Метод опорных векторов

Сбертех, МФТИ

Поиск оптимальной модели в задаче ML

- Чем более сложные модели мы выбираем для нашей выборки, тем меньше ошибка оценки
- Чем более сложные модели мы выбираем для нашей выборки, тем больше ошибка обучения

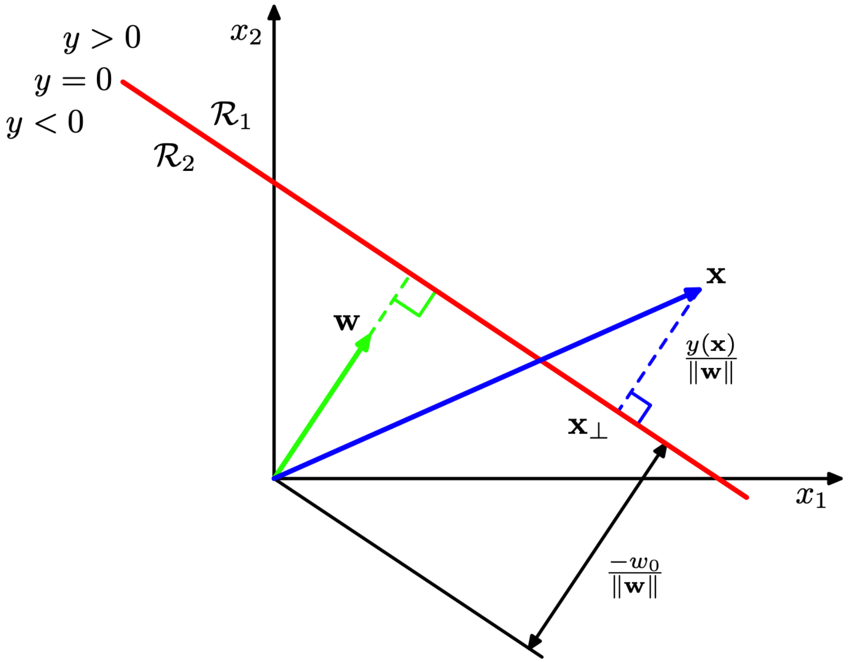
Как найти модель с наилучшими параметрами, сбалансировав между ошибкой обучения и ошибкой оценки?

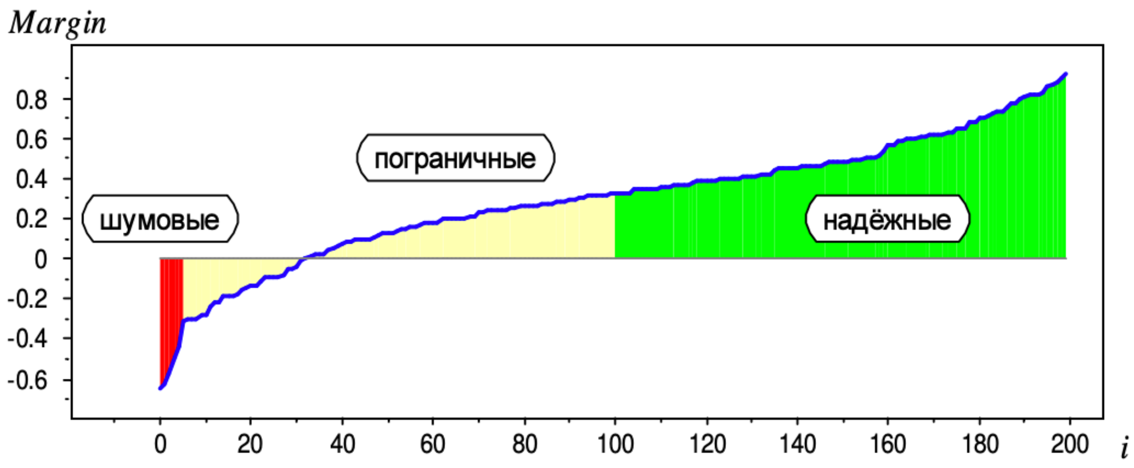
Модели классификации

Линейные модели классификации можно представить как разделяющую гиперплоскость размерностью $(D-1)$ в пространстве D

$$y(x) = w_0 + w_1 \mathrm{x}_1 + w_2 \mathrm{x}_2 = w^T x$$

- w – сдвиг плоскости относительно начала координат
- $w_1, \dots, w_n$ - направляющий вектор плоскости
- Расстояние от точки до разделяющей гиперплоскости(обозначается как margin) - $\rho = \frac{w^T x}{||w||}$



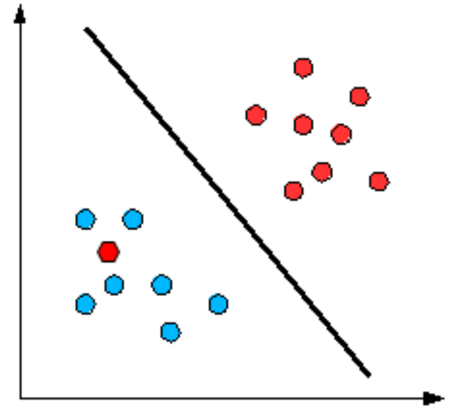


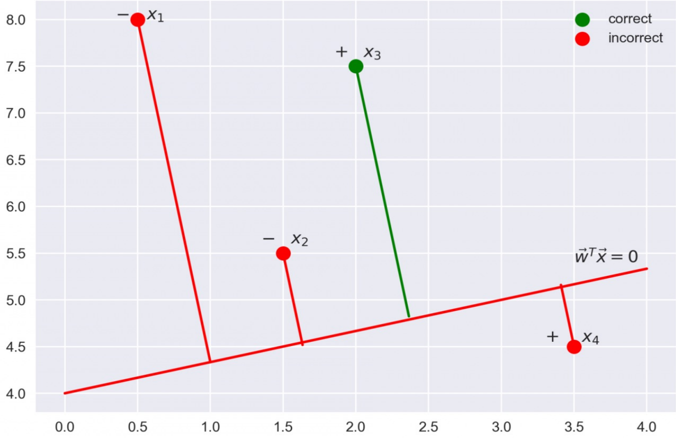
С помощью margin мы можем ранжировать объекты

Есть задача классификации – $y_i \in \{-1; +1\}$. Тогда расстояние между разделяющей прямой и задается уравнением

$$M_i = y_i(w^T x - w_0)$$

- **Если $M_i > 0$ - значит предсказание верное**
  1) $w^T x - w_0$ положительное, y_i = +1
  2) $w^T x - w_0$ отрицательное, y_i = -1
- **Если $M_i < 0$ – предсказание ошибочное**
  1) Говорим что $w^T x - w_0$ > 0, а на самом деле лейбл -1
  2) Говорим что $w^T x - w_0$ < 0, а на самом деле лейбл $+1$
- Чем больше M_i– тем более уверены мы в своем решении
- **Задача - максимизировать расстояние от разделяющей гиперплоскости размерности $(D-1)$ до каждой точки из обучающей выборки в пространстве - $\sum M_i \to max$**



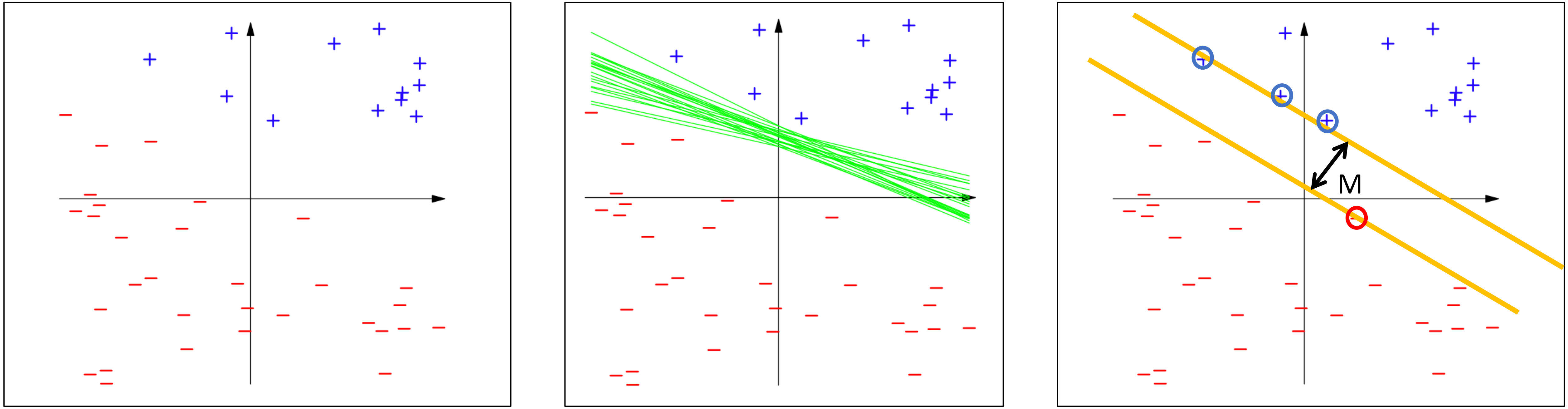


Положительный класс – $\{+1\}$, отрицательный класс – $\{-1\}$

$$f(x, w) = w^T x - w_0 = [0.75 \quad 1]\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 3 = 0.75x_1 + x_2 - 3$$

- Если $f(x_i, w) > 0$, значит предсказываемый объект над разделяющей прямой
  - Если $y_i f(x_i, w) > 0$ – знак класса такой же как и предсказание, значит решение верное(x_3)
- Если $f(x_i) < 0$, значит предсказываемый объект под разделяющей прямой
  - Если $y_i f(x_i, w) < 0$ – знак класса не соответствует знаку предсказания, значит решение неверное(x_1, x_2, x_4)
- Чем больше расстояние точки от прямой – тем выше уверенность

Разделяющая прямая



Можно построить много разных разделяющих прямых – но какая из их самая оптимальная?

Самая оптимальная – та, которая максимизирует зазор(M) между классами

Построение функции потерь для максимизации зазора

1) Модель классификации

$$a_\theta = sign(\theta^T x - \theta_0)$$

2) Ошибка минимальна при минимальном количестве неверно определённых объектов

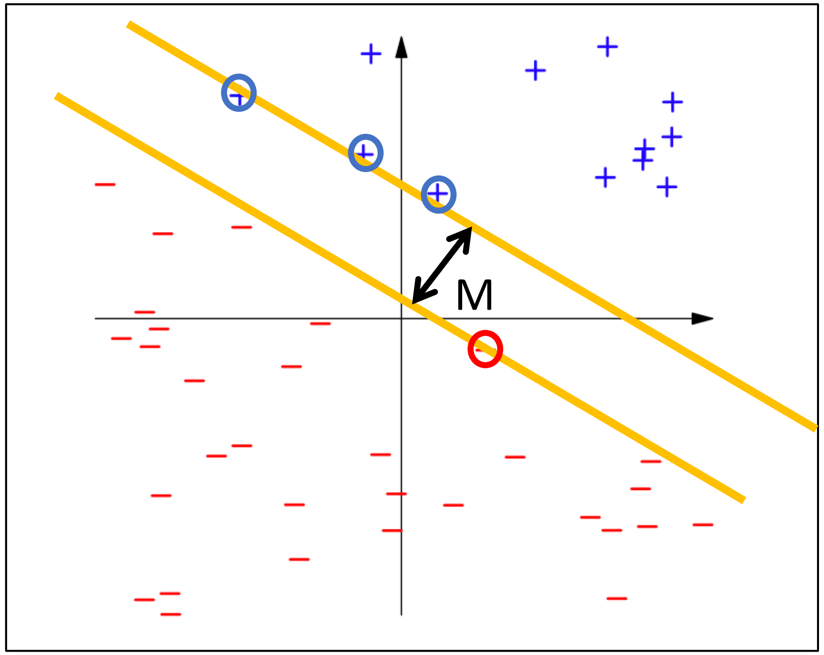
$$\sum_{i=1}^{N}[M_i < 0] \leq \frac{1}{n}\sum_{i=1}^{N} L_\theta(y_i, x_i)$$

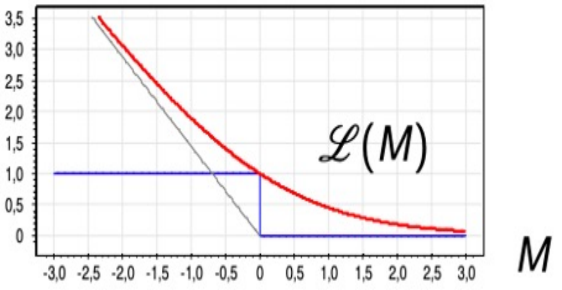
Как построить функцию ошибки L_θ?

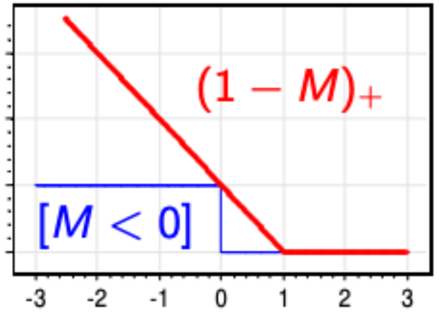
- Хотим штрафовать модель если $M_i = y_i(\theta^T x - \theta_0) \leq 1$ – значит у нас неуверенные/неправильные предсказания
- Хотим штрафовать модель за большие веса для борьбы с мультиколлинеарностью и переобучением

$$L_\theta(y_i, x_i) = \big(1 - y_i(\theta^T x - \theta_0)\big) + \lambda||\theta||_2$$
$$= \boxed{\max(0, 1 - M_i)} + \lambda||\theta||_2$$

Hinge loss







$$(1 - M)_+ = \max(0, 1 - M)$$

Hard-margin SVM

Пусть есть некоторый классификатор $a(x) = sign(\theta^T x - \theta_0)$ и некоторая линейно разделимая выборка x_i, y_i с помощью параметров θ:

$$\forall x_i, y_i : \; M_\theta(x_i, y_i) > 0$$

При этом все $M_\theta(x_i, y_i)$ нормированы, так что $\min M_\theta(x_i, y_i) = 1$

Уравнение разделяющей полосы:

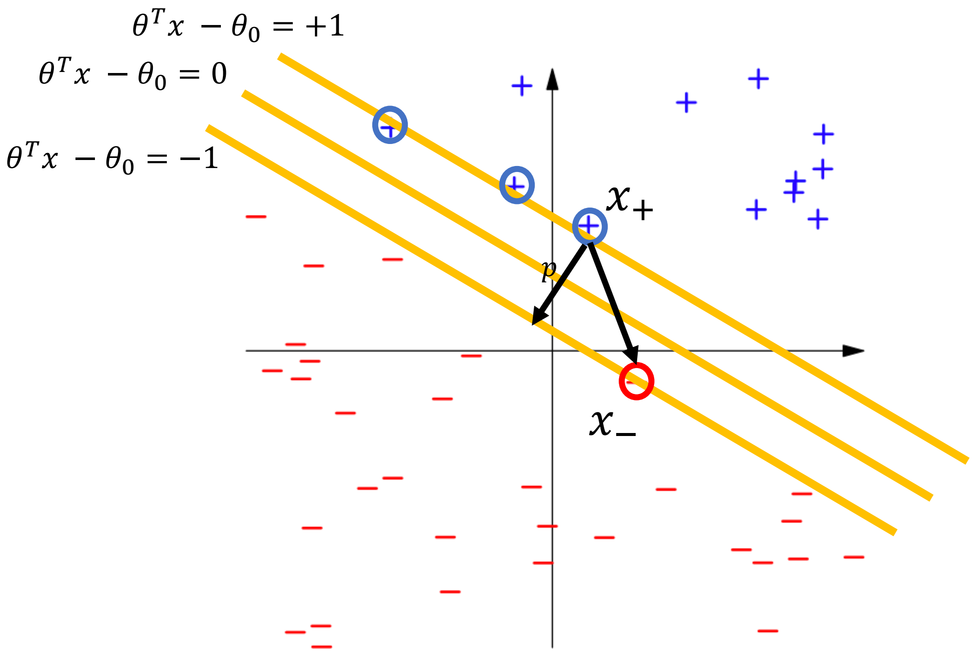
$$\theta^T x - \theta_0 = 0$$

Для объектов x_+ и x_-:

$$\begin{cases} \theta^T x_+ - \theta_0 = +1 \\ \theta^T x_- - \theta_0 = -1 \end{cases}$$

Вычтем одно уравнение из другого и нормируем веса

$$\boxed{\frac{(x_+ - x_-)\theta}{||\theta||}} = \frac{2}{||\theta||} \to max$$



$p = \frac{(x_+ - x_-)\theta}{||\theta||}$ - проекция вектора $x_+ - x_-$

Чем больше проекция → тем больше зазор → тем выше уверенность в верной классификации

Так как $\frac{2}{||\theta||}$, чем больше ширина полосы, тем меньше $||\theta||$ - отсюда **регуляризация**

Оптимизация с ограничениями при регуляризации

Рассмотрим задачу оптимизации с ограничениями:

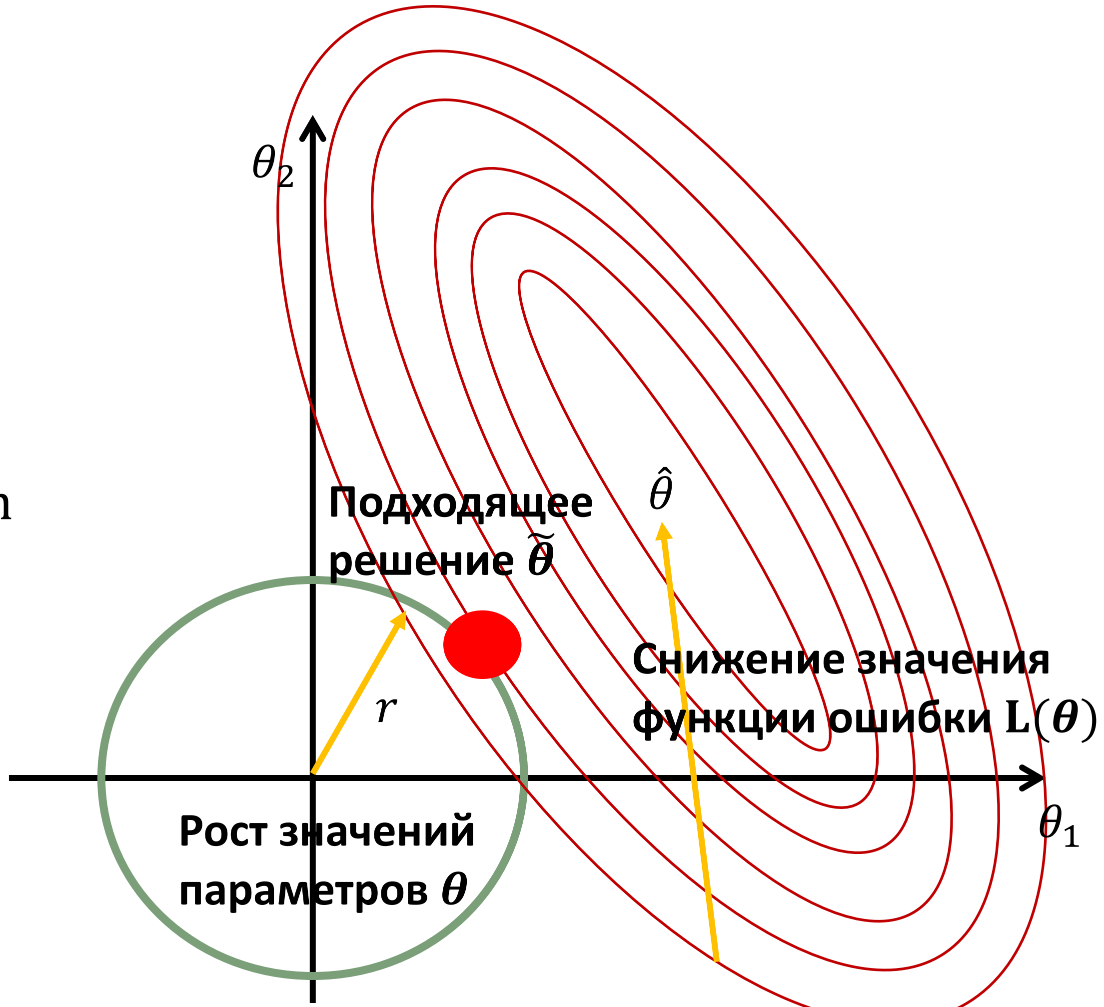
$$L(\theta) = \sum_{i=1}^{n} \left(y_i - \theta_1 x_i^1 - \theta_2 x_i^2\right) \to \min \qquad \text{s.t}\ \theta_1^2 + \theta_2^2 \le r^2$$

Для решения этой оптимизационной задачи с ограничениями воспользуемся методом множителей Лагранжа:

$$L(\theta, \lambda) = \sum_{i=1}^{n} \left(y_i - \theta_1 x_i^1 - \theta_2 x_i^2\right) + \lambda(\theta_1^2 + \theta_2^2 - r^2) \to \min_{\theta,\lambda}$$

Так как λ – это гиперпараметр (просто число) и r^2 - тоже, мы можем исключить их из задачи оптимизации. Получаем следующее выражение:

$$L(\theta) = \sum_{i=1}^{n} \left(y_i - \theta_1 x_i^1 - \theta_2 x_{x_i^2}\right) + \lambda(\theta_1^2 + \theta_2^2) \to \min_{\theta}$$



Решение Hard-margin SVM

Запишем функцию оптимизации с ограничениями:

$$\begin{cases} \frac{1}{2}\left|\left|\theta\right|\right|^2 \ \rightarrow min \\ M_\theta(x_i, y_i) \geq 1 \end{cases}$$

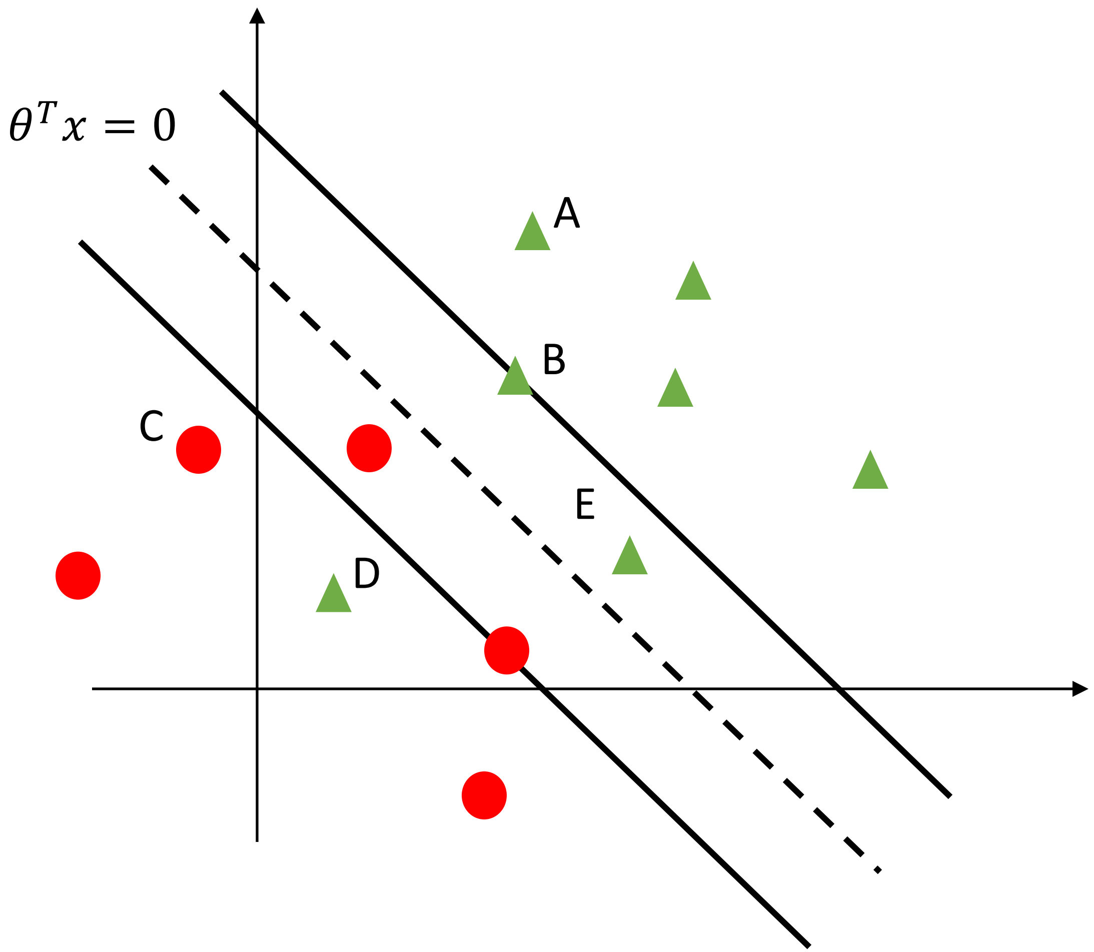
Запишем Лагранжиан:

$$L(x, \lambda) = f_0(x) - \sum_{i=1}^{N} \lambda_i f_i(x)$$

Выпишем оптимизационную задачу:

$$\min L(x, \lambda) = \frac{1}{2}\left|\left|\theta\right|\right|^2 \ - \sum_{i=1}^{N} \lambda_i [y_i(\theta^T x_i \ - \theta_0) - 1] = \frac{1}{2}\left|\left|\theta\right|\right|^2 \ - \sum_{i=1}^{N} \lambda_i [y_i(\theta^T x_i \ - \theta_0)] + \sum_{i=1}^{N} \lambda_i$$

- Это задача безусловной оптимизации - оптимизируем θ, λ – гиперпараметр.
- Это не гладкая функция - ее нельзя оптимизировать обычными численными методами.

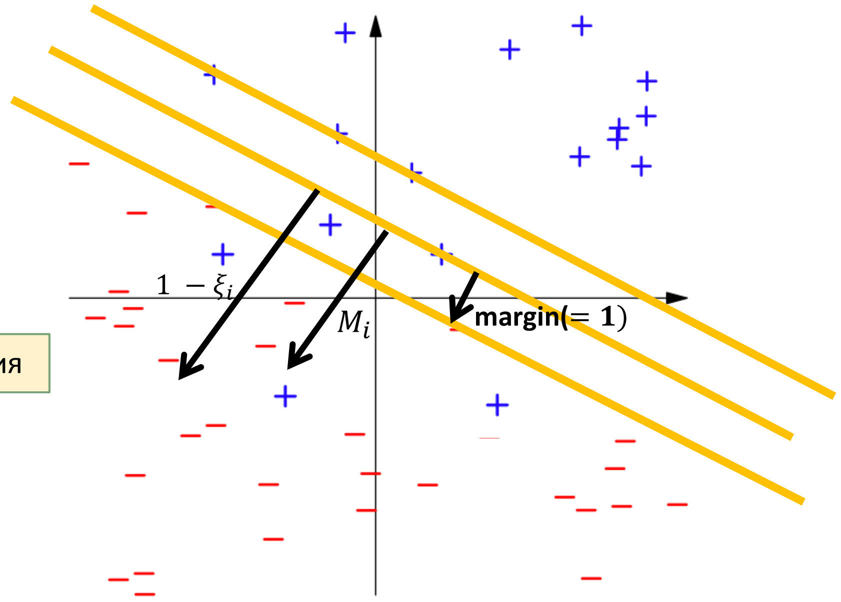

$\theta^T x = 0$
A
B
C
D
E

Soft-margin SVM

Теперь перейдем к линейно-неразделимому случаю. В таком случае, ограничение $M_\theta(x_i, y_i) \geq 1$ – не выполняется, задача не решается. Необходимо ослабить наши требования с помощью условия:

$$M_\theta(x_i, y_i) \geq 1 - \xi$$

Параметр ξ – **насколько мы готовы уменьшить текущий margin($= 1$), чтобы задача с ограничениями решилась.**



$$\begin{cases} \boxed{\frac{1}{2}||\theta||^2 + \mathrm{C}\sum_{i=1}^{N} \xi_i \to min} \\ \boxed{M_\theta(x_i, y_i) \geq 1 - \xi_i} \\ \xi_i \geq 0 \end{cases}$$

1) Уменьшение margin на ξ для выполнения условия

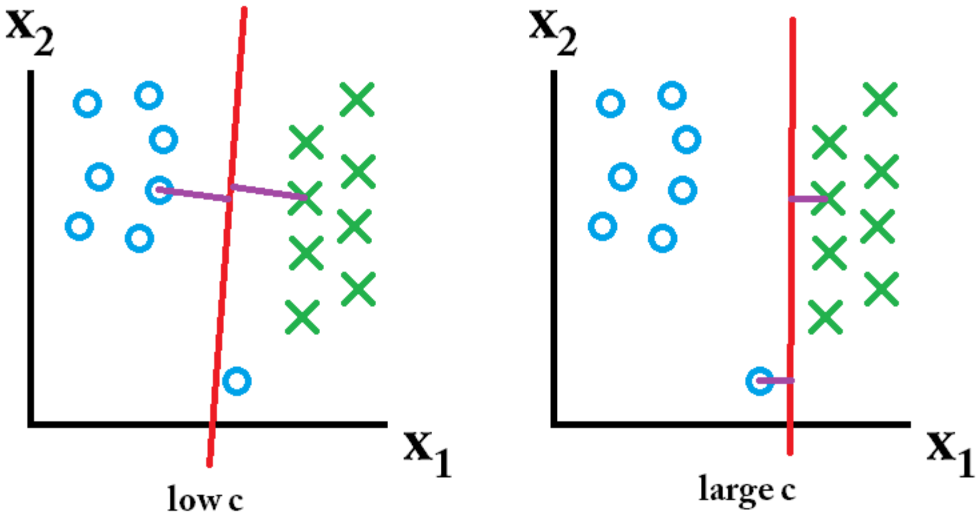
2) Но чтобы уменьшение на ξ не было большим, добавляем в функцию ошибки

- Чем больше C, тем меньше мы делаем разделяющую полосу(меньше уверенности в предсказаниях), и тем больше мы штрафуем за ошибки. Выпишем соответствующую задачу безусловной оптимизации:

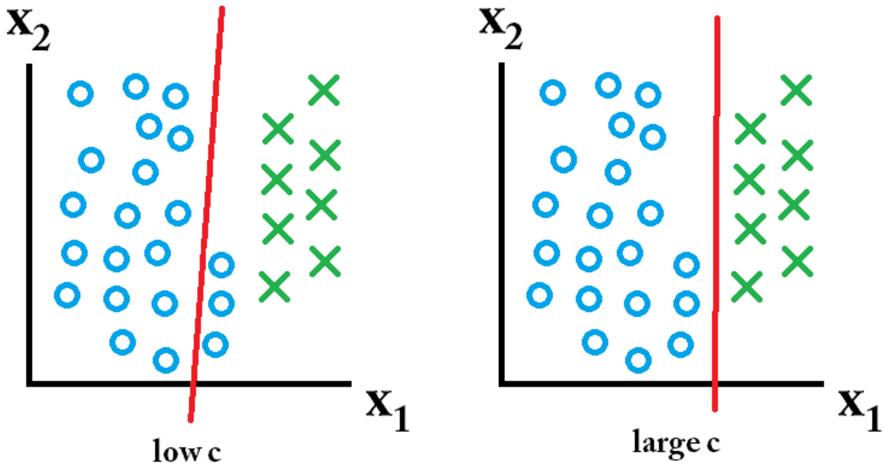
$$\begin{cases} M_\theta(x_i, y_i) \geq 1 - \xi_i \\ \xi_i \geq 0 \end{cases} \qquad \xi_i = \max(0, 1 - y_i(\theta^T x_i - \theta_0))$$

$$\frac{1}{2}||\theta||^2 + \mathrm{C}\sum_{i=1}^{N} \max(0, 1 - y_i(\theta^T x_i - \theta_0)) \to min$$

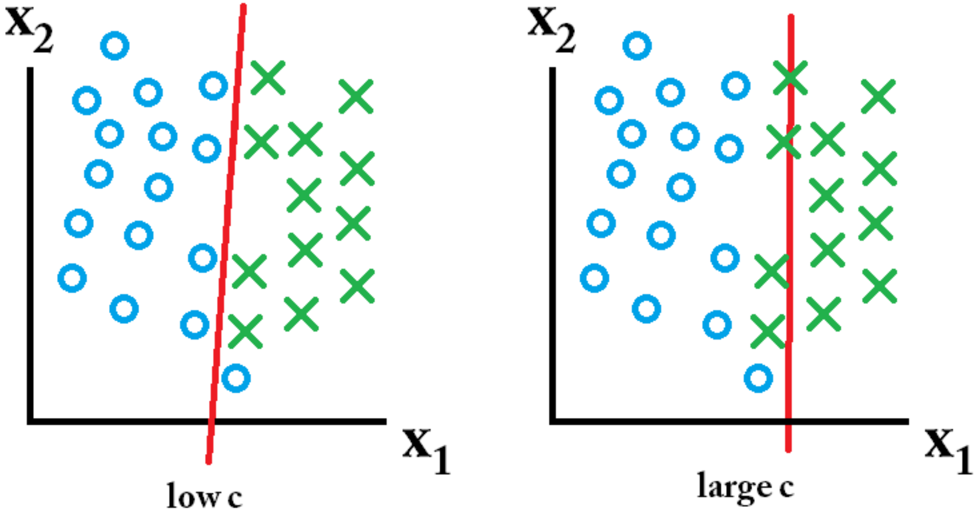
Параметр C



x_2
x_1
low c
x_2
x_1
large c



x_2
x_1
low c
x_2
x_1
large c



x_2
x_1
low c
x_2
x_1
large c

Двойственная задача в SVM

Запишем целевую функцию для минимизации:

$$\frac{1}{2}\left||\theta|\right|^2 + \mathrm{C}\sum_{i=1}^{N}\max(0, 1 - M_\theta(x_i, y_i)) = \frac{1}{2}\left||\theta|\right|^2 + C\sum_{i=1}^{N}\xi_i \to min$$

Выпишем ограничения этой функции и двойственные им переменные:

μ	$-\xi_i \leq 0$
λ	$0 \geq 1 - M_\theta(x_i, y_i) - \xi_i$

Запишем Лагранжиан и раскроем скобки:

$$\frac{1}{2}\left||\theta|\right|^2 + C\sum_{i=1}^{N}\xi_i + \sum_{i=1}^{N}\lambda_i(1 - M_\theta(x_i, y_i) - \xi_i) - \sum_{i=1}^{N}\mu_i\,\xi_i = \frac{1}{2}\left||\theta|\right|^2 - \sum_{i=1}^{N}\lambda_i\big(1 - M_\theta(x_i, y_i)\big) - \sum_{i=1}^{N}\xi_i(\lambda + \mu - C)$$

$$= \frac{1}{2}\left||\theta|\right|^2 - \sum_{i=1}^{N}\lambda_i\big(\cancel{1} - \cancel{1} + y_i(\theta^T x_i \; - \theta_0)\big) - \sum_{i=1}^{N}\xi_i(\lambda + \mu - C)$$

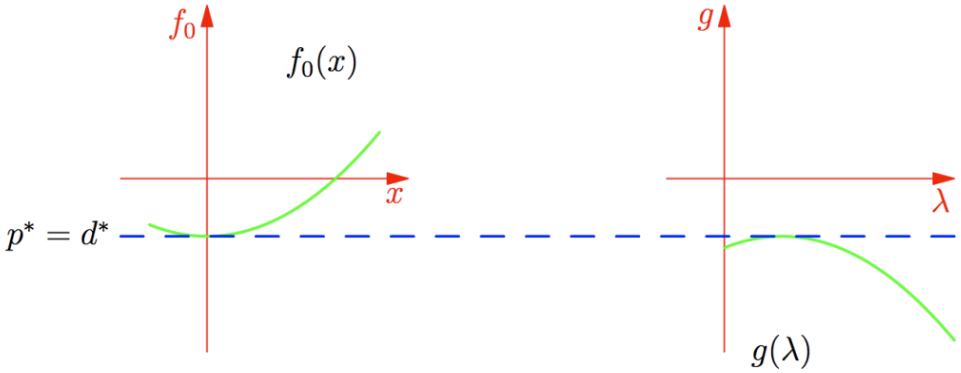
Возьмем производные по θ, θ_0, ξ и приравняем к 0 для поиска оптимальных параметров:

$$\frac{dL}{d\theta} = \theta \; - \sum_{i=1}^{N}\lambda_i y_i x_i = 0 \;\to \boldsymbol{\theta^*} = \sum_{i=1}^{N}\boldsymbol{\lambda_i y_i x_i} \quad \frac{dL}{d\theta_0} = \sum_{i=1}^{N}\boldsymbol{\lambda_i y_i} = \boldsymbol{0} \;\to \sum_{i=1}^{N}\lambda_i y_i = 0 \quad \frac{dL}{d\xi} = -\lambda - \mu + C \to \boldsymbol{\lambda + \mu = C}$$

Двойственная задача в SVM

Вместо минимизации функции ошибки по параметрам θ, мы будем оптимизировать по λ с учетом выведенных ранее тождеств. Нам необходимо, чтобы решение удовлетворяло следующим ограничениям:

$$\theta^* = \sum_{i=1}^{N} \lambda_i y_i x_i \qquad \sum_{i=1}^{N} \lambda_i y_i = 0 \qquad \lambda + \mu = C$$



$$\min f_0(x) = \max g(\lambda)$$

Выпишем исходную задачу оптимизации с ограничениями:

$$\min L_\theta = \frac{1}{2}||\theta||^2 - \sum_{i=1}^{N} \lambda_i [y_i(\theta^T x_i - \theta_0)] + \sum_{i=1}^{N} \lambda_i$$

И преобразуем ее в двойственную задачу([по теореме Каруша-Куна-Таккера]– решение этих двух задач эквивалентно), заменив и подставив $\theta^* = \sum_{i=1}^{N} \lambda_i y_i x_i$ вместо θ и подставив $\sum_{i=1}^{N} \lambda_i y_i = 0$:

$$\max L_\lambda = \frac{1}{2}\sum_{i=1}^{N} \lambda_i y_i x_i \lambda_i y_i x_i - \sum_{i=1}^{N} \lambda_i y_i x_i \lambda_i y_i x_i + \cancel{\sum_{i=1}^{N} \lambda_i y_i \theta_0} + \sum_{i=1}^{N} \lambda_i$$

$$\boxed{= \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N} \lambda_i y_i x_i \lambda_i y_i x_i, \sum_{i=1}^{N} \lambda_i y_i = 0, \lambda_i \geq 0, 0 \leq \lambda \leq C}$$

- При такой постановке задачи мы не зависим от θ.
- Дифференцируя по λ_i и приравнивая к 0 мы получим решение SVM. Точки, где $\lambda_i \neq 0$ – **опорные вектора.**
- $(\boldsymbol{x_i}, \boldsymbol{x_i})$ **- матрица скалярных произведений (Грамма)**

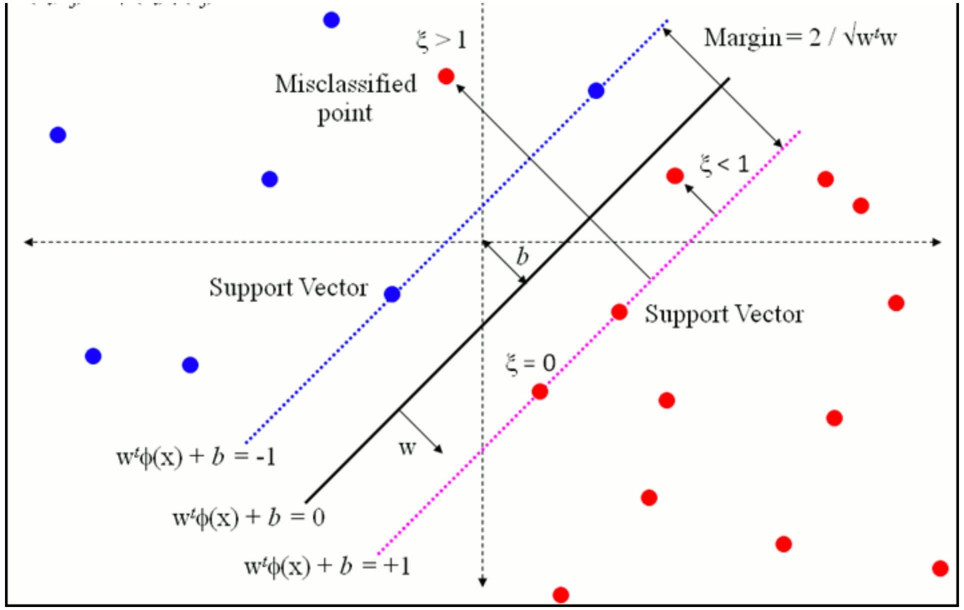
Понятие опорного вектора

1) Если $\lambda_i = 0;\ \xi_i = 0; M_i \geq 1$ - решение не зависит от этого объекта

2) Если $0 < \lambda_i < C;\ \xi_i = 0; M_i = 1$ - опорные объекты

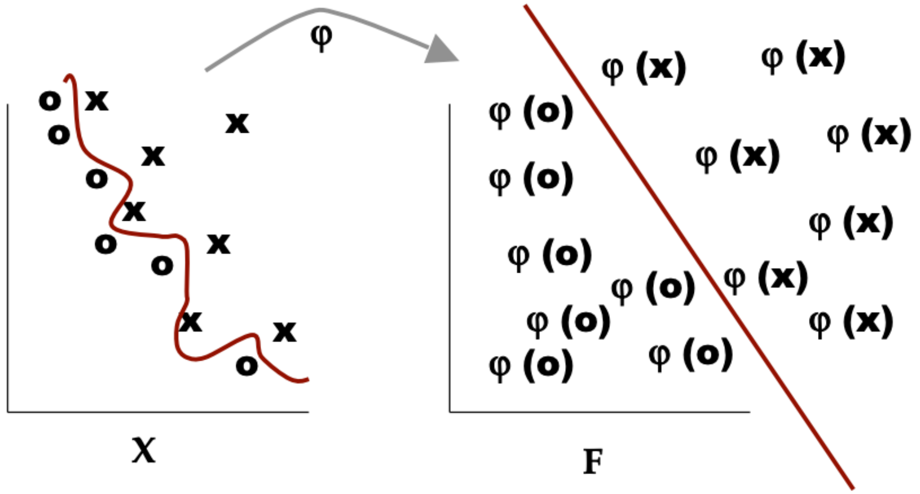
3) Если $\lambda_i = C;\ \xi_i > 0; M_i < 1$ - неверные опорные объекты

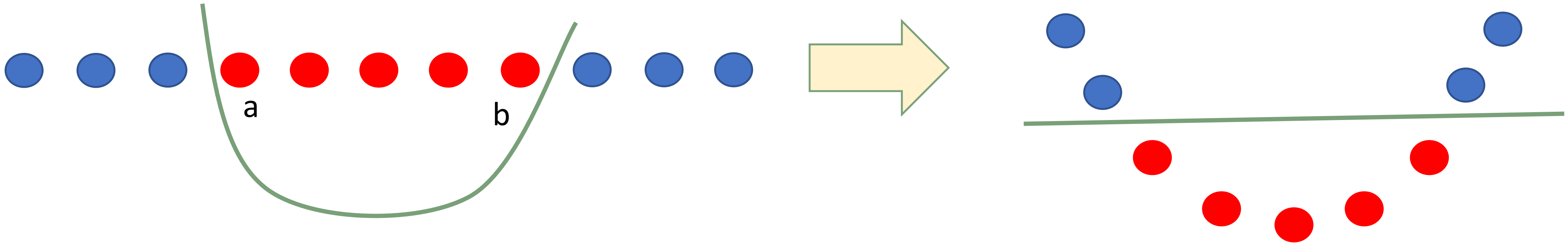
Объект $\boldsymbol{x_i}$ называется опорным, если $\boldsymbol{\lambda_i \neq 0}$



Нелинейный SVM

Ключевая идея – переведем объекты в более высокое пространство и разделим там. Для этого необходимо подобрать некоторую функцию перевода ϕ, которая сделает объекты в более высоком пространстве линейно разделимыми.





Нелинейный SVM

Для нахождения λ_i, мы должны найти попарные скалярные произведения объектов выборки в пространстве размера N.

$$\max L_\lambda = \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N} \lambda_i \lambda_i y_i y_i x_i x_i$$

Если выборка линейной не разделима в пространстве мы можем подобрать такую функцию K(**ядро**), что при некотором преобразовании ϕ, $K(x', x) = < \phi(x'), \phi(x) >$ выборка будет будет разделима в новом пространстве высшей размерности.

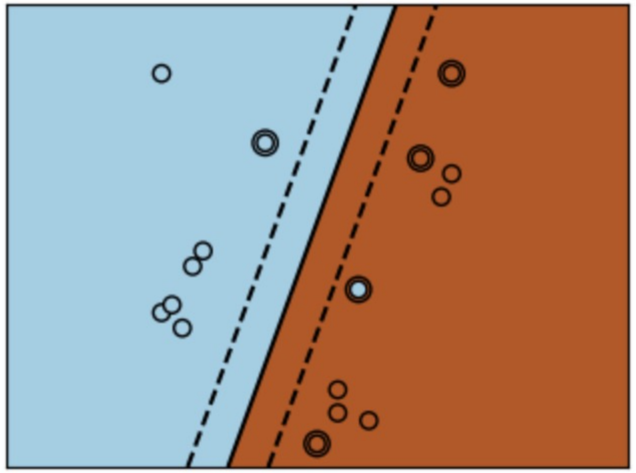
$$\max L_\lambda = \sum_{i=1}^{N} \lambda_i - \frac{1}{2}\sum_{i=1}^{N} \lambda_i \lambda_i y_i y_i K(x_i, x_i)$$

Пускай есть $\mathrm{u}, \mathrm{v} \in R^2$, $K(u, v) =< u, v >^2$, где $u = (u_1, u_2), v = (v_1, v_2)$. Найдем преобразование ϕ при котором $K(x', x) = < \phi(x'), \phi(x) >$.

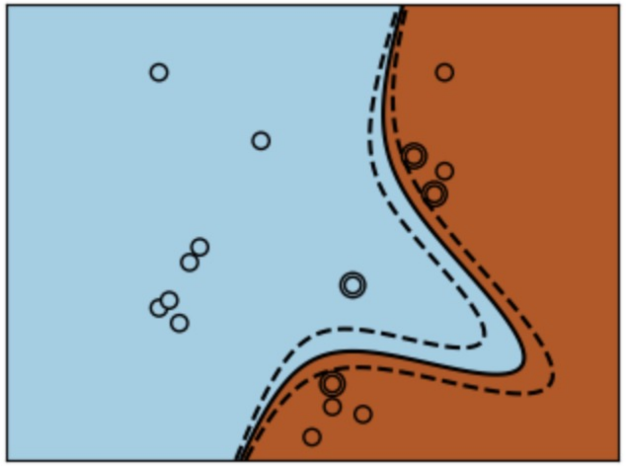
$$K(u, v) = < u, v >^2 = < (u_1, u_2)(v_1, v_2) >^2 = (u_1 v_1 + u_2 v_2)^2 = u_1^2 v_1^2 + u_2^2 v_2^2 + 2u_1 v_1 u_2 v_2$$
$$= < \left(u_1^2, u_2^2, \sqrt{2} u_1 v_2\right), \left(u_1^2, u_2^2, \sqrt{2} u_1 v_2\right) >$$

Таком образом, мы нашли преобразование ϕ, которое переводит векторы u, v в R^3 и возвращает их скалярное произведение в новом пространстве.

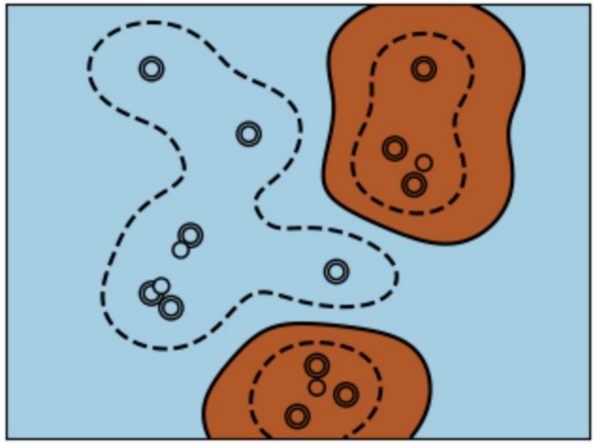
Виды ядер в SVM



Линейное ядро (x, x')



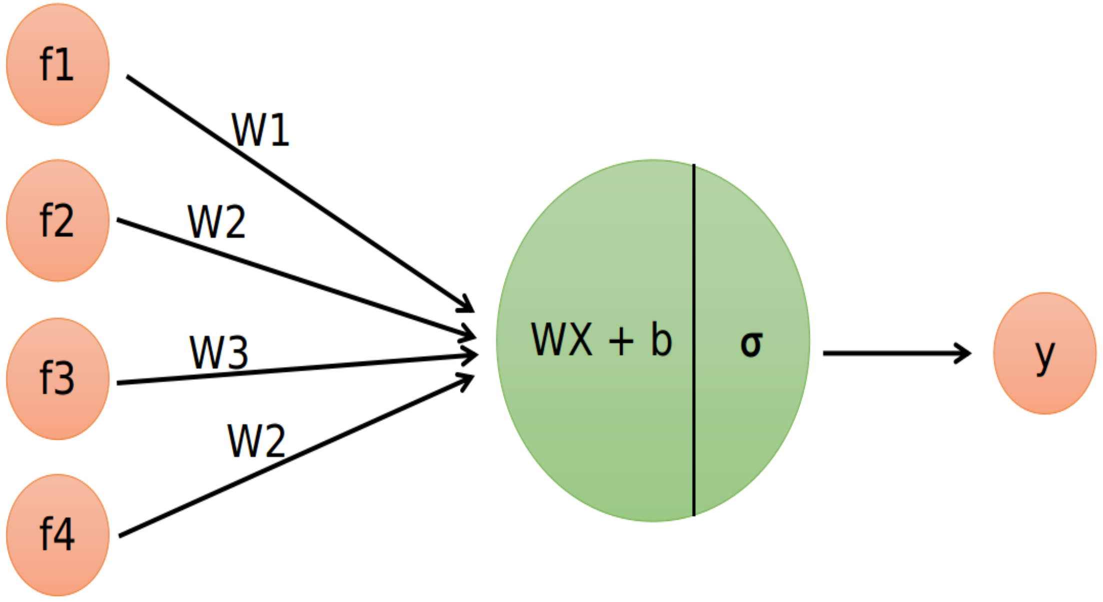
Полиномиальное ядро $(x, x)^2$



RBF ядро $(\gamma||x - x'||^2)$

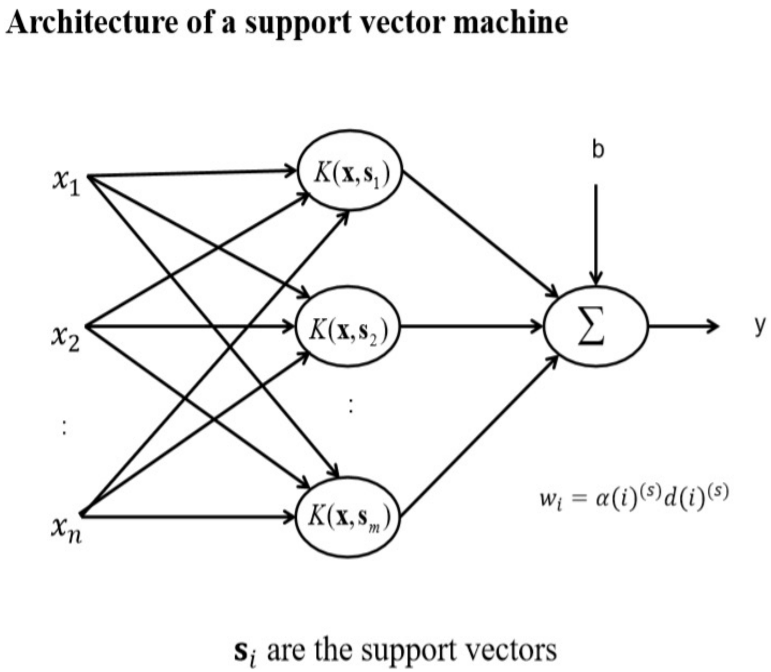
https://scikit-learn.org/stable/auto_examples/svm/plot_svm_kernels.html#sphx-glr-auto-examples-svm-plot-svm-kernels-py

Logistic Regression



- Максимизация правдоподобия
- Возвращает вероятность отнесения к положительному классу
- Гладкая функция для оптимизации
- Лучше интерпретируется

SVM



- Поиск оптимальной разделяющей гиперплоскости
- Имеет большую обобщающую способность
- Если $d \gg N$, быстрее обучается.
- Требует меньше данных(т.к часть этих данных окажется не нужна)
- Имеет точно оптимальное решение

SVM in nutshell

- Те, объекты, чьи λ не равны 0 при оптимизации – есть используемые опорные векторы
- Общая задача оптимизации при решении SVM

$$\frac{1}{2}\left||\theta|\right|^2 + \mathrm{C}\sum_{i=1}^{N}\max(0, 1 - y_i(\theta^T x_i \ - \theta_0)) \to min$$

- Двойственная задача Лагранжа

$$\sum_{i=1}^{N}\lambda_i - \frac{1}{2}\sum_{i=1}^{N}\lambda_i y_i x_i \lambda_i y_i x_i \to max, \sum_{i=1}^{N}\lambda_i y_i = 0, \lambda_i \geq 0, 0 \leq \lambda \leq C$$

1) Нужно находить d параметров в общей постановке задачи и N – в двойственной
2) Если $N \ << \ d$ – выгодней рассчитывать λ, чем θ

- Предсказание на объекте x рассчитывается – знак суммы положительных опорных объектов с отрицательными

$$a(x) = sign\left(\sum_{i=1}^{l}\lambda_i y_i (x \cdot x_i) + b\right), \theta^* = \sum_{i=1}^{N}\lambda_i y_i x_i \, , \theta_0 = \frac{1}{N}\sum_{i=1}^{N}(y_n \ - \sum_{j=1}^{N}\lambda_j y_j (x_i \cdot x_j))$$

- Метод опорных векторов, как и логистическая регрессия строит верхнюю оценку на функцию доли ошибок и добавляет к ней квадратичную регуляризацию.