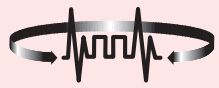
Машины на опорных векторах (SVM-классификация)

В.В.Вьюгин

ИППИ РАН



Support Vector Machine

Задачи решаемые SVM:

- Задачи классификации
- Задачи регрессии

Последовательность решения

- Настройка на обучающей выборке
- Применение (проверка) на тестовой (рабочей) выборке

Обучающая выборка

$$S = ((\bar{x}_1, y_1), \ \ldots, \ (\bar{x}_l, y_l)),$$

$\bar{x}_i \in \mathscr{R}^n$ – вектор евклидового пространства большой размерности n.

$y_i \in D$ – конечное множество. Часто $D = \{-1, 1\}$.

$y_i \in \mathscr{R}$ – в задаче многомерной регрессии.

Основное предположение

$(\bar{x}_i, y_i)$ – i.i.d согласно P (в частности, может быть $y_i = f(\bar{x}_i)$).

Ошибки обобщения будут оцениваться через P.

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

Задача классификации

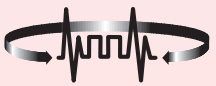
Задача классификации

$h : \mathscr{R}^n \to \{-1, 1\}$ – функция классификации.

$\mathrm{err}_P(h) = P\{(\bar{x}, y) : h(\bar{x}) \neq y\}$ – ошибка классификации..

$\mathrm{err}_S(h) = \frac{1}{l}|\{i : h(\bar{x}_i) \neq y_i, 1 \leq i \leq l\}|$ – доля ошибок классификации h на выборке S.

$\mathrm{err}_S(h) = 0$ или $h(\bar{x}_i) = y_i$ для всех $1 \leq i \leq l$ – гипотеза классификации h согласована с выборкой S.



Теория обобщения

Для произвольной гипотезы классификации h и $\varepsilon > 0$ имеем

$$P^l\{S : \mathrm{err}_S(h) = 0 \& \mathrm{err}_P(h) > \varepsilon\} = (1 - \mathrm{err}_P(h))^l \leq e^{-l\varepsilon}.$$

Пусть H – некоторый класс гипотез классификации. Если класс H конечный, то

$$P^l\{S : (\exists h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_P(h) > \varepsilon)\} \leq |H| e^{-l\varepsilon}.$$

PAC – переформулировки

С вероятностью $\leq \delta$ по случайной обучающей выборке S

$$\mathrm{err}_S(h) = 0 \;\&\; \mathrm{err}_P(h) > \frac{1}{l} \ln \frac{|H|}{\delta}.$$

С вероятностью $\geq 1 - |H|e^{-l\varepsilon}$ по S

$$\mathrm{err}_P(h) > \varepsilon \;\Rightarrow\; \mathrm{err}_S(h) > 0.$$

Бесконечный класс функций

Функция роста (сложность) класса H

$$B_H(l) = \max_{(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_l)} |\{(h(\bar{x}_1), h(\bar{x}_2), \ldots, h(\bar{x}_l)) : h \in H\}|.$$

Теорема 1:

Имеет место оценка

$$P^l\{S : (\exists\ h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_P(h) > \varepsilon)\} \leq 2B_H(2l)e^{-\frac{1}{2}\varepsilon l}.$$

VC – размерность класса функций

Теорема:
Для любого класса индикаторных функций H реализуется одна из двух возможностей:
1) $B_H(l) = 2^l$ для всех l;
2) Существует полностью разделимая выборка максимального размера d; в этом случае $B_H(l) = 2^l$ при $l \leq d$;

$$B_H(l) \leq \sum_{i=0}^{d} \binom{l}{i} < \left(\frac{el}{d}\right)^d$$

при $l > d$.
d – размерность Вапника–Червоненкиса класса H.

Теорема:

1) VC-размерность класса всех линейных функций классификации $h(\bar{x}) = \text{sign}((\bar{w} \cdot \bar{x}) + b)$ равна $n+1$.

2) VC-размерность класса всех однородных классификаторов над $h(\bar{x}) = \text{sign}(\bar{w} \cdot \bar{x})$ равна n.

3) Для класса всех однородных функций классификации над $\mathscr{R}^n$ при $l > n$ выполнено

$$G_{\mathscr{L}}(l) = \ln H_{\mathscr{L}}(l) = \ln \left(2 \sum_{i=0}^{n} \binom{l-1}{i} \right) < \\ < n \left(1 + \ln \left(\frac{l}{n} \right) \right) + \ln 2.$$

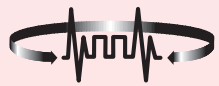
Доказательство теоремы 1:

Лемма:

Задан класс H функций классификации. Рассматриваются две случайные выборки S, S' длины l. Тогда для любого $\varepsilon > 0$ при $l \geq \frac{2}{\varepsilon^2}$ имеет место неравенство

$$P^l\{S : (\exists\ h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_P(h) > \varepsilon)\} \leq$$

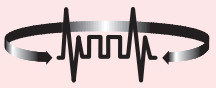
$$\leq\ 2P^{2l}\{SS' : (\exists h \in H)(\mathrm{err}_S(h) = 0 \& \mathrm{err}_{S'}(h) > \frac{1}{2}\varepsilon)\}.$$



Условная вероятность того, что функция $h \in H$ делает более $\frac{1}{2}\varepsilon l$ ошибок на выборке SS' с фиксированным составом Υ и все они сосредоточены на S'

$$P^{2l}\{SS' : (\exists\, h \in H)(\mathrm{err}_S(h) = 0 \,\&\, \mathrm{err}_{S'}(h) > \frac{1}{2}\varepsilon | \eta(SS') = \Upsilon\} \leq$$
$$\leq B_H(2l)e^{-\frac{1}{2}\varepsilon l}.$$

Левая часть – случайная величина (от Υ).
Можно проинтегрировать по Υ.

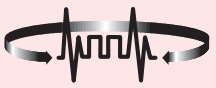


Следствие PAC - формулировка:
Пусть VC-размерность класса H равна d.
С вероятностью $\leq \delta$ по случайной обучающей выборке S

$$\mathrm{err}_S(h) = 0 \ \& \ \mathrm{err}_P(h) > \frac{2}{l}\left(d \ln \frac{2el}{d} + \ln \frac{2}{\delta}\right).$$

при $l > d$.
С вероятностью $\geq 1 - 2de^{-\frac{1}{2}\varepsilon l}$ по S

$$\mathrm{err}_P(h) > \varepsilon \ \Rightarrow \ \mathrm{err}_S(h) > 0.$$



Разделимая выборка

Выборка $S = ((\bar{x}_1, y_1), (\bar{x}_2, y_2), \ldots, (\bar{x}_l, y_l))$ – разделима с помощью гиперплоскости, если существует вектор $\bar{w}$ и числа c и $\gamma > 0$ такие, что

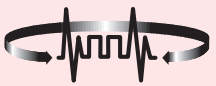
$$(\bar{w} \cdot \bar{x}_i) + c \geq \gamma \text{ при } y_i = 1,$$
$$(\bar{w} \cdot \bar{x}_i) + c \leq -\gamma \text{ при } y_i = -1.$$

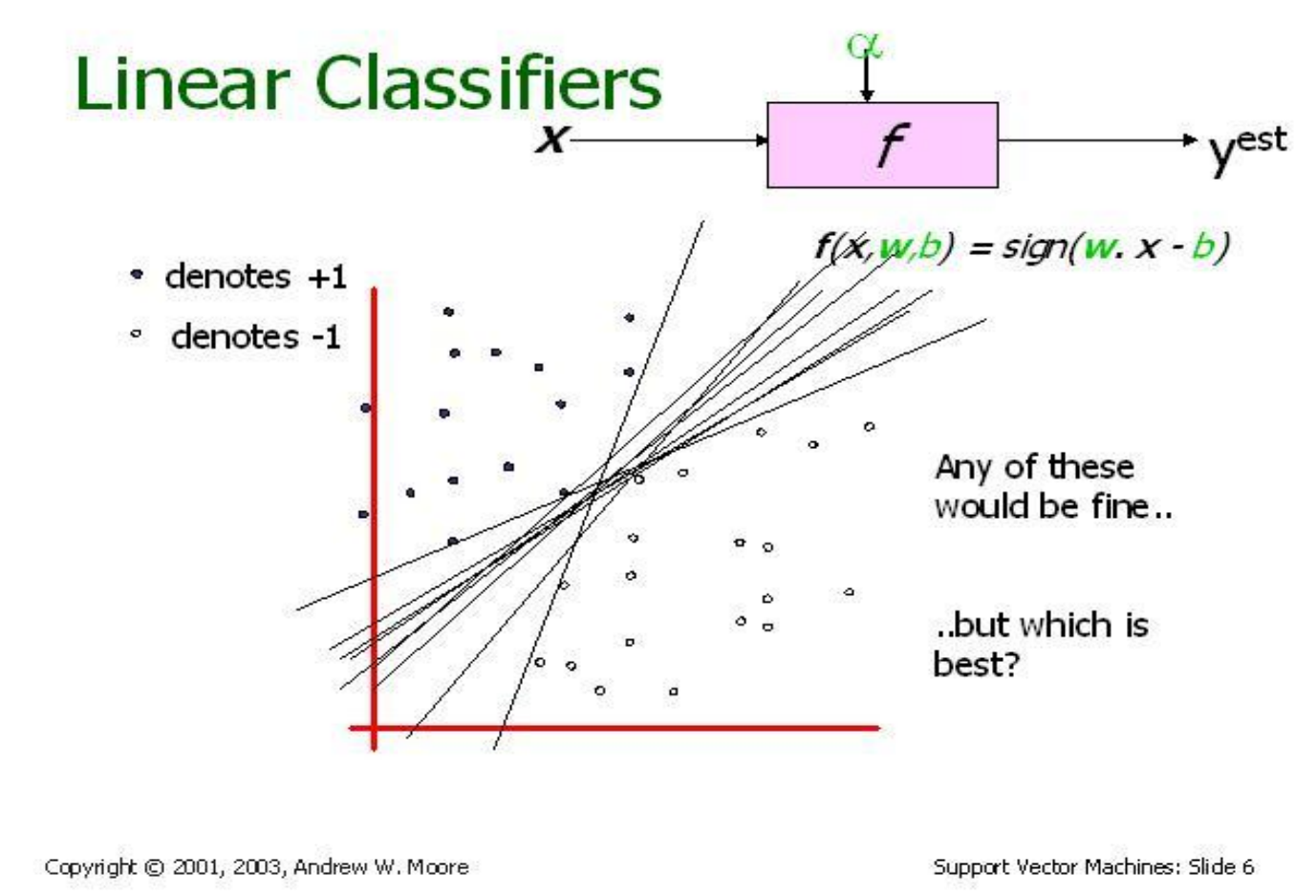
Выборка $S = ((\bar{x}_1, y_1), (\bar{x}_2, y_2), \ldots, (\bar{x}_l, y_l))$ – разделима с помощью гиперплоскости, если существует вектор $\bar{w}$ и числа b такие, что

$$(\bar{w} \cdot \bar{x}_i) + b \geq 1 \text{ при } y_i = 1,$$
$$(\bar{w} \cdot \bar{x}_i) + b \leq -1 \text{ при } y_i = -1.$$

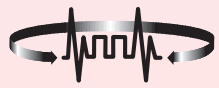
Эквивалентно

$$y_i((\bar{w} \cdot \bar{x}_i) + b) \geq 1.$$





Множество разделяющих гиперплоскостей



Расстояние между гиперплоскостями

$$(\bar{w} \cdot \bar{x}) + b = 1,$$
$$(\bar{w} \cdot \bar{x}) + b = -1$$

равно

$$\rho = \frac{2}{|\bar{w}|}.$$

Разделяющая гиперплоскость

$$(\bar{w} \cdot \bar{x}) + b = 0$$

находится между ними на расстоянии $1/|\bar{w}|$.

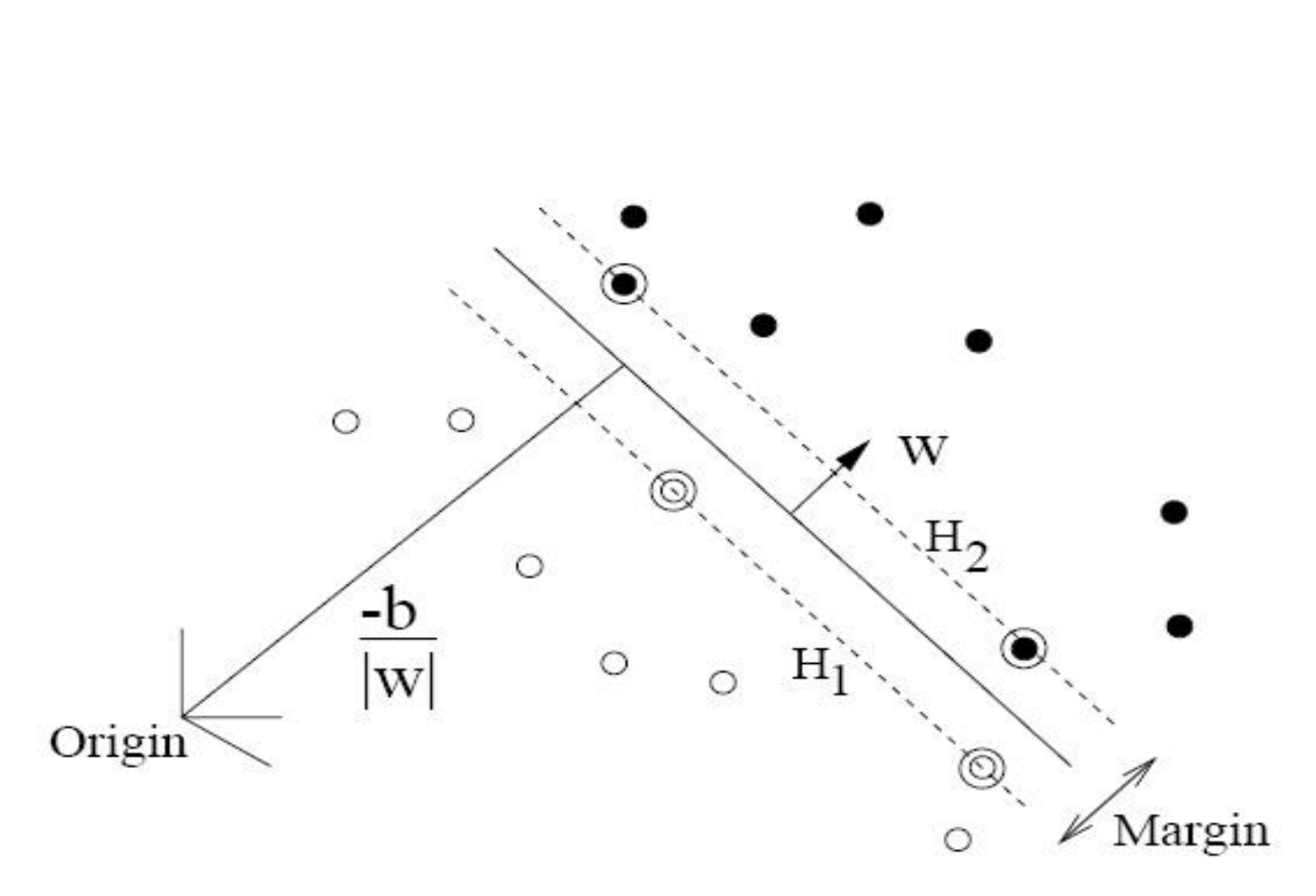
Максимизируем расстояние между граничными гиперплоскостями: $2/|\bar{w}|^2 \to \max$

Получаем задачу оптимизации

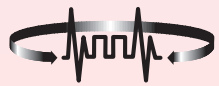
$$(\bar{w} \cdot \bar{w}) = \sum_{i=1}^{l} w_i^2 \to \min$$

$$\text{при условиях } y_i((\bar{w} \cdot \bar{x}_i) + b) \geq 1$$

при $i = 1, \ldots l$.



Оптимальная гиперплоскость



Лагранжиан

$$L(\bar{w}, b, \bar{\alpha}) = \frac{1}{2}(\bar{w} \cdot \bar{w}) - \sum_{i=1}^{l} \alpha_i (y_i((\bar{w} \cdot \bar{x}_i) + b) - 1),$$

где $\alpha_i \geq 0$ - множители Лагранжа.

Необходимое условие минимума лагранжиана имеет вид

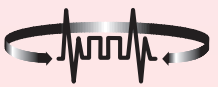
$$\frac{\partial L(\bar{w}, b, \bar{\alpha})}{\partial \bar{w}} = \bar{w} - \sum_{i=1}^{l} \alpha_i y_i \bar{x}_i = \bar{0},$$

$$\frac{\partial L(\bar{w}, b, \bar{\alpha})}{\partial b} = \sum_{i=1}^{l} \alpha_i y_i = 0.$$

Вектор весов разделяющей гиперплоскости

$$\bar{w} = \sum_{i=1}^{l} \alpha_i y_i \bar{x}_i,$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0.$$



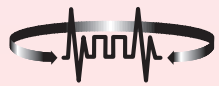
Прямая задача оптимизации: общий случай

$$f(\bar{w}) \to \min$$
$$\text{при условиях } \bar{g}(\bar{w}) \leq \bar{0},$$
$$\bar{h}(\bar{w}) = \bar{0}$$

Лагранжиан

$$L(\bar{w}, \bar{\alpha}, \bar{\beta}) = f(\bar{w}) + \sum_{i=1}^{m} \alpha_i g_i(\bar{w}) + \sum_{i=1}^{m} \beta_i h_i(\bar{w}) =$$
$$= f(\bar{w}) + \bar{\alpha}\bar{g}(\bar{w}) + \bar{\beta}\bar{h}(\bar{w}).$$

Теорема Куна–Таккера:
Пусть область допустимости, функция f – выпуклые, функции $\bar{h}$, $\bar{g}$ – аффинные.
Тогда вектор $\bar{w}^*$ является решением прямой задачи тогда и только тогда, когда существует пара $(\bar{\alpha}^*, \bar{\beta}^*)$, для которой



$$\frac{\partial L(\bar{w}^*, \bar{\alpha}^*, \bar{\beta}^*)}{\partial \bar{w}} = \bar{0},$$

$$\frac{\partial L(\bar{w}^*, \bar{\alpha}^*, \bar{\beta}^*)}{\partial \bar{\beta}} = \bar{0}, \tag{1}$$

$$\alpha_i^* g_i(\bar{w}^*) = 0,\ i = 1,\ \ldots,\ m, \tag{2}$$

$$g_i(\bar{w}^*) \leq 0,\ i = 1,\ \ldots,\ m,$$

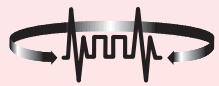
$$\alpha_i^* \geq 0,\ i = 1,\ \ldots,\ m.$$

Задача выпуклой оптимизации для SVM

Найти максимум квадратичного функционала

$$W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\bar{x}_i \cdot \bar{x}_j).$$

при $\alpha_i \geq 0$, $i = 1, \ldots l$.



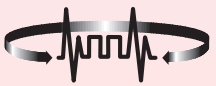
Решение задачи поиска оптимальной гиперплоскости:

$$\bar{w}_0 = \sum_{i=1}^{l} \alpha_i^0 y_i \bar{x}_i.$$

Оптимальные решения $\bar{w}_0$ и b_0 должны удовлетворять условиям Каруша - Куна - Таккера

$$\alpha_i^0 (y_i((\bar{w}_0 \cdot \bar{x}_i) + b_0) - 1) = 0$$

при $i = 1, \ldots l$.

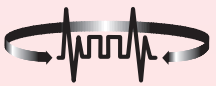


$\alpha_i^0 > 0$ может быть только для тех векторов, которые лежат на гиперплоскостях

$$(\bar{w}_0 \cdot \bar{x}_i) + b_0 = \pm 1.$$

Такие векторы называются ОПОРНЫМИ векторами (support vectors).

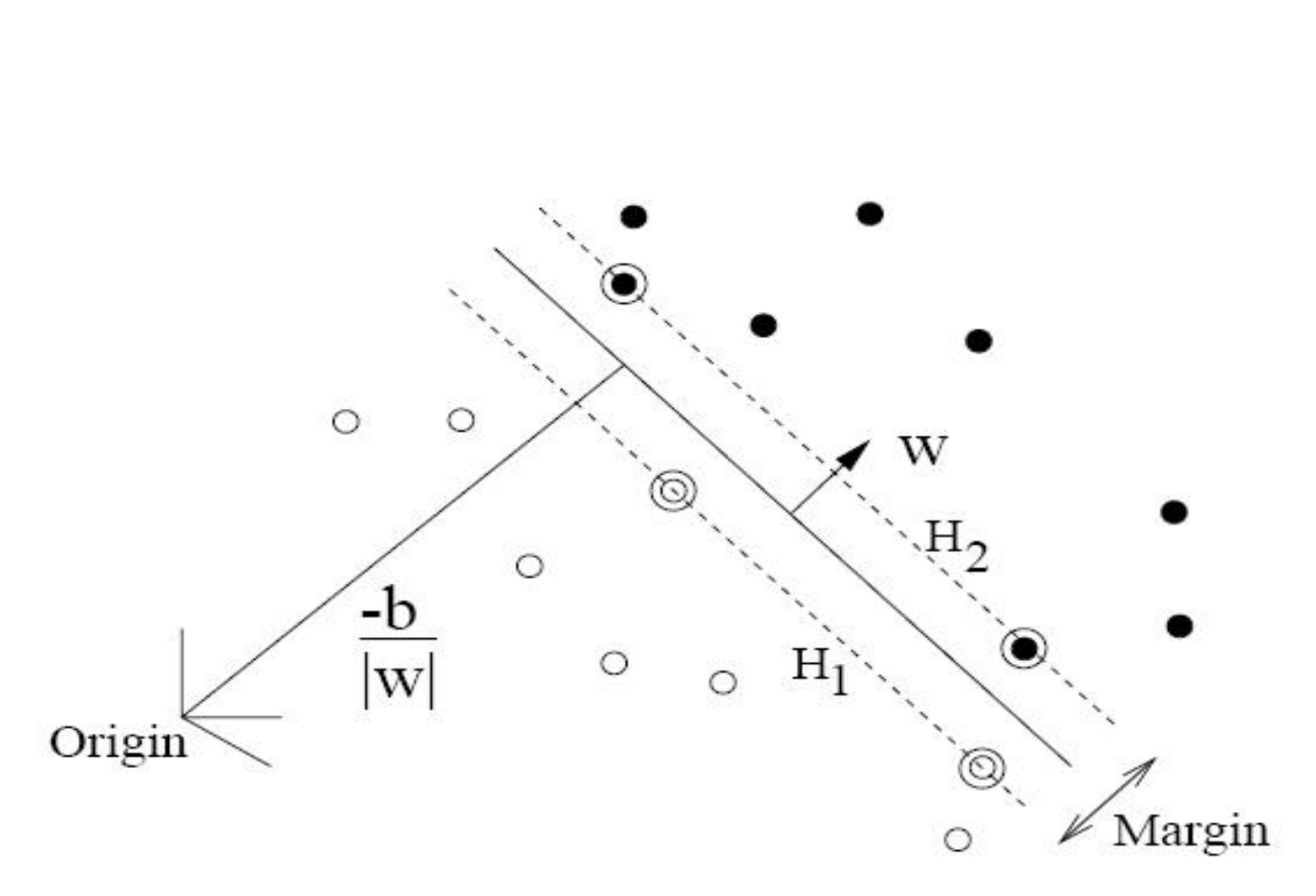
$$\bar{w}_0 = \sum_{s=1}^{k} \alpha_{i_s}^0 y_{i_s} \bar{x}_{i_s}.$$



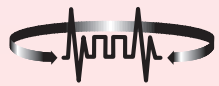
Оптимальная гиперплоскость имеет вид

$$\sum_{s=1}^{k} \alpha_{i_s}^0 y_{i_s} (\bar{x}_{i_s} \cdot \bar{x}) + b_0 = 0.$$

Остальные - не опорные векторы, при поиске оптимальной гиперплоскости можно не принимать во внимание, например, их можно изменить, при этом оптимальная гиперплоскость не изменится.



Разделяющие гиперплоскости и опорные векторы



Оценка вероятности ошибки обобщения через число опорных векторов

$$\mathrm{err}_P(h_{\hat{S}}) \leq \frac{d \ln l}{l},$$

где d - число опорных векторов.

SVM - метод в пространстве признаков

Нелинейное отображение

$$\bar{x} = (x_1, \dots x_n) \to \bar{\phi}(\bar{x}) = (\phi_1(\bar{x}), \dots \phi_N(\bar{x})). \tag{3}$$

Образ выборки

$$\bar{\phi}(\bar{x}_1), \dots, \bar{\phi}(\bar{x}_l)$$

в пространстве признаков $\mathscr{R}^N$.

Построим гиперплоскость в пространстве признаков $\mathscr{R}^N$

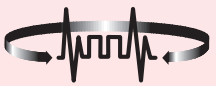
$$\sum_{j=1}^{N} w_j z_j + b = 0, \tag{4}$$

разделяющую вектора $\bar{\phi}(\bar{x}_1), \ldots, \bar{\phi}(\bar{x}_l)$.

Эта гиперплоскость имеет своим прообразом в пространстве $\mathscr{R}^n$, в общем случае нелинейную, поверхность

$$\sum_{j=1}^{N} w_j \phi_j(\bar{x}) + b = 0,$$

разделяющую вектора $\bar{x}_1, \ldots, \bar{x}_l$



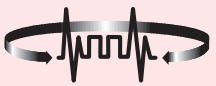
Вектор весов разделяющей гиперплоскости в пространстве признаков

$$\bar{w} = \sum_{i=1}^{l} \alpha_i^0 y_i \bar{\phi}(\bar{x}_i)$$

В координатах это представление имеет вид

$$w_j = \sum_{i=1}^{l} \alpha_i^0 y_i \phi_j(\bar{x}_i)$$

при $j = 1, \ldots, N$.



$$\sum_{j=1}^{N} w_j \phi_j(\bar{x}) + b =$$

$$\sum_{j=1}^{N} \left(\sum_{i=1}^{l} \alpha_i^0 y_i \phi_j(\bar{x}_i) \right) \phi_j(\bar{x}) + b =$$

$$\sum_{i=1}^{l} \alpha_i^0 y_i \sum_{j=1}^{N} \phi_j(\bar{x}) \phi_j(\bar{x}_i) + b =$$

$$\sum_{i=1}^{l} \alpha_i^0 y_i (\bar{\phi}(\bar{x}) \cdot \bar{\phi}(\bar{x}_i)) + b =$$

$$\sum_{i=1}^{l} \alpha_i^0 y_i K(\bar{x}, \bar{x}_i) + b = 0,$$

где

$$K(\bar{x}_i, \bar{x}) = (\bar{\phi}(\bar{x}_i) \cdot \bar{\phi}(\bar{x})).$$

Пример. Используем для классификации полиномы 2-ой степени от n переменных. Новые переменные

$$z_0 = 1, z_1 = x_1, \ \ldots, \ z_n = x_n,$$
$$z_{n+1} = x_1^2, \ \ldots, \ z_{2n} = x_n^2,$$
$$z_{2n+1} = x_1 x_2, \ \ldots, \ z_N = x_n x_{n-1}.$$

Нелинейное отображение

$$\bar{x} = (x_1, \ \ldots, \ x_n) \to \bar{\phi}(\bar{x}) = (1, z_1, \ \ldots, \ z_{N-1})$$

пространства $\mathscr{R}^n$ в пространство $\mathscr{R}^N$.

Пример. Прообразом разделяющей гиперплоскости в пространстве признаков $Z = \mathscr{R}^N$:

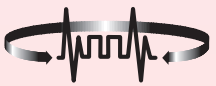
$$(\bar{w} \cdot \bar{z}) = 0$$

при отображении $\bar{x} \to \bar{z}$ является поверхность второго порядка в исходном пространстве $\mathscr{R}^n$:

$$(\bar{w} \cdot \bar{\phi}(\bar{x})) = \sum_{i=1}^{N-1} w_i z_i + w_0 =$$

$$= w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=n+1}^{2n} w_i x_i^2 + \sum_{i=2n+1}^{N-1} w_i x_{j_i} x_{k_i} = 0,$$

где (j_i, k_i) – пара натуральных чисел с номером i в какой-нибудь взаимно однозначной нумерации всех пар натуральных чисел $\leq n$.



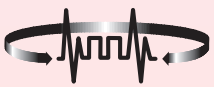
Пример. Отображение из $\mathscr{R}^n$ в $\mathscr{R}^N$ в удобном виде:

$$\bar{\phi}(\bar{x}) = \bar{\phi}(x_1, \ldots, x_n) =$$
$$(1, \sqrt{2}x_1, \ldots, \sqrt{2}x_n, x_1^2, \ldots x_n^2, \sqrt{2}x_1x_2, \ldots \sqrt{2}x_{n-1}x_n).$$

Тогда

$$K(\bar{x}, \bar{y}) = (\bar{\phi}(\bar{x}) \cdot \bar{\phi}(\bar{y}) =$$
$$1 + \sum_{i=1}^{n} 2x_iy_i + \sum_{i=1}^{n} x_i^2y_i^2 + \sum_{i \neq j} 2x_ix_jy_iy_j = (1 + \bar{x} \cdot \bar{y})^2.$$

Получаем $K(\bar{x}, \bar{y}) = (1 + \bar{x} \cdot \bar{y})^2$ - полиномиальное ядро второго порядка.



Пример. Функция классификации соответствующая оптимальной разделяющей гиперплоскости в пространстве признаков, имеет вид

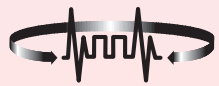
$$f(\bar{x}) = \sum_{i=1}^{l} \alpha_i^0 y_i (1 + \bar{x}_i \cdot \bar{x})^2 + b$$

Примеры ядер

$(\bar{x}, \bar{y}) = (\bar{x} \cdot \bar{y})^d$,

$K(\bar{x}, \bar{y}) = ((\bar{x} \cdot \bar{y}) + c)^d$ - полиномиальные ядра.

$K(\bar{x}, \bar{y}) = \exp(-\sum_{i=1}^{n} (x_i - y_i)^2 / \sigma^2)$ - гауссовское ядро.



Экспоненциальное ядро $K(\bar{u},\bar{v}) = \exp((\bar{u}\cdot\bar{v})/\sigma^2)$.

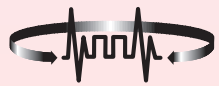
Ряд Тейлора: $\exp((\bar{u}\cdot\bar{v})) = \sum_{k=0}^{\infty} \frac{(\bar{u}\cdot\bar{v})^k}{k!}$

Экспоненциальное ядро трансформируется в Гауссово ядро

$$\frac{K(\bar{u},\bar{v})}{\sqrt{K(\bar{u},\bar{u})K(\bar{v},\bar{v})}} =$$

$$\frac{\exp((\bar{u}\cdot\bar{v})/\sigma^2)}{\sqrt{\exp((\bar{u}\cdot\bar{u})/\sigma^2)\exp((\bar{v}\cdot\bar{v})/\sigma^2)}} =$$

$$\exp(-|\bar{u}-\bar{v}|^2/2\sigma^2).$$

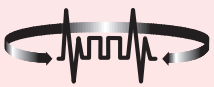


Неотделимая выборка

Ищем вектора $\bar{w}$, $\bar{\xi}$ и число b, так чтобы

$$(\bar{w} \cdot \bar{w}) + C \sum_{i=1}^{l} \xi_i^2 \to \min$$
$$y_i((\bar{w} \cdot \bar{x}_i) + b_0) \geq 1 - \xi_i,$$
$$\xi_i \geq 0$$

при $i = 1, \ldots l$. Константа C определяет баланс между двумя частями функционала.

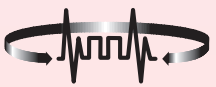


Лагранжиан задачи имеет вид

$$L(\bar{w}, b, \bar{\xi}, \bar{\alpha}) = \frac{1}{2}(\bar{w} \cdot \bar{w}) + \frac{C}{2} \sum_{i=1}^{l} \xi_i^2 -$$

$$\sum_{i=1}^{l} \alpha_i (y_i((\bar{w} \cdot \bar{x}_i) + b) - 1 + \xi_i),$$

где $\alpha_i \geq 0$ - множители Лагранжа.



Соответствующая двойственная задача формулируется путем дифференцирования лагранжиана

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha})}{\partial \bar{w}} = \bar{w} - \sum_{i=1}^{l} y_i \alpha_i \bar{x}_i = \bar{0},$$

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha})}{\partial \bar{\xi}} = C\bar{\xi} - \bar{\alpha} = \bar{0},$$

$$\frac{\partial L(\bar{w}, b, \bar{\xi}, \bar{\alpha})}{\partial b} = \sum_{i=1}^{l} y_i \alpha_i,$$

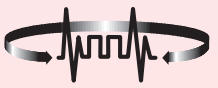
а также подстановкой этих соотношений в лагранжиан.

$$L(\bar{w}, b, \bar{\xi}, \bar{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j (\bar{x}_i \cdot \bar{x}_j) - \frac{1}{2C} (\bar{\alpha} \cdot \bar{\alpha}).$$

Условия Каруша - Куна - Таккера имеют вид

$$\alpha_i (y_i ((\bar{w} \cdot \bar{x}_i) + b) - 1 + \xi_i) = 0$$

при $i = 1, \ldots l$.

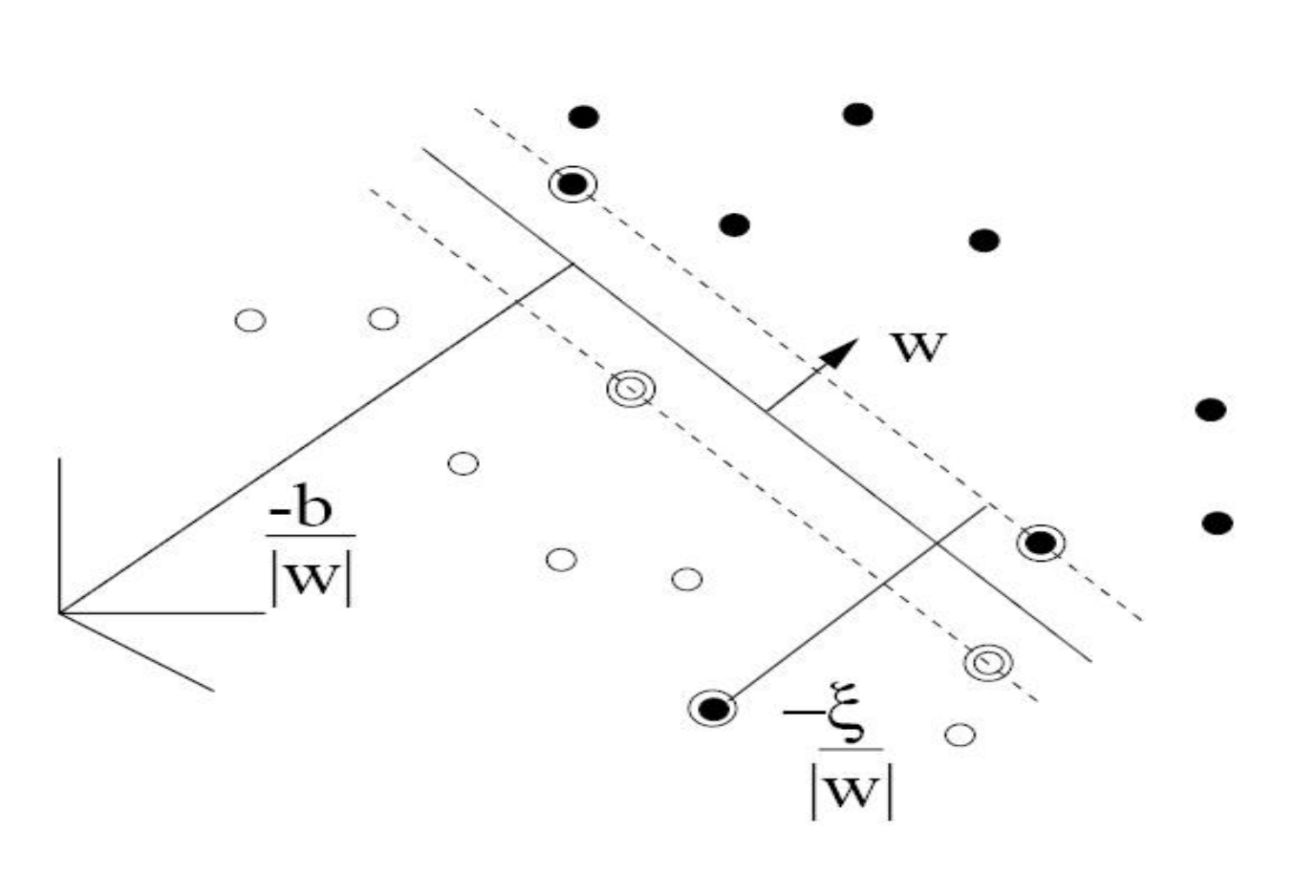


Задача оптимизации в линейной норме

Ищем вектора $\bar{w}$, $\bar{\xi}$ и число b, так чтобы

$$(\bar{w} \cdot \bar{w}) + C \sum_{i=1}^{l} \xi_i \to \min$$
$$y_i(((\bar{w} \cdot \bar{x}_i) + b_0) \geq 1 - \xi_i,$$
$$\xi_i \geq 0$$

при $i = 1, \ldots l$. Константа C определяет баланс между двумя частями функционала.



Оптимальная гиперплоскость для неразделимой выборки

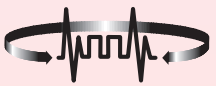
$\alpha_i = 0$, если

$$y_i((\bar{w} \cdot \bar{x}_i) + b) > 1, \qquad \xi_i = 0.$$

Опорные векторы – это те $\bar{x}_i$, где $\alpha_i > 0$, для них

$$y_i((\bar{w} \cdot \bar{x}_i) + b) \leq 1, \qquad \xi_i \geq 0,$$

при $i = 1, \ldots l$.
В числе опорных векторов могут находятся вектора, расположенные на границах гиперплоскостей и неправильно классифицируемые вектора.



Задача оптимизации для пространства с ядром $K(\bar{x}_i, \bar{x}_j)$

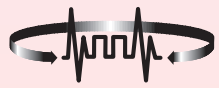
$$W(\bar{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(\bar{x}_i, \bar{x}_j) \to \max$$

$$\text{при условиях } \sum_{i=1}^{l} y_i \alpha_i = 0,$$

$$C \geq \alpha_i \geq 0 \; i = 1, \; \ldots, \; l.$$

$f(\bar{x}) = \sum\limits_{i=1}^{l} y_i \alpha_i^* K(\bar{x}_i, \bar{x}) + b^*$ – разделяющая поверхность.

Оценки ошибки обобщения не зависящие от размерности пространства



Число покрытия класса функций

ε-покрытие множества функций $\mathscr{F}$ относительно множества $X = \{\bar{x}_1, \ldots, \bar{x}_l\}$ – это конечное множество функций $\mathscr{B} \subseteq \mathscr{F}$ такое, что для любого $f \in \mathscr{F}$ существует $g \in \mathscr{B}$ такая, что $|f(\bar{x}_i) - g(\bar{x}_i)| < \varepsilon$ при $i = 1, \ldots, l$.
$\mathscr{N}(\mathscr{F}, X, \varepsilon)$ – размер $|\mathscr{B}|$ минимального такого покрытия – число покрытия.

$$\mathscr{N}(\mathscr{F}, l, \varepsilon) = \max_{|X|=l} \mathscr{N}(\mathscr{F}, X, \varepsilon)$$

$\log \mathscr{N}(\mathscr{F}, l, \varepsilon)$ – (ε, l)-энтропия класса функций $\mathscr{F}$.

Число упаковки класса функций $\mathscr{F}$

Множество функций $\mathscr{B} \subseteq \mathscr{F}$ – ε-различимо относительно множества $X = \{\bar{x}_1, \ldots, \bar{x}_l\}$, если $\forall f, g \in \mathscr{B}$, $f \neq g$,

$$\max_{x_i \in X} |f(x_i) - g(x_i)| > \varepsilon.$$

$\mathscr{M}(\mathscr{F}, \varepsilon, X)$ – размер $|\mathscr{B}|$ максимального ε-различимого на X подмножества $\mathscr{B} \subseteq \mathscr{F}$ – число упаковки.

Лемма:
$\mathscr{M}(\mathscr{F}, X, 2\varepsilon) \leq \mathscr{N}(\mathscr{F}, X, \varepsilon) \leq \mathscr{M}(\mathscr{F}, X, \varepsilon)$.

Множество $X = \{\bar{x}_1, \ldots, \bar{x}_l\}$ – γ-разделимо (γ-shattered), если существуют вещественные числа $r_1, \ldots, r_l$, такие что $\forall E \subseteq X \exists f_E \in \mathscr{F}$, так что

- $f_E(\bar{x}_i) > r_i + \gamma$, если $\bar{x}_i \in E$,
- $f_E(\bar{x}_i) \leq r_i - \gamma$, если $\bar{x}_i \notin E$.

Если $r_i = r$ для всех i, то – это одноуровневая γ-разделимость.

Пороговая (fat-) размерность

$\mathrm{fat}_{\mathscr{F}}(\gamma)$ – пороговая размерность или fat-размерность класса $\mathscr{F}$ равна мощности $|X|$ самого большого по количеству элементов γ-разделимого множества $X \subseteq \mathscr{R}^n$ с помощью функций из класса $\mathscr{F}$.
Пороговая (fat-) размерность может быть бесконечной.

По определению $\mathrm{fat}_{\mathscr{F}}(\gamma) \leq VCdim(H_{\mathscr{F}})$, где $H_{\mathscr{F}}$ – класс индикаторных функций, построенных по функциям из $\mathscr{F}$.

Оценка энтропии через fat-размерность

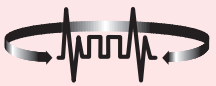
$\mathrm{fat}_{\mathscr{F}}(\gamma)$ – fat-размерность класса $\mathscr{F}$.

Теорема:

Пусть $\mathscr{F}$ – класс функций типа $\mathscr{R}^n \to [a, b]$, где $a < b$. Тогда

$$\log \mathscr{N}(\mathscr{F}, l, \gamma) \leq 1 + d \log \frac{2el(b-a))}{d\gamma} \log \frac{4l(b-a)^2}{\gamma^2},$$

где $d = \mathrm{fat}_{\mathscr{F}}(\gamma/4)$.



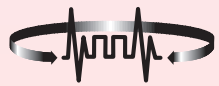
Идея доказательства:

Дискретизация:
$f^{\alpha}(x) = \lfloor \frac{f(x)}{\alpha} \rfloor$, где $f(x) \in [0,1]$. Тогда $f^{\alpha}(x) \in \{0,1,\ldots,\lfloor \frac{1}{\alpha} \rfloor\}$.
$\mathscr{F}^{\alpha} = \{f^{\alpha} : f \in \mathscr{F}\}$.

Класс $\mathscr{F}$ строго разделяет множество $X = \{\bar{x}_1, \ldots, \bar{x}_l\}$, если существуют вещественные числа $r_1, \ldots, r_l$, такие что
$\forall E \subseteq X \exists f_E \in \mathscr{F}$, так что

- $f_E(\bar{x}_i) > r_i + 1$, если $\bar{x}_i \in E$,
- $f_E(\bar{x}_i) \leq r_i - 1$, если $\bar{x}_i \notin E$.

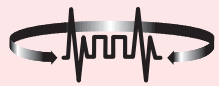
Строгая размерность:
$Sdim(\mathscr{F})$ – равно размеру наибольшего строго разделимого множества X.



Лемма 1:

1) $Sdim(\mathscr{F}^{\alpha}) \leq \mathrm{fat}_{\mathscr{F}}(\alpha/2)$

2) $\mathscr{M}(\mathscr{F}, X, \alpha) \leq \mathscr{M}(\mathscr{F}^{\alpha/2}, X, 2)$

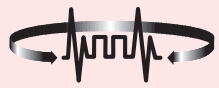


Лемма 2:

Пусть $|X| = l$, $B = \{0, 1, \ldots, b\}$, $\mathscr{F} \subseteq B^X$, $Sdim(\mathscr{F}) = d$. Тогда

$$\mathscr{M}(\mathscr{F}, X, 2) < 2(l(b+1)^2)^{\lceil \log y \rceil},$$

где $y = \sum_{i=1}^{d} \binom{l}{i} b^i$.



Лемма 3:

Пусть $\mathscr{F} \subseteq [0,1]^{\mathscr{R}^n}$, $0 < \alpha < 1$, $d = \mathrm{fat}_{\mathscr{F}}(\alpha/4)$. Тогда

$$\mathscr{N}(\mathscr{F}, l, \alpha) < 2\left(l\left(\frac{2}{\alpha}+1\right)^2\right)^{\lceil d \log(\frac{2el}{d\alpha})\rceil}.$$

Доказательство. $\mathscr{N}(\mathscr{F}, l, \alpha) = \sup_{|X|=l} \mathscr{N}(\mathscr{F}, X, \alpha) \leq$

$\sup_{|X|=l} \mathscr{M}(\mathscr{F}, X, \alpha) \leq \sup_{|X|=l} \mathscr{M}(\mathscr{F}^{\alpha/2}, X, 2) < 2(l(b+1)^2)^{\lceil \log y \rceil}$, где

$b = \lfloor \frac{2}{\alpha} \rfloor$, $d' = Sdim(\mathscr{F}^{\alpha/2}) \leq \mathrm{fat}_{\mathscr{F}}(\alpha/4) = d$,

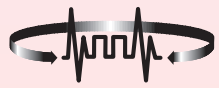
$y = \sum_{i=1}^{d'} \binom{l}{i} b^i \leq \sum_{i=1}^{d} \binom{l}{i} b^i \leq b^d \left(\frac{el}{d}\right)^d$, $\log y \leq \log(bel/d)$.

Оценка ошибки

Для произвольных $\varepsilon > 0$ и $\gamma > 0$

$$P^l\{S : (\exists\, f \in \mathscr{F})(\mathrm{err}_S(f) = 0 \& m_S(f) \geq \gamma \& \mathrm{err}_P(f) > \varepsilon)\} \leq$$
$$\leq 2\mathscr{N}(\mathscr{F}, 2l, \gamma/2)e^{-\frac{1}{2}\varepsilon l},$$

где $m_S(f) = \min\limits_{i=1,\ldots l} y_i f(\bar{x}_i)$.



PAC-формулировка

$\mathscr{F}$ – класс вещественных функций, $\gamma > 0$, $\delta > 0$.
Для любого распределения вероятностей P на $\mathscr{R} \times \{-1, 1\}$ с вероятностью $1 - \delta$ на случайной выборке S длины l для любой вещественной функции $f \in \mathscr{F}$, для которой $m_S(f) > \gamma$,

$$\mathrm{err}_P(f) \leq \frac{2}{l}\left(\ln \mathscr{N}(\mathscr{F}, 2l, \gamma/2) + \ln\frac{2}{\delta}\right),$$

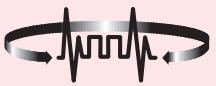
где $\mathrm{err}_P(f) = P\{\mathrm{sign}(f(\bar{x})) \neq y\} = P\{yf(\bar{x}) \leq 0\}$.

Следствие из оценки энтропии:

С вероятностью $1-\delta$ для любой гипотезы $f \in \mathscr{F}$, для которой $m_S(f) \geq \gamma$,

$$\mathrm{err}_P(f) \leq \frac{2}{l}\left(d \log \frac{8el}{d\gamma} \log \frac{32l}{\gamma^2} + \log \frac{4}{\delta}\right)$$

при $l > d$, где $d = \mathrm{fat}_{\mathscr{F}}(\gamma/8)$.



Оценка ошибки для ограниченных линейных функций:
X – шар радиуса R в n-мерном евклидовом пространстве и $\mathscr{F}$ – класс линейных однородных пороговых функций $f(\bar{x}) = (\bar{w} \cdot \bar{x})$, где $|\bar{w}| \leq 1$ и $\bar{x} \in X$. Тогда

$$\text{fat}_{\mathscr{F}}(\gamma) \leq \left(\frac{R}{\gamma}\right)^2.$$

В результате получаем оценку ошибки (dimension free)

$$\text{err}_P(f) \approx \frac{1}{l}\left(\frac{R}{\gamma}\right)^2,$$

где $\text{err}_P(f) = P\{yf(\bar{x}) \leq 0\}$.

Вероятность ошибки для SVM с квадратичной нормой и с переменными мягкого отступа

$$\mathrm{err}_P(f) = P\{(\bar{x}, y) : yf(\bar{x}) < 0\} \leq$$
$$\frac{c}{l}((|\bar{w}|^2 R^2 + |\bar{\xi}|^2)\log^2 l + \log\frac{1}{\delta}),$$

где $\mathrm{err}_P(f) = P\{yf(\bar{x}) \leq 0\}$.

Классификация с K классами (разделимая выборка)

Выборка $S = ((x_1^1, y_1^1), \ldots, (x_{l_k}^1, y_{l_k}^1), \ldots, (x_1^K, y_1^K), \ldots, (x_{l_K}^K, y_{l_K}^K))$ – K – меток классов.

Функции классификации:

$f_k(\bar{x}) = (\bar{w}^k \cdot \bar{x}) + b_k,\ k = 1, \ldots, K$

Задача оптимизации:

$$\sum_{k=1}^{K} (\bar{w}^k \cdot \bar{w}^k) = \sum_{k=1}^{K} \sum_{i=1}^{l} (w_i^k)^2 \to \min$$

при условиях $y_i((\bar{w}^k \cdot \bar{x}_i^k) + b_k) - y_i((\bar{w}^m \cdot \bar{x}_i^k) + b_m) \geq 1$

при $m, k = 1, \ldots K,\ m \neq k,\ i = 1, \ldots l_k$.

Классификация с K классами (неразделимая выборка)

Выборка $S = ((x_1^1, y_1^1), \ldots, (x_{l_k}^1, y_{l_k}^1), \ldots, (x_1^K, y_1^K), \ldots, (x_{l_K}^K, y_{l_K}^K))$ – K – меток классов.

Функции классификации:

$f_k(\bar{x}) = (\bar{w}^k \cdot \bar{x}) + b_k,\ k = 1, \ldots, K$

Задача оптимизации:

$$\sum_{k=1}^{K} (\bar{w}^k \cdot \bar{w}^k) + \sum_{i,j=1, i \neq j} \xi_{i,j} \to \min$$

при условиях $y_i((\bar{w}^k \cdot \bar{x}_i^k) + b_k) - y_i((\bar{w}^m \cdot \bar{x}_i^k) + b_m) \geq 1 - \xi_{k,m}$

$$\xi_{k,m} \geq 0$$

при $m, k = 1, \ldots K,\ m \neq k,\ i = 1, \ldots l_k$.