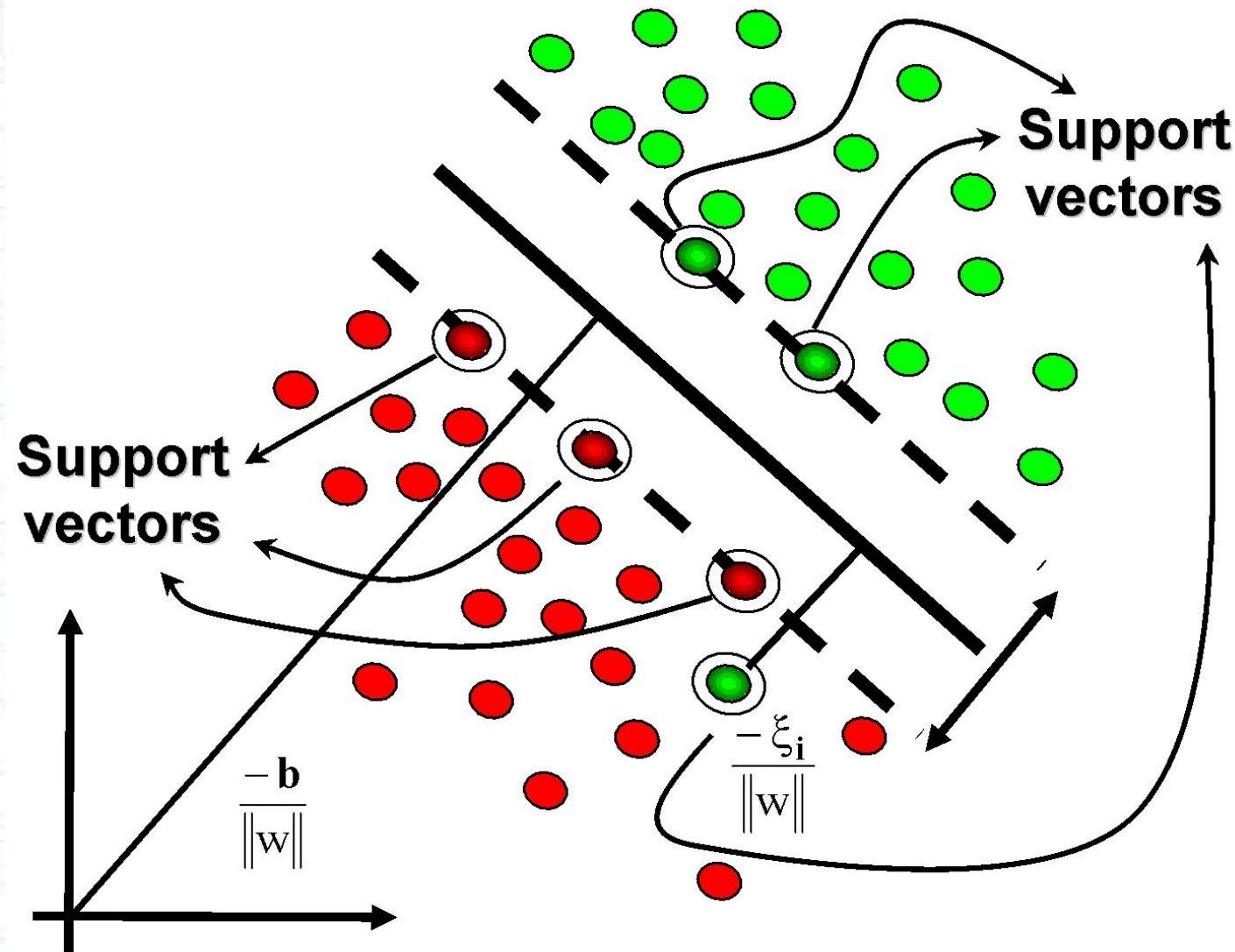


Машинное обучение

Метод опорных векторов (SVM)



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:

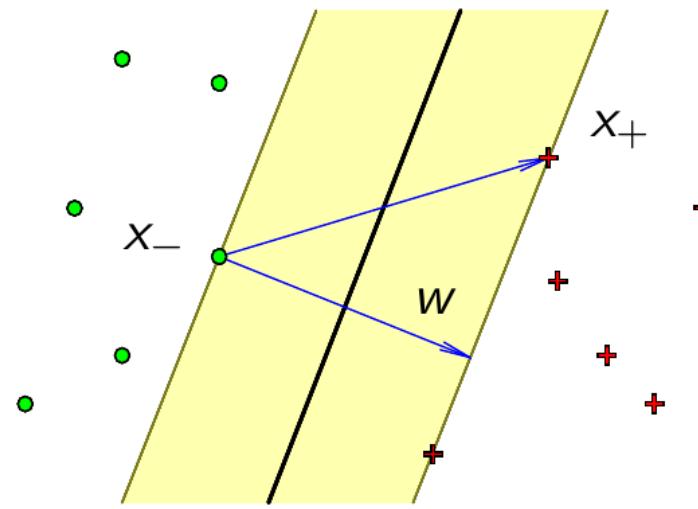
$$a(x, w) = \text{sign}(\langle w, x \rangle - w_0)$$

- Допустим, что обучающая выборка линейно разделима:

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

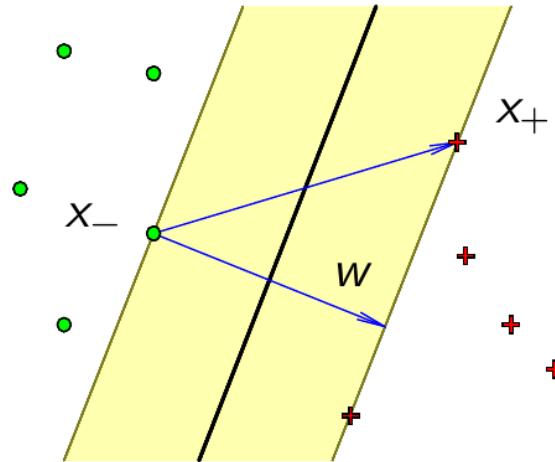
- w и w_0 определены с точностью до множителя \Rightarrow нормируем $\min_{i=1, \dots, \ell} M_i(w, w_0) = 1$
- Ширина полосы:

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



Метод опорных векторов для линейно разделимой выборки

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



Что делать, если выборка
не разделима гиперплоскостью?

Случай линейно неразделимой выборки

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

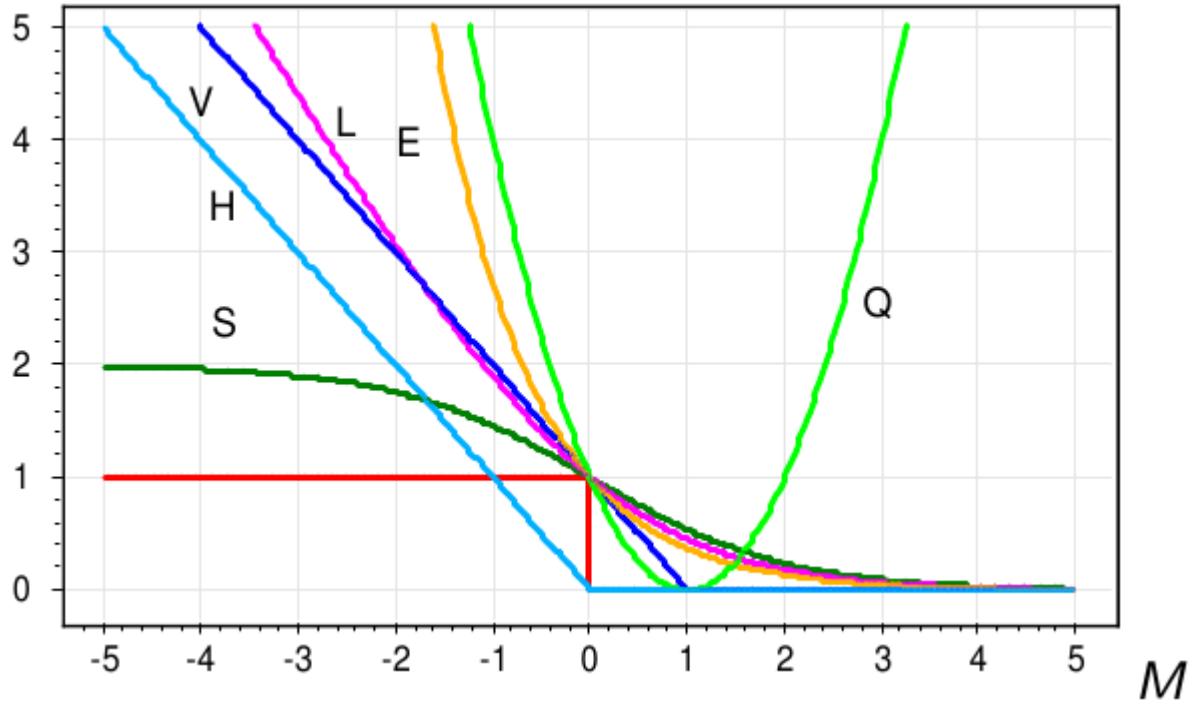
Так как $\xi_i \geq 0$ и $\xi_i \geq 1 - M_i$, то в силу минимизации суммы ξ_i

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

Следовательно, наша задача эквивалентна минимизации функционала

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

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:

$$\begin{cases} f(x) \rightarrow \min_x; \\ g_i(x) \leq 0, \quad i = 1, \dots, m; \\ h_j(x) = 0, \quad j = 1, \dots, k. \end{cases}$$

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

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial x} = 0, \quad \mathcal{L}(x; \mu, \lambda) = f(x) + \sum_{i=1}^m \mu_i g_i(x) + \sum_{j=1}^k \lambda_j h_j(x); \\ g_i(x) \leq 0; \quad h_j(x) = 0; \quad (\text{исходные ограничения}) \\ \mu_i \geq 0; \quad (\text{двойственные ограничения}) \\ \mu_i g_i(x) = 0; \quad (\text{условие дополняющей нежёсткости}) \end{cases}$$

Применение условий ККТ к задаче SVM

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

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

λ_i — переменные, двойственные к ограничениям $M_i \geq 1 - \xi_i$;

η_i — переменные, двойственные к ограничениям $\xi_i \geq 0$.

$$\begin{cases} \frac{\partial \mathcal{L}}{\partial w} = 0, \quad \frac{\partial \mathcal{L}}{\partial w_0} = 0, \quad \frac{\partial \mathcal{L}}{\partial \xi} = 0; \\ \xi_i \geq 0, \quad \lambda_i \geq 0, \quad \eta_i \geq 0, \quad i = 1, \dots, \ell; \\ \lambda_i = 0 \text{ либо } M_i(w, w_0) = 1 - \xi_i, \quad i = 1, \dots, \ell; \\ \eta_i = 0 \text{ либо } \xi_i = 0, \quad i = 1, \dots, \ell; \end{cases}$$

Необходимые условия седловой точки функции Лагранжа

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

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

Необходимые условия седловой точки функции Лагранжа:

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

$$\frac{\partial \mathcal{L}}{\partial w_0} = - \sum_{i=1}^{\ell} \lambda_i y_i = 0 \implies \sum_{i=1}^{\ell} \lambda_i y_i = 0;$$

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

Типы объектов

Типизация объектов:

1. $\lambda_i = 0; \eta_i = C; \xi_i = 0; M_i \geq 1.$
— периферийные (неинформационные) 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

$$\begin{cases} -\mathcal{L}(\lambda) = -\sum_{i=1}^{\ell} \lambda_i + \frac{1}{2} \sum_{i=1}^{\ell} \sum_{j=1}^{\ell} \lambda_i \lambda_j y_i y_j \langle x_i, x_j \rangle \rightarrow \min_{\lambda}; \\ 0 \leq \lambda_i \leq C, \quad i = 1, \dots, \ell; \\ \sum_{i=1}^{\ell} \lambda_i y_i = 0. \end{cases}$$

Решение прямой задачи выражается через решение двойственной:

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

Линейный классификатор:

$$a(x) = \operatorname{sign}\left(\sum_{i=1}^{\ell} \lambda_i y_i \langle x_i, x \rangle - w_0 \right).$$

Преимущество перехода к двойственной задаче

- Двойственная задача не зависит от размерности пространства X
- Размерность вектора λ совпадает с числом объектов, но подавляющее большинство координат равны или 0 или C

Обучение SVM

1. Find α^1 as the initial feasible solution. Set $k = 1$.
2. If α^k is an optimal solution of (1), stop. Otherwise, find a *two-element* working set $B = \{i, j\} \subset \{1, \dots, l\}$. Define $N \equiv \{1, \dots, l\} \setminus B$ and α_B^k and α_N^k to be sub-vectors of α^k corresponding to B and N , respectively.
3. Solve the following sub-problem with the variable α_B :

$$\begin{aligned}
 \min_{\alpha_B} \quad & \frac{1}{2} [\alpha_B^T \quad (\alpha_N^k)^T] \begin{bmatrix} Q_{BB} & Q_{BN} \\ Q_{NB} & Q_{NN} \end{bmatrix} \begin{bmatrix} \alpha_B \\ \alpha_N^k \end{bmatrix} - [\mathbf{e}_B^T \quad \mathbf{e}_N^T] \begin{bmatrix} \alpha_B \\ \alpha_N^k \end{bmatrix} \\
 &= \frac{1}{2} \alpha_B^T Q_{BB} \alpha_B + (-\mathbf{e}_B + Q_{BN} \alpha_N^k)^T \alpha_B + \text{constant} \\
 &= \frac{1}{2} [\alpha_i \quad \alpha_j] \begin{bmatrix} Q_{ii} & Q_{ij} \\ Q_{ij} & Q_{jj} \end{bmatrix} \begin{bmatrix} \alpha_i \\ \alpha_j \end{bmatrix} + (-\mathbf{e}_B + Q_{BN} \alpha_N^k)^T \begin{bmatrix} \alpha_i \\ \alpha_j \end{bmatrix} + \text{constant}
 \end{aligned} \tag{2}$$

subject to $0 \leq \alpha_i, \alpha_j \leq C,$
 $y_i \alpha_i + y_j \alpha_j = -\mathbf{y}_N^T \alpha_N^k,$

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

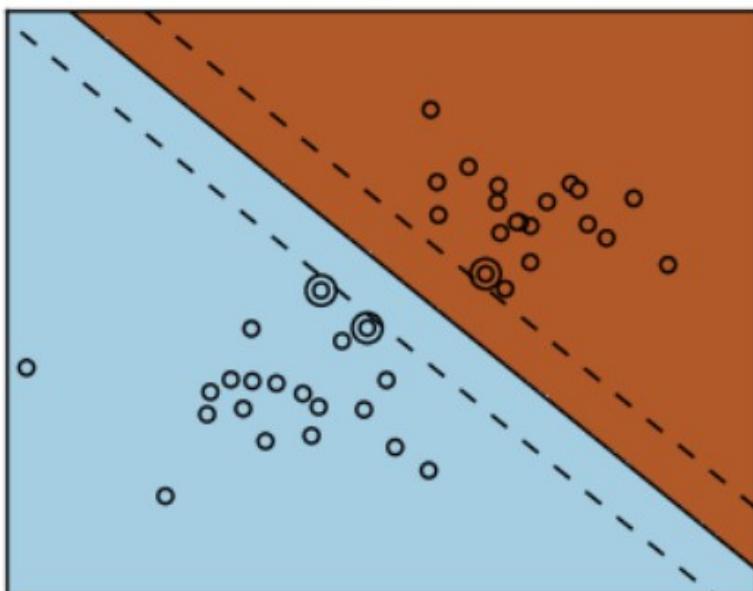
4. Set α_B^{k+1} to be the optimal solution of (2) and $\alpha_N^{k+1} \equiv \alpha_N^k$. Set $k \leftarrow k + 1$ and goto Step 2.

Влияние константы С на решение SVM

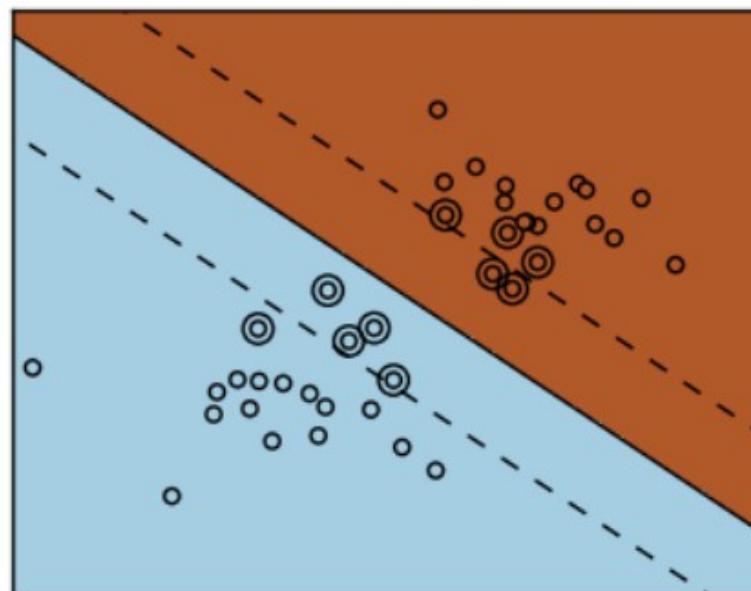
SVM — аппроксимация и регуляризация эмпирического риска:

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

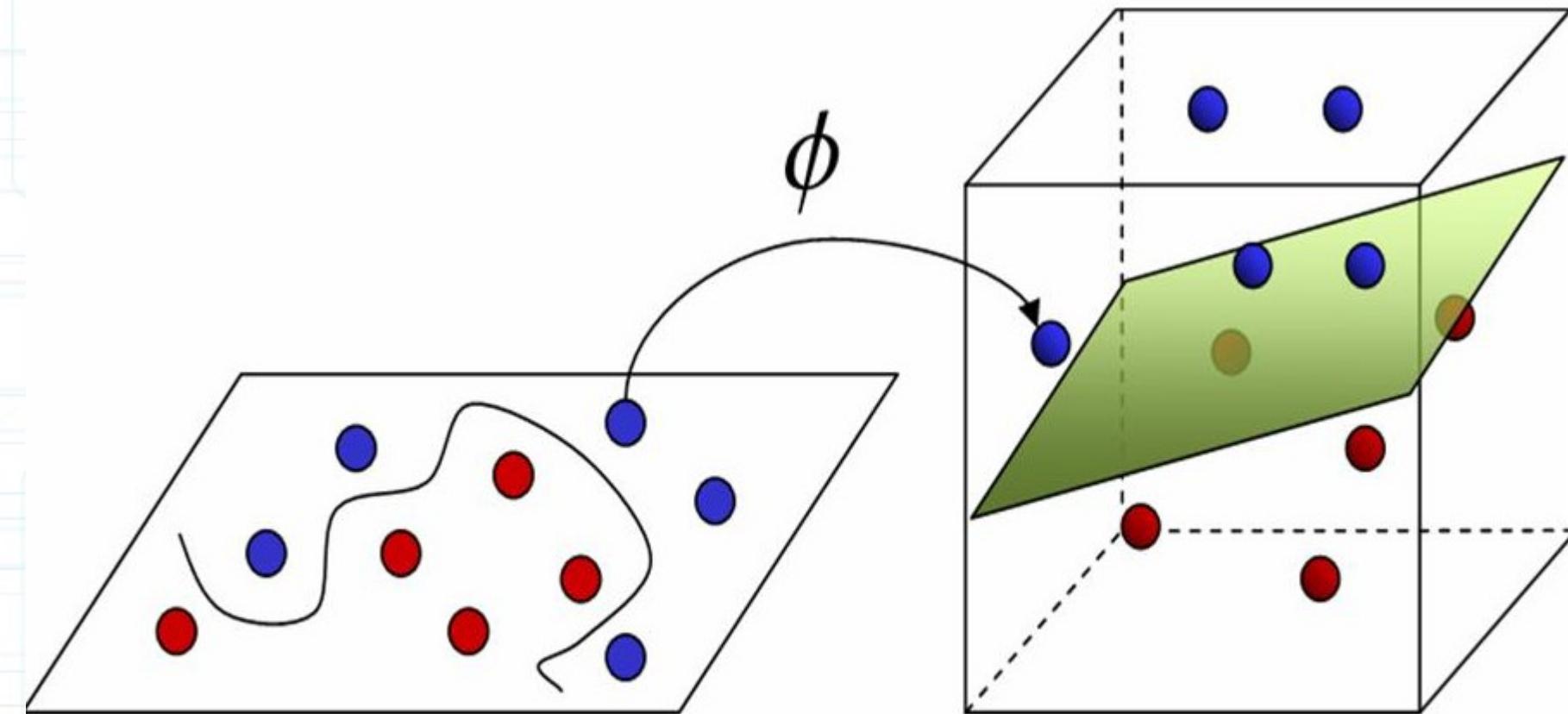
большой С
слабая регуляризация



малый С
сильная регуляризация



Нелинейное обобщение SVM Расширение пространства

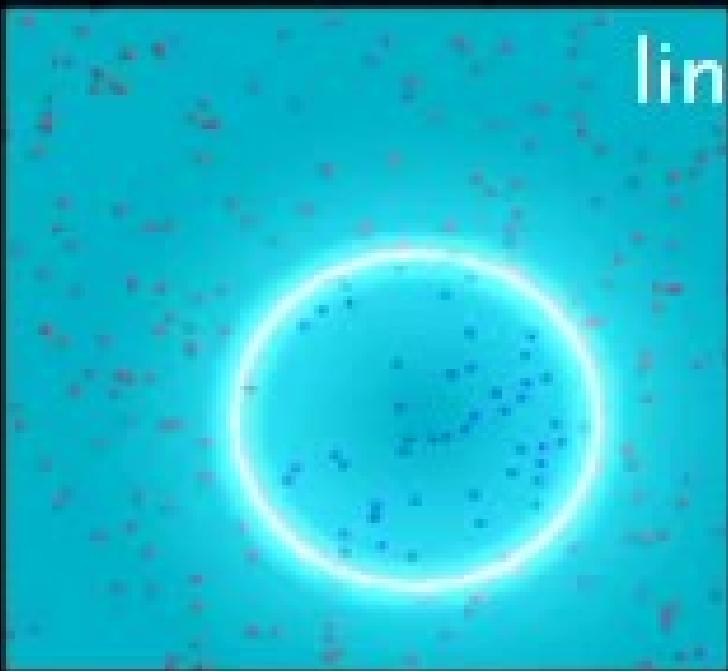


Input Space

Feature Space

Видео-демонстрация

The blue/red
dots are not
linearly separable



Полиномиальные ядра

$$\psi: (u_1, u_2) \mapsto (u_1^2, u_2^2, \sqrt{2}u_1u_2)$$

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

$$\begin{aligned} K(u, v) &= \langle u, v \rangle^2 = \langle (u_1, u_2), (v_1, v_2) \rangle^2 = \\ &= (u_1 v_1 + u_2 v_2)^2 = u_1^2 v_1^2 + u_2^2 v_2^2 + 2u_1 v_1 u_2 v_2 = \\ &= \langle (u_1^2, u_2^2, \sqrt{2}u_1u_2), (v_1^2, v_2^2, \sqrt{2}v_1v_2) \rangle. \end{aligned}$$

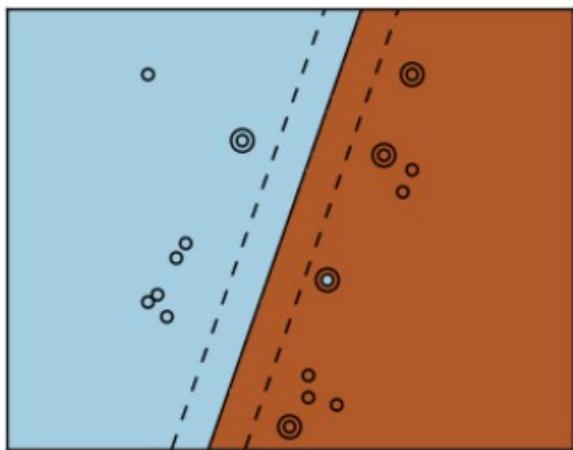
В общем случае:

$$K(x, x') = (\langle x, x' \rangle + 1)^d$$

Примеры классификаций с различными ядрами

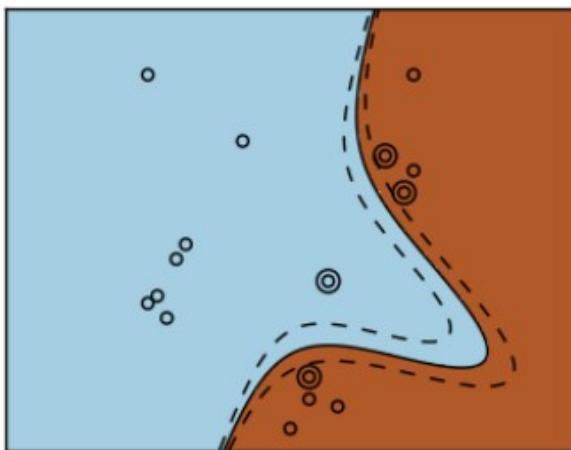
линейное

$$\langle x, x' \rangle$$



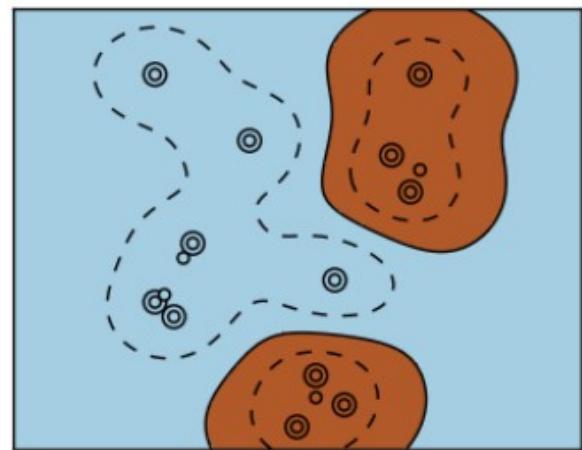
полиномиальное

$$(\langle x, x' \rangle + 1)^d, \quad d=3$$

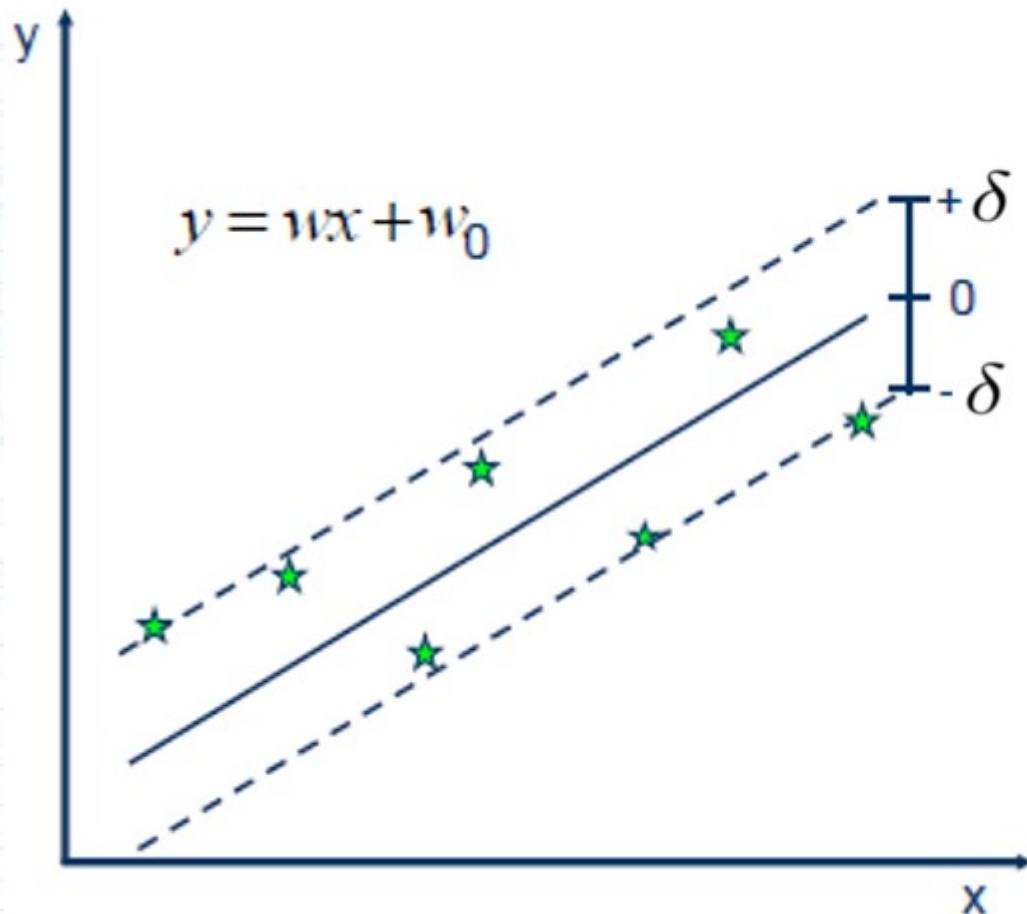


гауссовское (RBF)

$$\exp(-\beta \|x - x'\|^2)$$



SVM-регрессия



- Задача:

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

- Ограничения

$$y_i - w \cdot x_i - w_0 \leq \delta$$

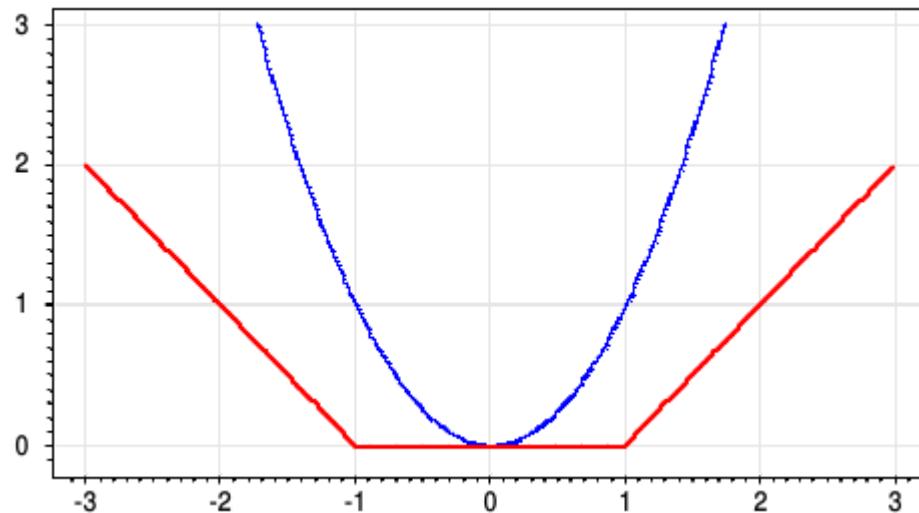
$$w \cdot x_i + w_0 - y_i \leq \delta$$

Эквивалентная постановка задачи

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

Сравнение с обычной регрессией (МНК)

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



Задача решается путём замены переменных и сведения к задаче квадратичного программирования

Решение задачи оптимизации

Замена переменных:

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

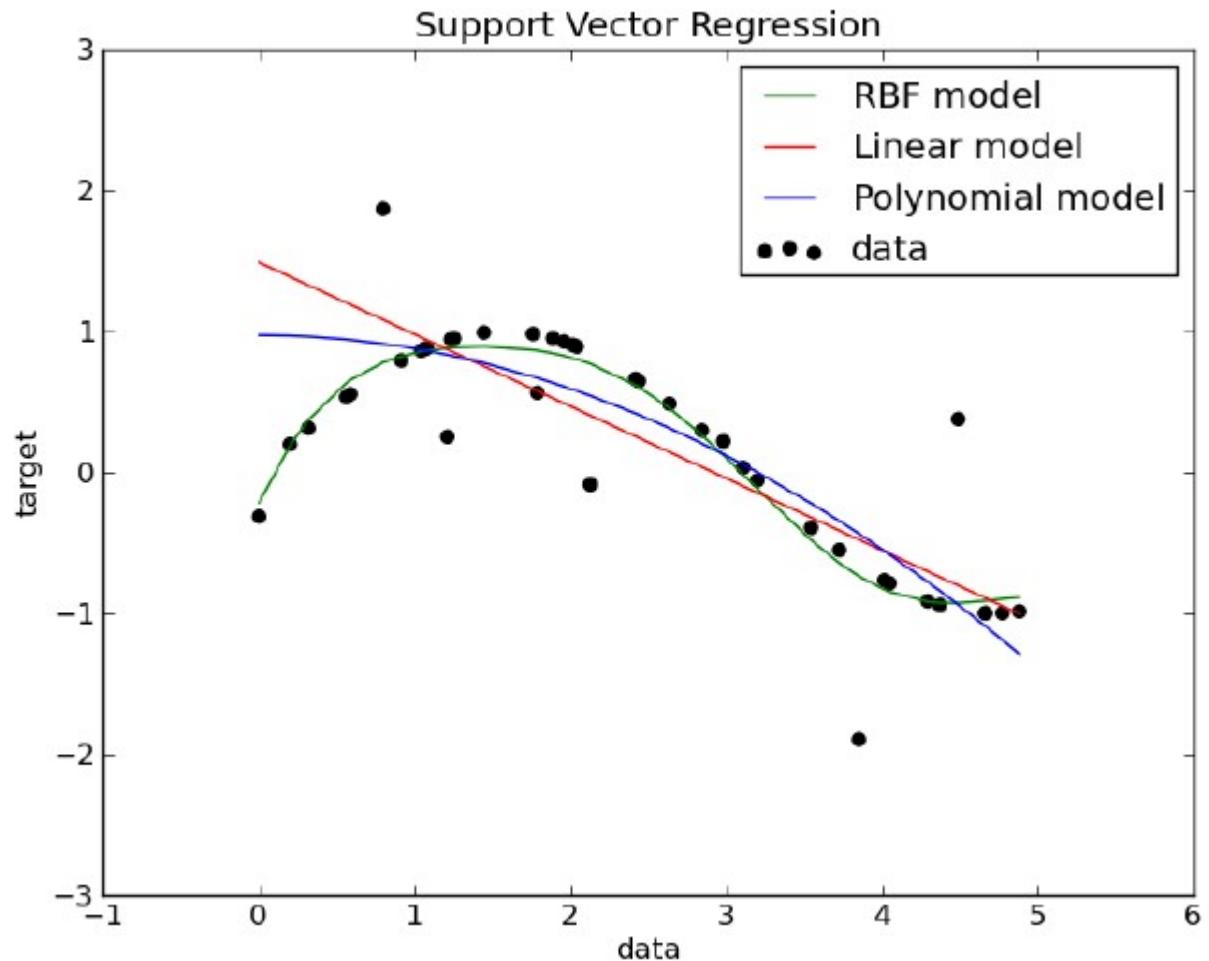
$$\xi_i^- = (-\langle w, x_i \rangle + w_0 + y_i - \delta)_+;$$

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

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)

Апроксимация эмпирического риска с L_1 -регуляризацией:

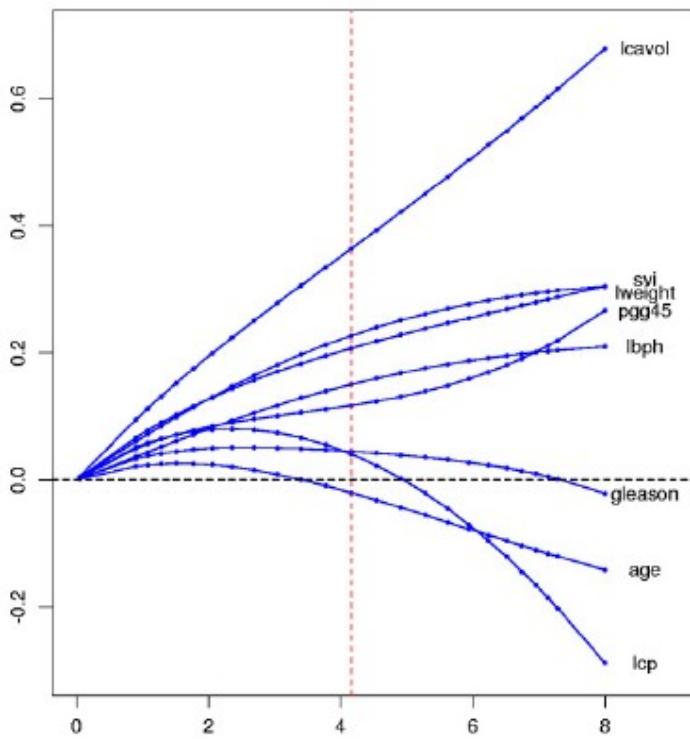
$$\sum_{i=1}^{\ell} (1 - M_i(w, w_0))_+ + \mu \sum_{j=1}^n |w_j| \rightarrow \min_{w, w_0}$$

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

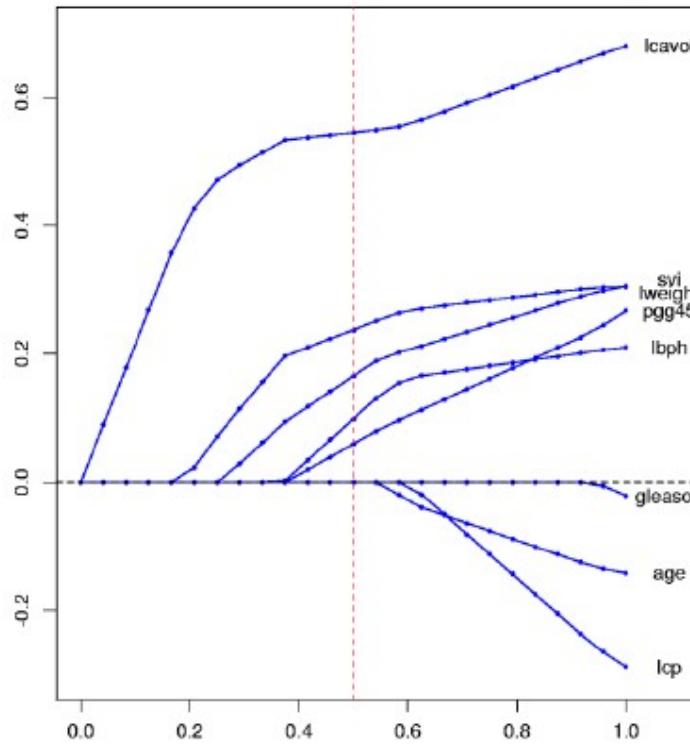
Сравнение L_2 и L_1 регуляризации

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)

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

Elastic Net менее жёстко отбирает признаки.

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

