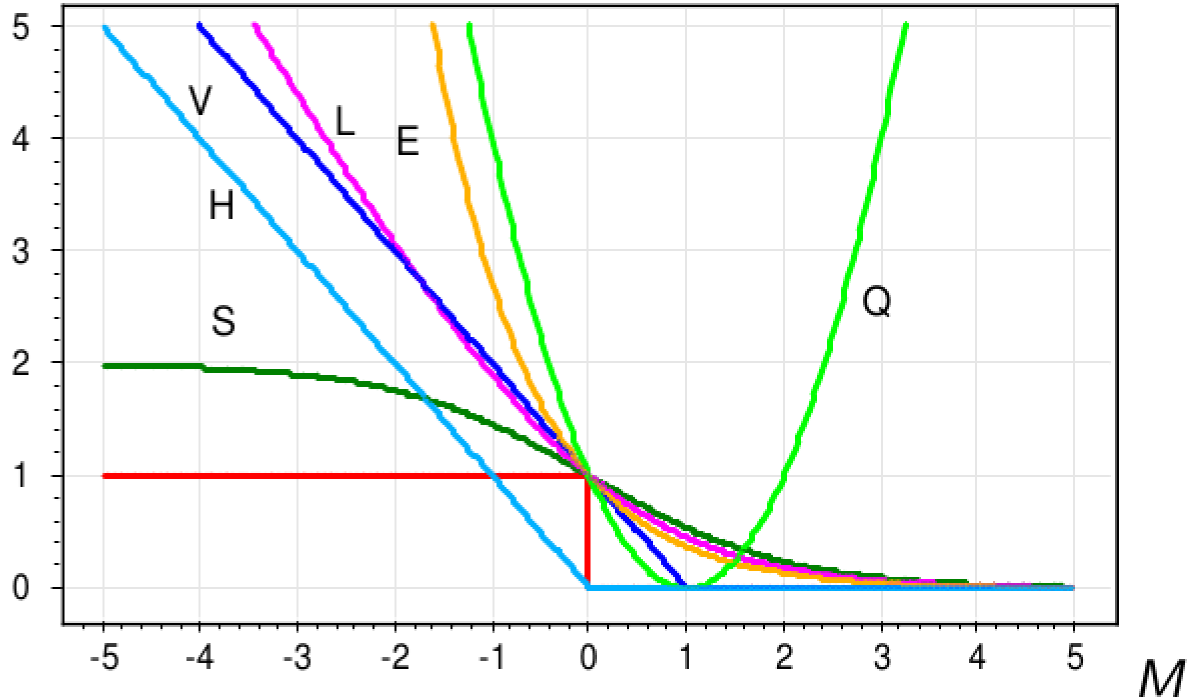
Так как $\xi_i \geqslant 0$ и $\xi_i \geqslant 1-M_i$ **то в силу минимизации суммы** ξ_i

$$\xi_i = (1-M_i)_+$$

Следовательно, наша задача эквивалентна минимизации функционала

$$Q(w,w_0) = \sum_{i=1}^{\ell}\big(1-M_i(w,w_0)\big)_+ + \frac{1}{2C}\|w\|^2 \to \min_{w,w_0}$$

Часто используемые функции потерь



$V(M) = (1 - M)_+$ — кусочно-линейная (SVM);

$H(M) = (-M)_+$ — кусочно-линейная (Hebb's rule);

$L(M) = \log_2(1 + e^{-M})$ — логарифмическая (LR);

$Q(M) = (1 - M)^2$ — квадратичная (FLD);

$S(M) = 2(1 + e^{M})^{-1}$ — сигмоидная (ANN);

$E(M) = e^{-M}$ — экспоненциальная (AdaBoost);

$[M < 0]$ — пороговая функция потерь.

Условия Каруша–Куна–Таккера

Задача математического программирования:

$$\begin{cases} f(x) \to \min\limits_{x}; \\ g_i(x) \leqslant 0, \quad i = 1, \dots, m; \\ h_j(x) = 0, \quad j = 1, \dots, k. \end{cases}$$

Необходимые условия. Если x — точка локального минимума, то существуют множители μ_i, $i = 1, \dots, m$, λ_j, $j = 1, \dots, k$:

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial x} = 0, \quad \mathscr{L}(x; \mu, \lambda) = f(x) + \sum\limits_{i=1}^{m} \mu_i g_i(x) + \sum\limits_{j=1}^{k} \lambda_j h_j(x); \\ g_i(x) \leqslant 0; \; h_j(x) = 0; \; \text{(исходные ограничения)} \\ \mu_i \geqslant 0; \; \text{(двойственные ограничения)} \\ \mu_i g_i(x) = 0; \; \text{(условие дополняющей нежёсткости)} \end{cases}$$

Применение условий ККТ к задаче SVM

Функция Лагранжа: $\mathscr{L}(w, w_0, \xi; \lambda, \eta) =$

$$= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{\ell} \lambda_i \bigl(M_i(w, w_0) - 1\bigr) - \sum_{i=1}^{\ell} \xi_i \bigl(\lambda_i + \eta_i - C\bigr),$$

λ_i — переменные, двойственные к ограничениям $M_i \geqslant 1 - \xi_i$;
η_i — переменные, двойственные к ограничениям $\xi_i \geqslant 0$.

$$\begin{cases} \dfrac{\partial \mathscr{L}}{\partial w} = 0, \quad \dfrac{\partial \mathscr{L}}{\partial w_0} = 0, \quad \dfrac{\partial \mathscr{L}}{\partial \xi} = 0; \\ \xi_i \geqslant 0, \quad \lambda_i \geqslant 0, \quad \eta_i \geqslant 0, \quad i = 1, \ldots, \ell; \\ \lambda_i = 0 \ \text{ либо } \ M_i(w, w_0) = 1 - \xi_i, \quad i = 1, \ldots, \ell; \\ \eta_i = 0 \ \text{ либо } \ \xi_i = 0, \quad i = 1, \ldots, \ell; \end{cases}$$

Необходимые условия седловой точки функции Лагранжа

Функция Лагранжа: $\mathscr{L}(w, w_0, \xi; \lambda, \eta) =$

$$= \frac{1}{2}\|w\|^2 - \sum_{i=1}^{\ell} \lambda_i \bigl(M_i(w, w_0) - 1\bigr) - \sum_{i=1}^{\ell} \xi_i \bigl(\lambda_i + \eta_i - C\bigr),$$

Необходимые условия седловой точки функции Лагранжа:

$$\frac{\partial \mathscr{L}}{\partial w} = w - \sum_{i=1}^{\ell} \lambda_i y_i x_i = 0 \quad \Longrightarrow \quad w = \sum_{i=1}^{\ell} \lambda_i y_i x_i;$$

$$\frac{\partial \mathscr{L}}{\partial w_0} = -\sum_{i=1}^{\ell} \lambda_i y_i = 0 \quad \Longrightarrow \quad \sum_{i=1}^{\ell} \lambda_i y_i = 0;$$

$$\frac{\partial \mathscr{L}}{\partial \xi_i} = -\lambda_i - \eta_i + C = 0 \quad \Longrightarrow \quad \eta_i + \lambda_i = C, \quad i = 1, \dots, \ell.$$

Типы объектов

Типизация объектов:

1. $\lambda_i = 0; \;\; \eta_i = C; \;\; \xi_i = 0; \;\; M_i \geqslant 1.$
   — периферийные (неинформативные) объекты.

2. $0 < \lambda_i < C; \;\; 0 < \eta_i < C; \;\; \xi_i = 0; \;\; M_i = 1.$
   — опорные граничные объекты.

3. $\lambda_i = C; \;\; \eta_i = 0; \;\; \xi_i > 0; \;\; M_i < 1.$
   — опорные-нарушители.

- Объект x_i называется опорным, если $\lambda_i \neq 0$.

Двойственная задача

$$\begin{cases} -\mathscr{L}(\lambda) = -\sum\limits_{i=1}^{\ell} \lambda_i + \frac{1}{2} \sum\limits_{i=1}^{\ell} \sum\limits_{j=1}^{\ell} \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle \;\to\; \min\limits_{\lambda}; \\ 0 \leqslant \lambda_i \leqslant C, \quad i = 1, \ldots, \ell; \\ \sum\limits_{i=1}^{\ell} \lambda_i y_i = 0. \end{cases}$$

Решение прямой задачи выражается через решение двойственной:

$$\begin{cases} w = \sum\limits_{i=1}^{\ell} \lambda_i y_i x_i; \\ w_0 = \langle w, x_i \rangle - y_i, \quad \text{для любого } i\text{: } \lambda_i > 0,\ M_i = 1. \end{cases}$$

Линейный классификатор:

$$a(x) = \operatorname{sign}\Bigl(\sum\limits_{i=1}^{\ell} \lambda_i y_i \langle x_i, x \rangle - w_0 \Bigr).$$

Обучение SVM

1. Find $\boldsymbol{\alpha}^1$ as the initial feasible solution. Set $k = 1$.

2. If $\boldsymbol{\alpha}^k$ is an optimal solution of (1), stop. Otherwise, find a *two-element* working set $B = \{i, j\} \subset \{1, \ldots, l\}$. Define $N \equiv \{1, \ldots, l\} \backslash B$ and $\boldsymbol{\alpha}_B^k$ and $\boldsymbol{\alpha}_N^k$ to be sub-vectors of $\boldsymbol{\alpha}^k$ corresponding to B and N, respectively.

3. Solve the following sub-problem with the variable $\boldsymbol{\alpha}_B$:

$$
\begin{aligned}
\min_{\boldsymbol{\alpha}_B} \quad & \frac{1}{2} \begin{bmatrix} \boldsymbol{\alpha}_B^T & (\boldsymbol{\alpha}_N^k)^T \end{bmatrix} \begin{bmatrix} Q_{BB} & Q_{BN} \\ Q_{NB} & Q_{NN} \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_B \\ \boldsymbol{\alpha}_N^k \end{bmatrix} - \begin{bmatrix} \mathbf{e}_B^T & \mathbf{e}_N^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha}_B \\ \boldsymbol{\alpha}_N^k \end{bmatrix} \\
& = \frac{1}{2} \boldsymbol{\alpha}_B^T Q_{BB} \boldsymbol{\alpha}_B + (-\mathbf{e}_B + Q_{BN} \boldsymbol{\alpha}_N^k)^T \boldsymbol{\alpha}_B + \text{ constant} \\
& = \frac{1}{2} \begin{bmatrix} \alpha_i & \alpha_j \end{bmatrix} \begin{bmatrix} Q_{ii} & Q_{ij} \\ Q_{ij} & Q_{jj} \end{bmatrix} \begin{bmatrix} \alpha_i \\ \alpha_j \end{bmatrix} + (-\mathbf{e}_B + Q_{BN} \boldsymbol{\alpha}_N^k)^T \begin{bmatrix} \alpha_i \\ \alpha_j \end{bmatrix} + \text{ constant} \\
\text{subject to} \quad & 0 \le \alpha_i, \alpha_j \le C, \qquad\qquad (2) \\
& y_i \alpha_i + y_j \alpha_j = -\mathbf{y}_N^T \boldsymbol{\alpha}_N^k,
\end{aligned}
$$

where $\begin{bmatrix} Q_{BB} & Q_{BN} \\ Q_{NB} & Q_{NN} \end{bmatrix}$ is a permutation of the matrix Q.

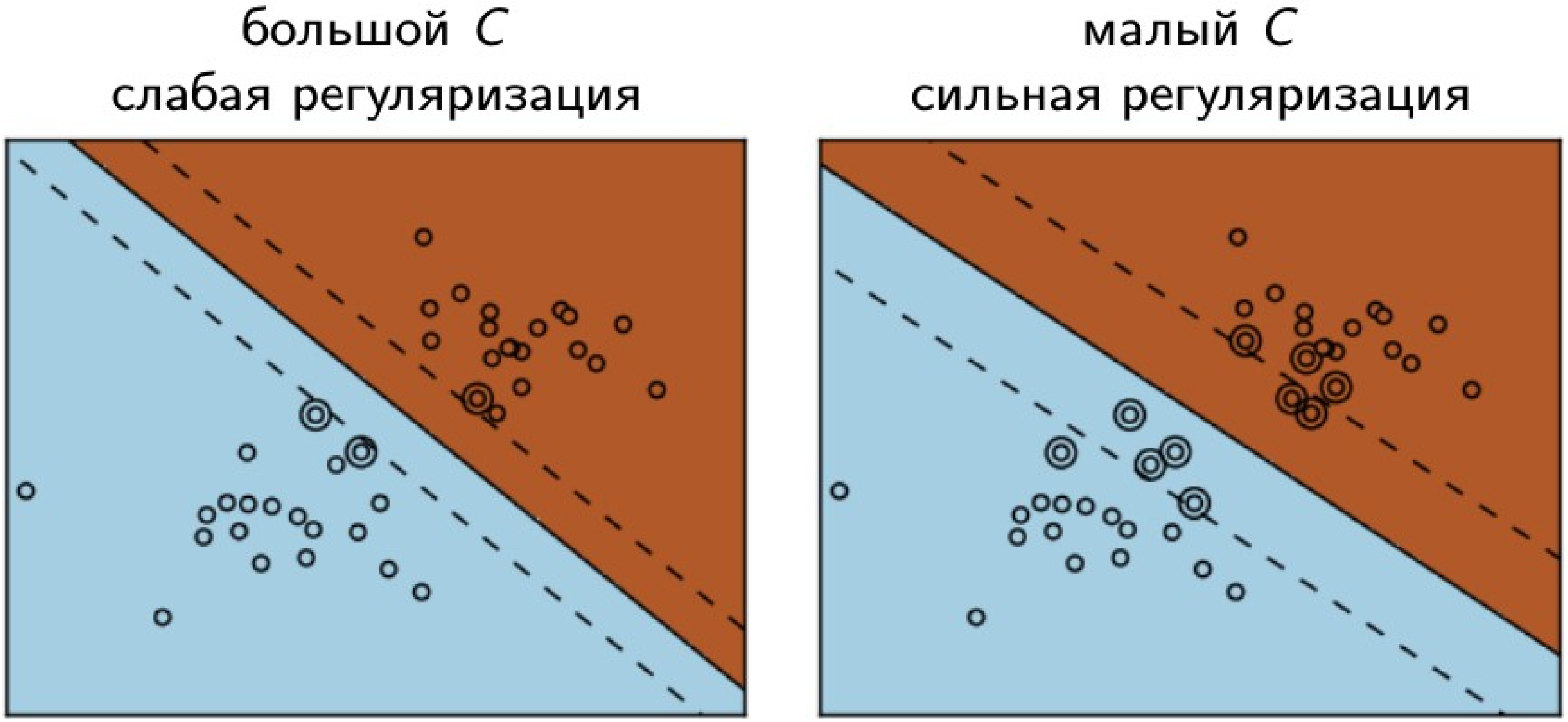
4. Set $\boldsymbol{\alpha}_B^{k+1}$ to be the optimal solution of (2) and $\boldsymbol{\alpha}_N^{k+1} \equiv \boldsymbol{\alpha}_N^k$. Set $k \leftarrow k + 1$ and goto Step 2.

R.-E. Fan, P.-H. Chen, and C.-J. Lin. Working set selection using second order information for training SVM. Journal of Machine Learning Research 6, 1889-1918, 2005

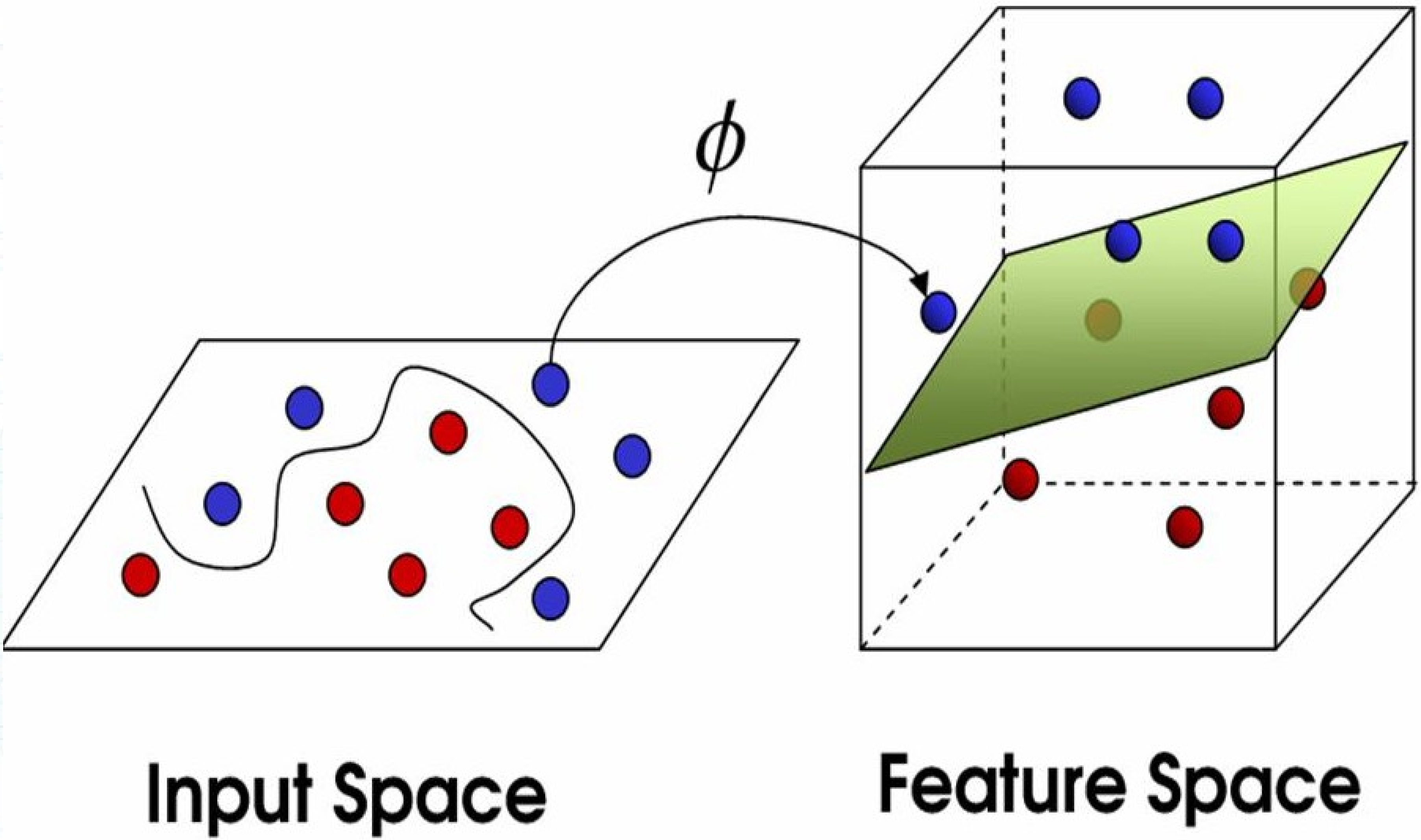
Влияние константы C на решение SVM

SVM — аппроксимация и регуляризация эмпирического риска:

$$\sum_{i=1}^{\ell} \left(1 - M_i(w, w_0)\right)_+ + \frac{1}{2C}\|w\|^2 \to \min_{w,w_0}.$$



Нелинейное обобщение SVM
Расширение пространства



Видео-демострация

Полиномиальные ядра

$$\psi\colon (u_1, u_2) \mapsto \left(u_1^2, u_2^2, \sqrt{2}u_1 u_2\right)$$

$$K(x, x') = \langle \psi(x), \psi(x') \rangle_H$$

$$\begin{aligned} K(u, v) = \langle u, v \rangle^2 &= \langle (u_1, u_2), (v_1, v_2) \rangle^2 = \\ &= (u_1 v_1 + u_2 v_2)^2 = u_1^2 v_1^2 + u_2^2 v_2^2 + 2u_1 v_1 u_2 v_2 = \\ &= \langle (u_1^2, u_2^2, \sqrt{2}u_1 u_2), (v_1^2, v_2^2, \sqrt{2}v_1 v_2) \rangle. \end{aligned}$$

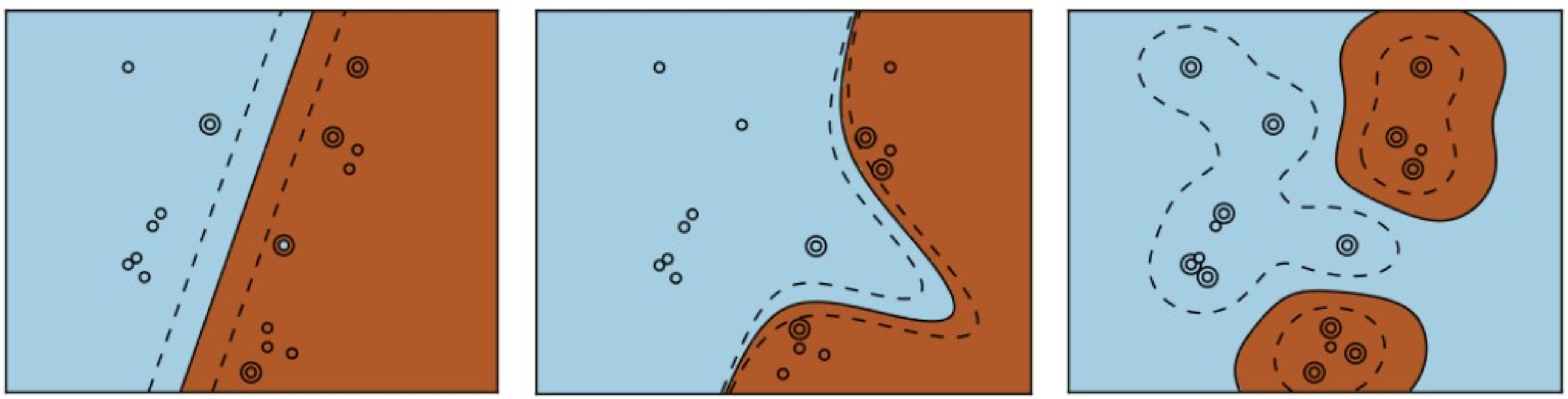
В общем случае: $K(x, x') = \left(\langle x, x' \rangle + 1\right)^d$

Примеры классификаций с различными ядрами

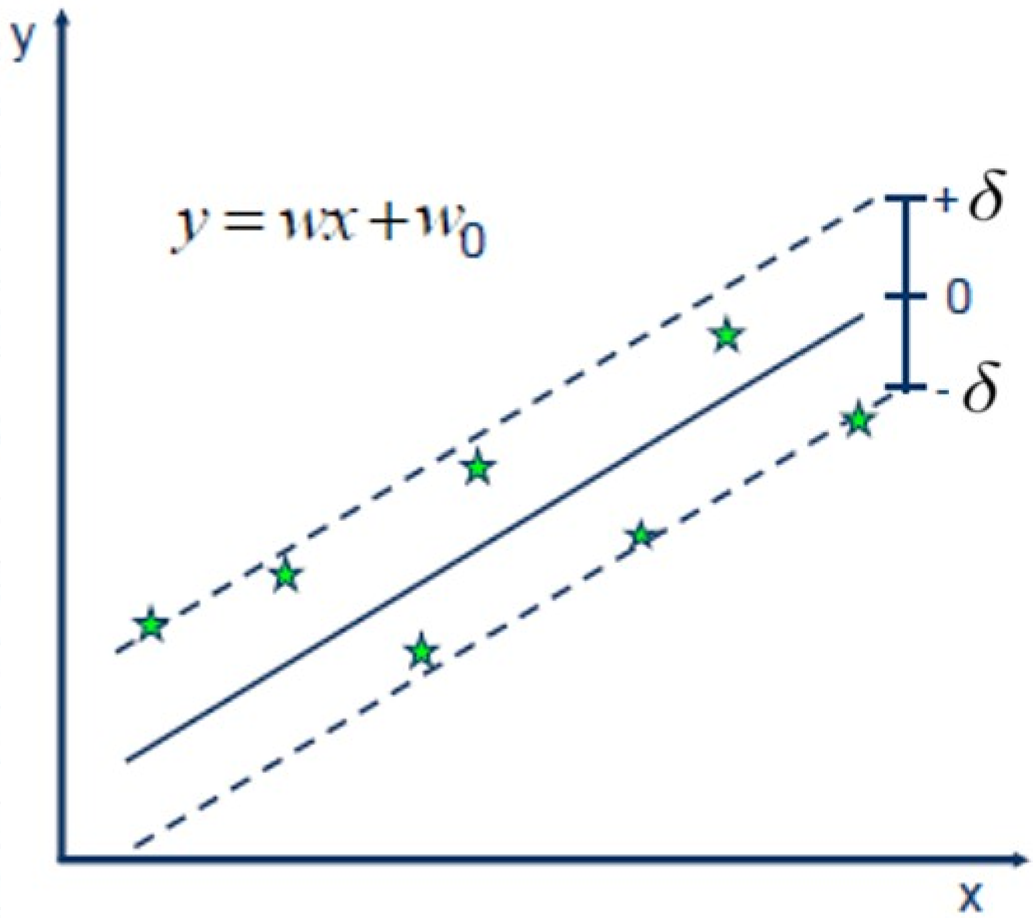
линейное $\langle x, x' \rangle$

полиномиальное $\left(\langle x, x' \rangle + 1\right)^d, \quad d{=}3$

гауссовское (RBF) $\exp\left(-\beta \|x - x'\|^2\right)$



SVM-регрессия



- Задача:

$$\min \frac{1}{2}\|w\|^2$$

- Ограничения

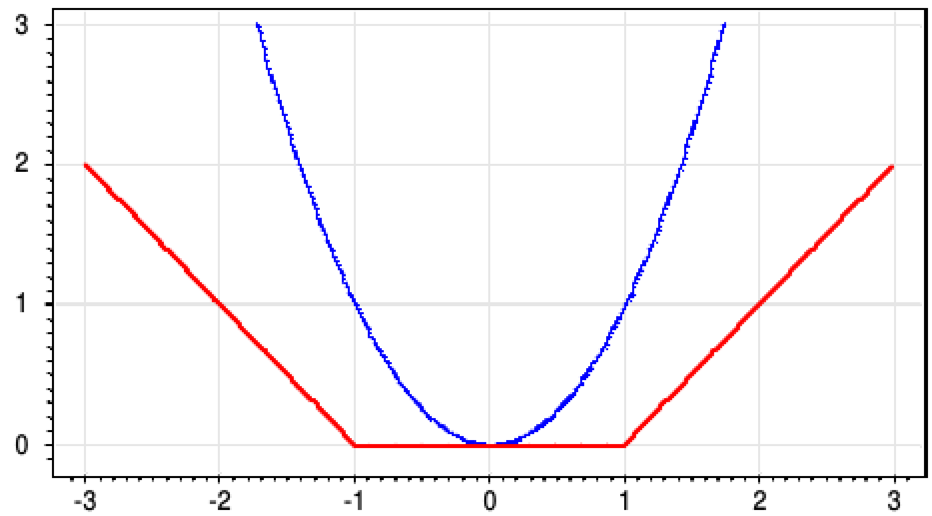
$$y_i - wx_i - w_0 \le \delta$$
$$wx_i + w_0 - y_i \le \delta$$

Эквивалентная постановка задачи

$$\sum_{i=1}^{\ell} \left(|\langle w, x_i\rangle - w_0 - y_i| - \delta\right)_+ + \frac{1}{2C}\|w\|^2 \to \min_{w, w_0}$$

Сравнение с обычной регрессией (МНК)

Функция потерь: $\mathscr{L}(\varepsilon) = \left(|\varepsilon| - \delta\right)_+$ в сравнении с $\mathscr{L}(\varepsilon) = \varepsilon^2$



Задача решается путём замены переменных и сведения к задаче квадратичного программирования

Решение задачи оптимизации

Замена переменных:

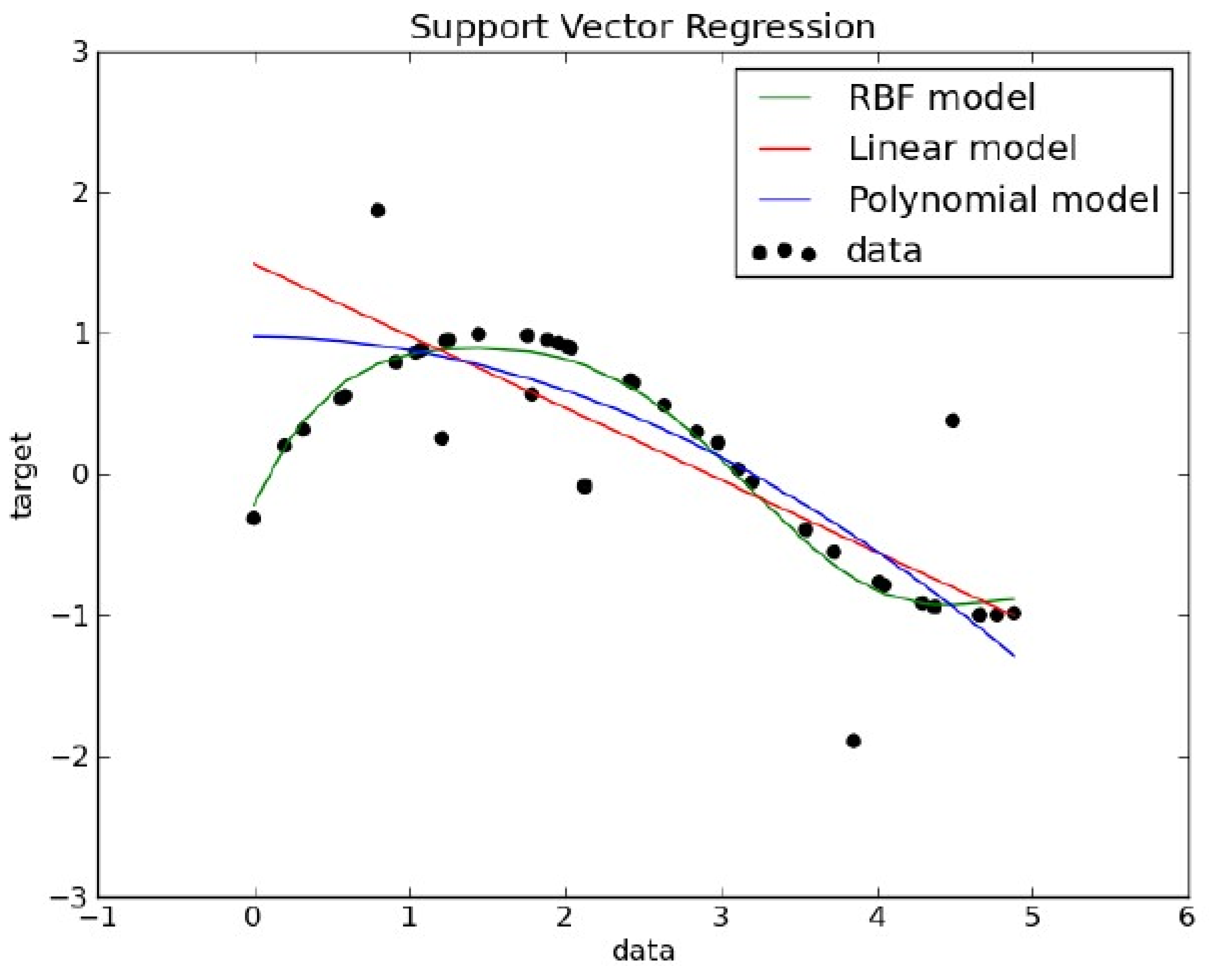
$$\xi_i^+ = (\langle w, x_i\rangle - w_0 - y_i - \delta)_+ \,;$$
$$\xi_i^- = (-\langle w, x_i\rangle + w_0 + y_i - \delta)_+ \,;$$

$$\begin{cases} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}(\xi_i^+ + \xi_i^-) \to \min\limits_{w, w_0, \xi^+, \xi^-}; \\ y_i - \delta - \xi_i^- \leqslant \langle w, x_i\rangle - w_0 \leqslant y_i + \delta + \xi_i^+, \quad i = 1, \dots, \ell; \\ \xi_i^- \geqslant 0, \quad \xi_i^+ \geqslant 0, \quad i = 1, \dots, \ell. \end{cases}$$

Это задача квадратичного программирования с линейными ограничениями-неравенствами, решается также сведением к двойственной задаче.

Сравнение

- Сравнение SVM-регрессии с гауссовским (RBF) ядром, линейной и полиномиальной регрессией:



1-norm SVM (LASSO SVM)

Аппроксимация эмпирического риска с L_1 -регуляризацией:

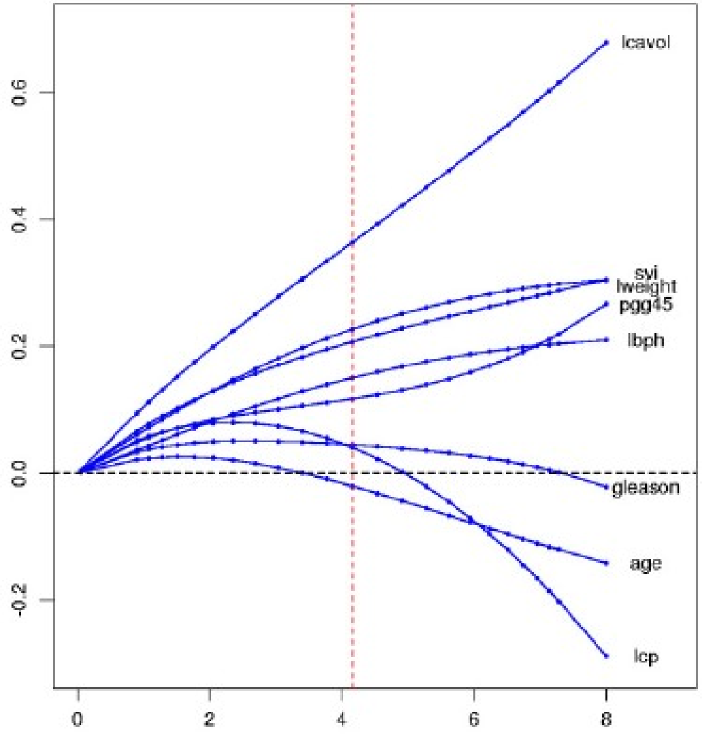
$$\sum_{i=1}^{\ell} \left(1 - M_i(w, w_0)\right)_+ + \mu \sum_{j=1}^{n} |w_j| \to \min_{w, w_0}$$

Отбор признаков с параметром селективности μ: чем больше μ, тем меньше признаков останется

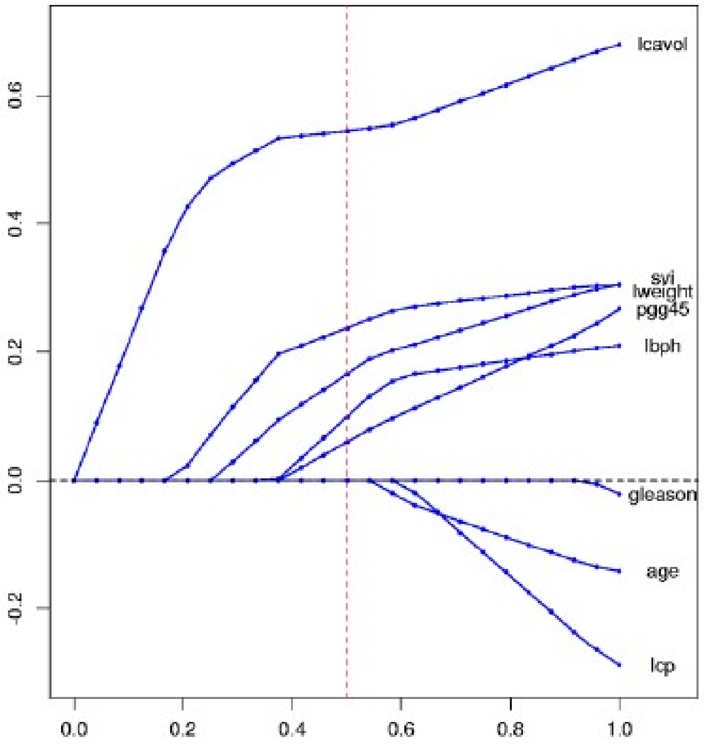
Сравнение L_2 и L_1 регуляризации

Зависимость весов w_j от коэффициента $\frac{1}{\mu}$

L_2 регуляризатор: $\mu \sum_j w_j^2$



L_1 регуляризатор: $\mu \sum_j |w_j|$



Задача из UCI: prostate cancer (диагностика рака)

Doubly Regularized SVM (Elastic Net SVM)

$$C\sum_{i=1}^{\ell}\bigl(1 - M_i(w, w_0)\bigr)_+ + \mu\sum_{j=1}^{n}|w_j| + \frac{1}{2}\sum_{j=1}^{n}w_j^2 \;\to\; \min_{w,w_0}$$

Elastic Net менее жёстко отбирает признаки.

Зависимости весов w_j от коэффициента $\log\frac{1}{\mu}$: