Байесовские постановки задачи обучения распознаванию образов по методу опорных векторов с управляемой селективностью отбора информативных признаков объектов

Татарчук Александр Игоревич
Моттль Вадим Вячеславович
Вычислительный центр РАН

Типовая задача распознавания образов в множествах объектов реального мира

Наблюдателя интересует некоторое множество объектов реального мира: $\omega \in \Omega$.

Пара характеристик:

$$\begin{cases} \mathbf{x}(\omega) \in \mathbb{X} - \text{наблюдаемая}, \\ y(\omega) \in \mathbb{Y} = \{-1,1\} - \text{скрытая (индекс класса)}. \end{cases}$$

Обучающая выборка:

$$\Omega' \subset \Omega : \left\{ (\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N) \right\}.$$

Решающая функция:

$$\hat{y}(\mathbf{x}) : \mathbb{X} \to \mathbb{Y},$$

$$\hat{y}(\mathbf{x}(\omega)) \neq y(\omega) - \text{ошибка}.$$

Линейный подход к обучению распознавания двух классов

Вектор действительных признаков $\mathbf{x}(\omega) \in \mathbb{R}^n$ погружает множество объектов в $\mathbb{R}^n$.

Евклидова метрика в $\mathbb{R}^n$:

$$\rho(\mathbf{x}', \mathbf{x}'') = \| \mathbf{x}' - \mathbf{x}'' \| = \left((\mathbf{x}' - \mathbf{x}'')^T (\mathbf{x}' - \mathbf{x}'') \right)^{1/2}.$$

Разделяющая гиперплоскость:

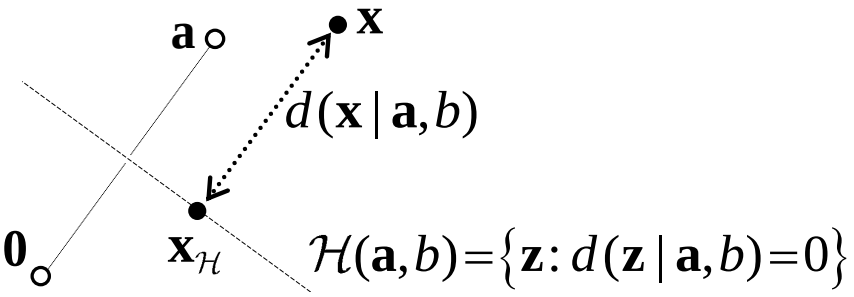
$$\mathcal{H}(\mathbf{a}, b) = \left\{ \mathbf{x} \in \mathbb{R}^n : \mathbf{a}^T \mathbf{x} + b = 0 \right\},\ \mathbf{a} \in \mathbb{R}^n.$$

Решающая функция:

$$d(\mathbf{x} \mid \mathbf{a}, b) = \mathbf{a}^T \mathbf{x} + b \gtrless 0.$$

Свойство решающей функции:

$$| d(\mathbf{x} \mid \mathbf{a}, b) | = \rho(\mathbf{x}, \mathbf{x}_{\mathcal{H}}),\ \mathbf{x}, \mathbf{x}_{\mathcal{H}} \in \mathbb{R}^n,\ \| \mathbf{a} \| = 1.$$

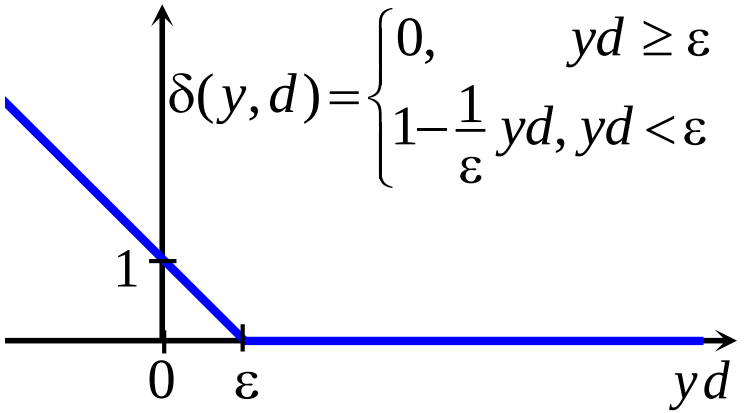


Некоторая конкретизация гипотезы компактности Э.М. Бравермана

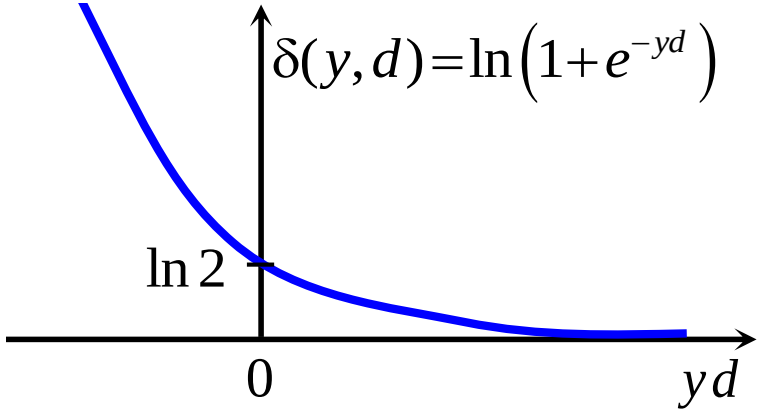
Функция потерь: Степень несоответствия значения решающей функции скрытой характеристике объекта

Индекс класса объекта $y \in \{-1, 1\}$

Метод опорных векторов



Метод логистической регрессии



Принцип обучения

Разделение обучающей выборки:

- с наибольшим зазором $\varepsilon \to \max$,
- с наименьшей суммой потерь

$$\sum_{j=1}^{N} \delta(y_j, d(\mathbf{x}_j \mid \mathbf{a}, b)) \to \min(\mathbf{a}, b).$$

} противоречивые требования

Предложенный невыпуклый критерий обучения

$$\begin{cases} \varepsilon^{-2} + C\sum_{j=1}^{N} \delta_j \to \min(\mathbf{a}, b, \boldsymbol{\delta}, \varepsilon), \;\; \mathbf{a}^T\mathbf{a} = 1, \\ y_j(\mathbf{a}^T\mathbf{x}_j + b) \geq \varepsilon(1-\delta_j),\; \delta_j \geq 0,\; j = 1,\ldots,N,\; \varepsilon \geq 0. \end{cases}$$

Утверждение: Предложенный критерий эквивалентен классическому критерию метода опорных векторов

$$\begin{cases} \mathbf{a}^T\mathbf{a} + C\sum_{j=1}^{N} \delta_j \to \min(\mathbf{a}, b, \boldsymbol{\delta}), \\ y_j(\mathbf{a}^T\mathbf{x}_j + b) \geq 1-\delta_j,\; \delta_j \geq 0,\; j = 1,\ldots,N. \end{cases}$$

Свойства решающей функция для нового объекта:

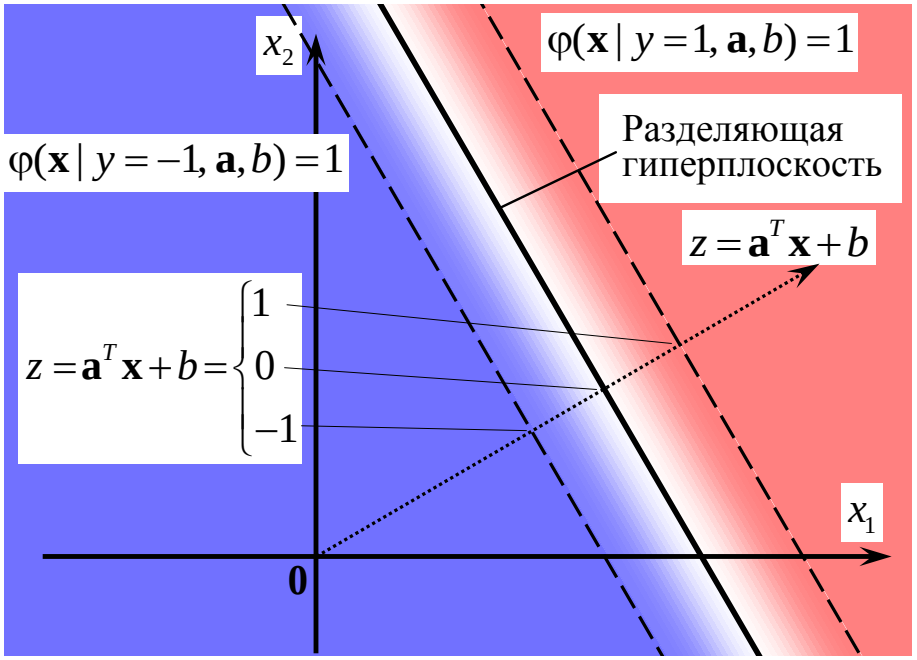
$$d(\mathbf{x} \mid \hat{\mathbf{a}}, \hat{b}) = \hat{\mathbf{a}}^T\mathbf{x} + \hat{b} = \sum_{j:\; \hat{\lambda}_j > 0} \hat{\lambda}_j y_j(\mathbf{x}_j^T\mathbf{x}) + b \gtrless 0$$

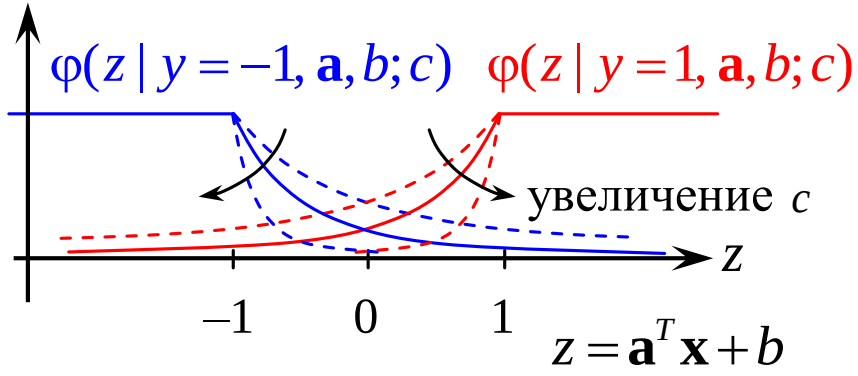
Требуются только **скалярные произведения** $\mathbf{x}_j^T\mathbf{x}$ нового вектора признаков и **только с опорными векторами.**

Предложенная вероятностная модель наблюдения объектов в признаковом пространстве

Предположение 1: Объекты двух классов $y = \pm 1$ распределены преимущественно по разные стороны некоторой гиперплоскости в пространстве признаков $\mathbf{x} \in \mathbb{R}^n$.

$$\varphi(\mathbf{x}|y,\mathbf{a},b;c) \propto \begin{cases} 1, & y(\mathbf{a}^T\mathbf{x}+b) \geq 1, \\ \exp\left[-c\left(1-y(\mathbf{a}^T\mathbf{x}+b)\right)\right], & y(\mathbf{a}^T\mathbf{x}+b) < 1, \end{cases} \qquad \mathbf{a} \in \mathbb{R}^n,\ b \in \mathbb{R},\ c > 0.$$





Предположение 2: Несобственная совместная априорная плотность параметров гиперплоскости $(\mathbf{a}, b)$ не зависит от b

$$\Psi(\mathbf{a},b) \propto \Psi(\mathbf{a}).$$

Байесовский критерий обучения для пары несобственных плотностей распределения

Критерий обучения: Максимизация апостериорной плотности вероятностей параметров гиперплоскости

$$(\hat{\mathbf{a}}, \hat{b} \mid X, Y; c) = \arg\max_{\mathbf{a},b} \left[\ln \Psi(\mathbf{a}) + \sum_{j=1}^{N} \ln \varphi(\mathbf{x}_j \mid y_j, \mathbf{a}, b; c) \right].$$

Теорема 1. *Критерий обучения для заданных плотностей эквивалентен задаче*

$$\begin{cases} -\ln \Psi(\mathbf{a}) + c \sum_{j=1}^{N} \delta_j \to \min(\mathbf{a}, b, \boldsymbol{\delta}), \\ y_j (\mathbf{a}^T \mathbf{x}_j + b) \geq 1 - \delta_j, \; \delta_j \geq 0, \; j = 1, \ldots, N. \end{cases}$$

Свойства критерия:

Если $\ln \Psi(\mathbf{a})$ вогнута, то критерий выпуклый и точка минимума $(\hat{\mathbf{a}}, \hat{b}, \hat{\boldsymbol{\delta}})$ единственна.

Активные ограничения $y_j (\mathbf{a}^T \mathbf{x}_j + b) = 1 - \delta_j$ определяют опорные векторы $\mathbf{x}_j$.

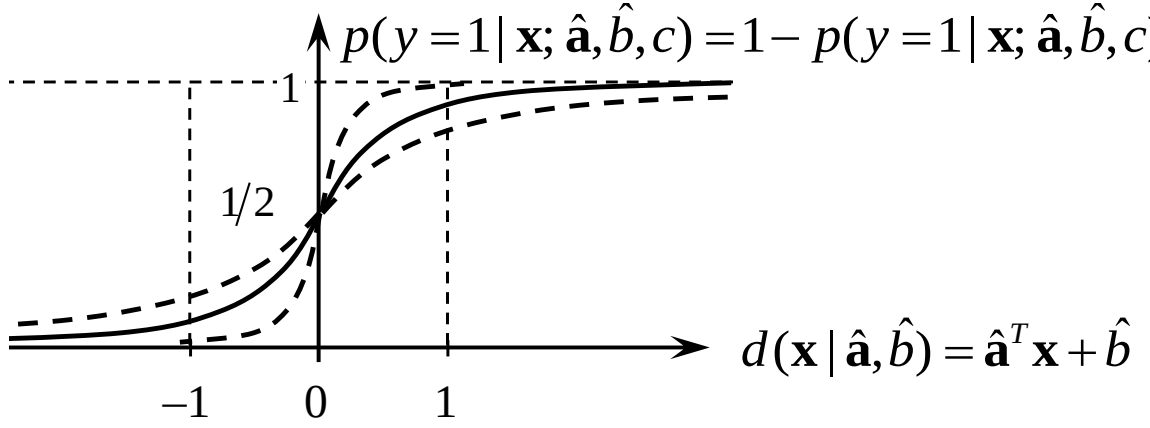
Решающая функция зависит только от опорных векторов.

Апостериорные вероятности классов: для случая равных априорных вероятностей

Теорема 2: *Для случая равных априорных вероятностей* $q_{-1} = q_1 = 1/2$ *апостериорные вероятности классов будут иметь вид сигмоидоподобной функции*

$$p(y=1|\mathbf{x},\hat{\mathbf{a}},\hat{b};c) = 1 - p(y=-1|\mathbf{x},\hat{\mathbf{a}},\hat{b};c) = \begin{cases} \left\{1+\exp(c)\exp[-c(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})]\right\}^{-1}, & \hat{\mathbf{a}}^T\mathbf{x}+\hat{b} < -1, \\ \left\{1+\exp[-2c(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})]\right\}^{-1}, & -1 \le \hat{\mathbf{a}}^T\mathbf{x}+\hat{b} \le 1, \\ \left\{1+\exp(-c)\exp[-c(\hat{\mathbf{a}}^T\mathbf{x}+\hat{b})]\right\}^{-1}, & \hat{\mathbf{a}}^T\mathbf{x}+\hat{b} > 1. \end{cases}$$

Графическое представление апостериорных вероятностей классов



Параметр $c > 0$ управляет крутизной сигмоидоподобного обобщения ступенчатой решающей функции $\hat{\mathbf{a}}^T\mathbf{x}+\hat{b} \gtrless 0$.

Известные частные случаи критериев обучения

Нормальная плотность	$\Psi(\mathbf{a}) = \Psi(a_1,...,a_n) = \prod_{i=1}^{n} \mathcal{N}(a_i \mid 0, r) = \frac{1}{r^{n/2}(2\pi)^{n/2}} \exp\left(-\frac{1}{2r}\mathbf{a}^T\mathbf{a}\right)$
Метод опорных векторов	$\begin{cases} \mathbf{a}^T\mathbf{a} + C\sum_{j=1}^{N}\delta_j \to \min(\mathbf{a}, b), \quad C = 2rc, \\ y_j(\mathbf{a}^T\mathbf{x}_j + b) \geq 1 - \delta_j,\ \delta_j \geq 0,\ j = 1,...,N. \end{cases}$
Плотность Лапласа	$\Psi(\mathbf{a}\mid\beta,\mu) = (2r)^{-n/2}\exp\left(-(r/2)^{-1/2}\sum_{i=1}^{n}\mid a_i \mid\right).$
Lasso SVM (1-norm SVM)	$\begin{cases} \sum_{i=1}^{n}\mid a_i \mid + c(r/2)^{1/2}\sum_{j=1}^{N}\delta_j \to \min(\mathbf{a}, b, \boldsymbol{\delta}), \\ y_j(\mathbf{a}^T\mathbf{x}_j + b) \geq 1 - \delta_j,\ \delta_j \geq 0,\ j = 1,...,N. \end{cases}$
Специальная плотность	$\Psi(\mathbf{a}\mid\beta,\mu) = D^n \exp\left[-\sum_{i=1}^{n}(\beta a_i^2 + \mu\mid a_i \mid)\right],\ \beta \geq 0,\ \mu \geq 0.$
Нормирующая константа (доказано Е. Черноусовой)	$D = \left[2\sqrt{\frac{\pi}{\beta}}\exp\left(\frac{\mu^2}{4\beta}\right)\Phi\left(\frac{\mu}{\sqrt{2\beta}}\right)\right]^{-1},\ \Phi(u) = \frac{1}{\sqrt{2\pi}}\int_u^{\infty}\exp\left(-\frac{z^2}{2}\right)dz.$
Elastic Net SVM (DrSVM)	$\begin{cases} \beta\sum_{i=1}^{n}a_i^2 + \mu\sum_{i=1}^{n}\mid a_i \mid + c\sum_{j=1}^{N}\delta_j \to \min(\mathbf{a}, b, \boldsymbol{\delta}), \\ y_j(\mathbf{a}^T\mathbf{x}_j + b) \geq 1 - \delta_j,\ \delta_j \geq 0,\ j = 1,...,N. \end{cases}$

Первый предлагаемый метод: Метод релевантных признаков

Принцип: Селекция признаков выполняется путем совместного байесовского оценивания направляющего вектора и дисперсий его компонент.

Нормальные условные распределения параметров гиперплоскости

$$\Psi(\mathbf{a}\,|\,\mathbf{r}) = \prod_{i=1}^{n} \mathcal{N}(a_i\,|\,0, r_i) = \left(\prod_{i=1}^{n} r_i\right)^{-1/2} (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n} (1/r_i)\, a_i^2\right).$$

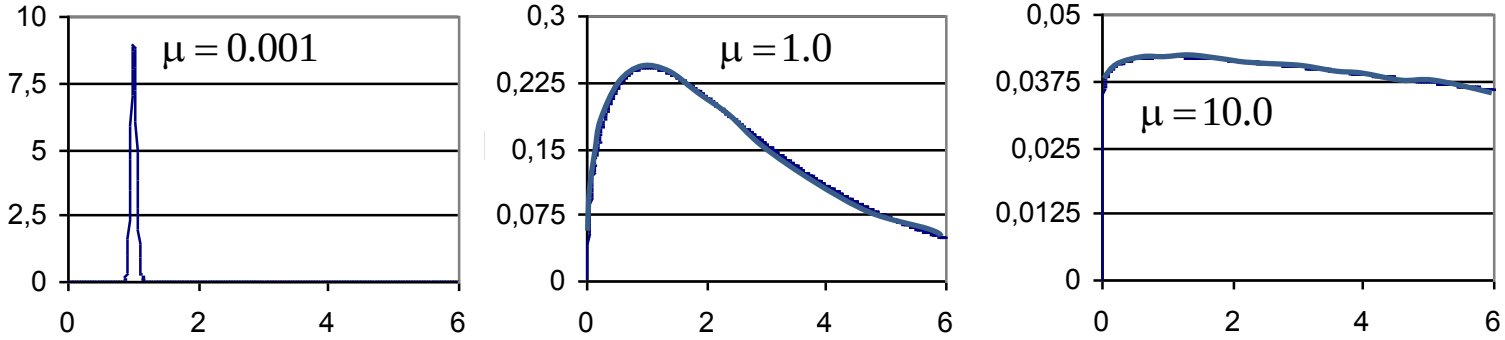
Независимые гамма распределения обратных дисперсий:

$$G(\mathbf{r}\,|\,\mu) = \prod_{i=1}^{n} \gamma(1/r_i\,|\,\mu),$$

$$\gamma(1/r_i\,|\,\mu) \propto (1/r_i)^{1/2\mu} \exp\left[-(1/2\mu)(1/r_i)\right],$$

$\mu \geq 0$ – **параметр селективности.**

Зависимость гамма распределения от параметра селективности



Критерий обучения

Теорема 3. *Критерий обучения эквивалентен оптимизационной задаче*

$$\begin{cases} \sum_{i=1}^{n}\left[(1/r_i)\left(a_i^2+(1/\mu)\right)+\left((1/\mu)+1+\mu\right)\ln r_i\right]+C\sum_{j=1}^{N}\delta_j \to \min(\mathbf{a},b,\mathbf{r},\boldsymbol{\delta}), \\ y_j\left(\mathbf{a}^T\mathbf{x}+b\right)\geq 1-\delta_j,\ \delta_j\geq 0,\ j=1,\ldots,N, \quad C=2c. \end{cases}$$

Чем больше дисперсия r_i, тем больше вес i-го признака в решающем правиле.

$$d(\mathbf{x}\,|\,\hat{\mathbf{a}},\hat{b},\hat{\mathbf{r}}) = \sum_{i=1}^{n} r_i \sum_{j:\,\lambda_j>0} y_j\lambda_j x_{ij}x_i + b \underset{<}{>} 0.$$

Итерационный алгоритм обучения

Покоординатная оптимизация двух групп переменных	$\mathbf{r}^k=(r_i^k=1,\ i=1,\ldots,n) \Rightarrow (\mathbf{a}^k,b^k),\ (\delta_j^k,\ j=1,\ldots,N),$ $\mathbf{a}^k \Rightarrow \mathbf{r}^{k+1},\quad \mathbf{r}^{k+1} \Rightarrow (\mathbf{a}^{k+1},b^{k+1})$. Начальное приближение $\mathbf{r}^0=\mathbf{1}$.
Шаг 1: $\mathbf{r}^k \Rightarrow (\mathbf{a}^k,b^k)$	$\begin{cases} (\mathbf{a}^k,b^k,\boldsymbol{\delta}^k)=\arg\min\limits_{\mathbf{a},b,\boldsymbol{\delta}} \sum_{i=1}^{n}(1/r^k)\,a_i^2 + C\sum_{j=1}^{N}\delta_j \to \min(\mathbf{a},b), \\ y_j\left(\mathbf{a}^T\mathbf{x}+b\right)\geq 1-\delta_j,\ \delta_j\geq 0,\ j=1,\ldots,N. \end{cases}$
Шаг 2: $\mathbf{a}^k \Rightarrow \mathbf{r}^{k+1}$	**Теорема 4**. *Новые значения дисперсий* $r_i^{k+1}=\arg\min\limits_{r_i}\left[(1/r_i)\left((a_i^k)^2+(1/\mu)\right)+\left((1/\mu)+1+\mu\right)\ln r_i\right]=\dfrac{(a_i^k)^2+1/\mu}{\mu+1+1/\mu}$
Критерий останова	$\|\,\mathbf{a}^k-\mathbf{a}^{k+1}\,\|<\varepsilon$. Алгоритм обычно сходится за 10-15 шагов.

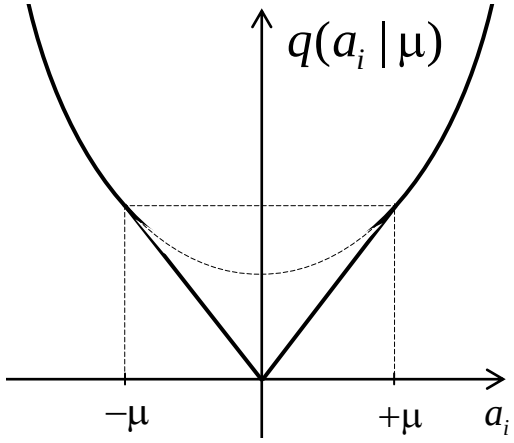
Второй предлагаемый метод: Метод опорных признаков

Принцип: Априорная плотность $\Psi(\mathbf{a})$ есть комбинация распределения Лапласа при небольших значениях нормы $|a| \le \mu$ и нормального распределения при бóльших $|a_i| > \mu$.

Априорное распределение независимых компонент направляющего вектора

$$\psi(a_i \mid \mu) \propto \exp\bigl(-q(a_i \mid \mu)\bigr), \quad i = 1, \ldots, n,$$

$$q(a_i \mid \mu) = \begin{cases} 2\mu |a_i|, & |a_i| \le \mu, \\ \mu^2 + a_i^2, & |a_i| > \mu. \end{cases}$$



$\mu \ge 0$ – **параметр селективности**.

Теорема 5. *Критерий обучения эквивалентен оптимизационной задаче*

$$\begin{cases} J_{SFM}(\mathbf{a}, b, \boldsymbol{\delta} \mid C, \mu) = 2\mu \sum\limits_{|a_i| \le \mu} |a_i| + \sum\limits_{|a_i| > \mu} (\mu^2 + a_i^2) + C \sum\limits_{j=1}^{N} \delta_j \to \min(\mathbf{a}, b, \boldsymbol{\delta}), \\ y_j \left(\mathbf{a}^T \mathbf{x} + b\right) \le 1 - \delta_j, \; \delta_j \ge 0, \; j = 1, \ldots, N. \end{cases}$$

Задача выпуклой оптимизации.

Двойственная задача оптимизационная задача обучения

Теорема 5. *Критерий обучения имеет двойственную оптимизационную задачу*

$$\begin{cases} L(\boldsymbol{\lambda},\boldsymbol{\xi}\,|\,C,\mu)=\sum_{j=1}^{N}\lambda_j-\sum_{i\in I}\sum_{i=2}^{n}(1/2)\xi_i\to\max(\boldsymbol{\lambda},\boldsymbol{\xi}),\\ \xi_i\geq 0,\ \ \xi_i\geq\left(\sum_{j=1}^{N}y_j x_{ij}\hat{\lambda}_j\right)^2-\mu^2,\ \ i\in I=\{1,\ldots,n\},\\ \sum_{j=1}^{N}y_j\lambda_j=0,\ \ 0\leq\lambda_j\leq C/2,\ \ j=1,\ldots,N. \end{cases}$$

Решение $(\hat{\lambda}_1,\ldots,\hat{\lambda}_N)$ *определяет разбиение множества компонент направляющего:*

$$\begin{cases} \hat{a}_i=\sum_{j:\hat{\lambda}_j>0}y_j\hat{\lambda}_j x_{ij}, & i\in I^+=\left\{i\in I:\left(\sum_{j=1}^{N}y_j x_{ij}\hat{\lambda}_j\right)^2-\mu^2>0\right\},\\ \hat{a}_i=\hat{\eta}_i\sum_{j:\hat{\lambda}_j>0}y_j\hat{\lambda}_j x_{ij}, & i\in I^0=\left\{i\in I:\left(\sum_{j=1}^{N}y_j x_{ij}\hat{\lambda}_j\right)^2-\mu^2=0\right\},\\ \hat{a}_i=0, & i\in I^-=\left\{i\in I:\left(\sum_{j=1}^{N}y_j x_{ij}\hat{\lambda}_j\right)^2-\mu^2<0\right\}. \end{cases}$$

опорные признаки

удаленные признаки

Теорема 6. *Итоговое решение о классе нового объекта имеет вид*

$$d(\mathbf{x}\,|\,\hat{\boldsymbol{\lambda}},\eta_i,i\in I^0,b)=\underbrace{\sum_{j:\hat{\lambda}_j>0}y_j\hat{\lambda}_j\underbrace{\left[\sum_{i\in I^+}x_{ij}x_i+\sum_{i\in I^0}\hat{\eta}_i x_{ij}x_i\right]}_{\textbf{опорные признаки}}}_{\text{опорные векторы}}+\hat{b}\gtrless 0,$$

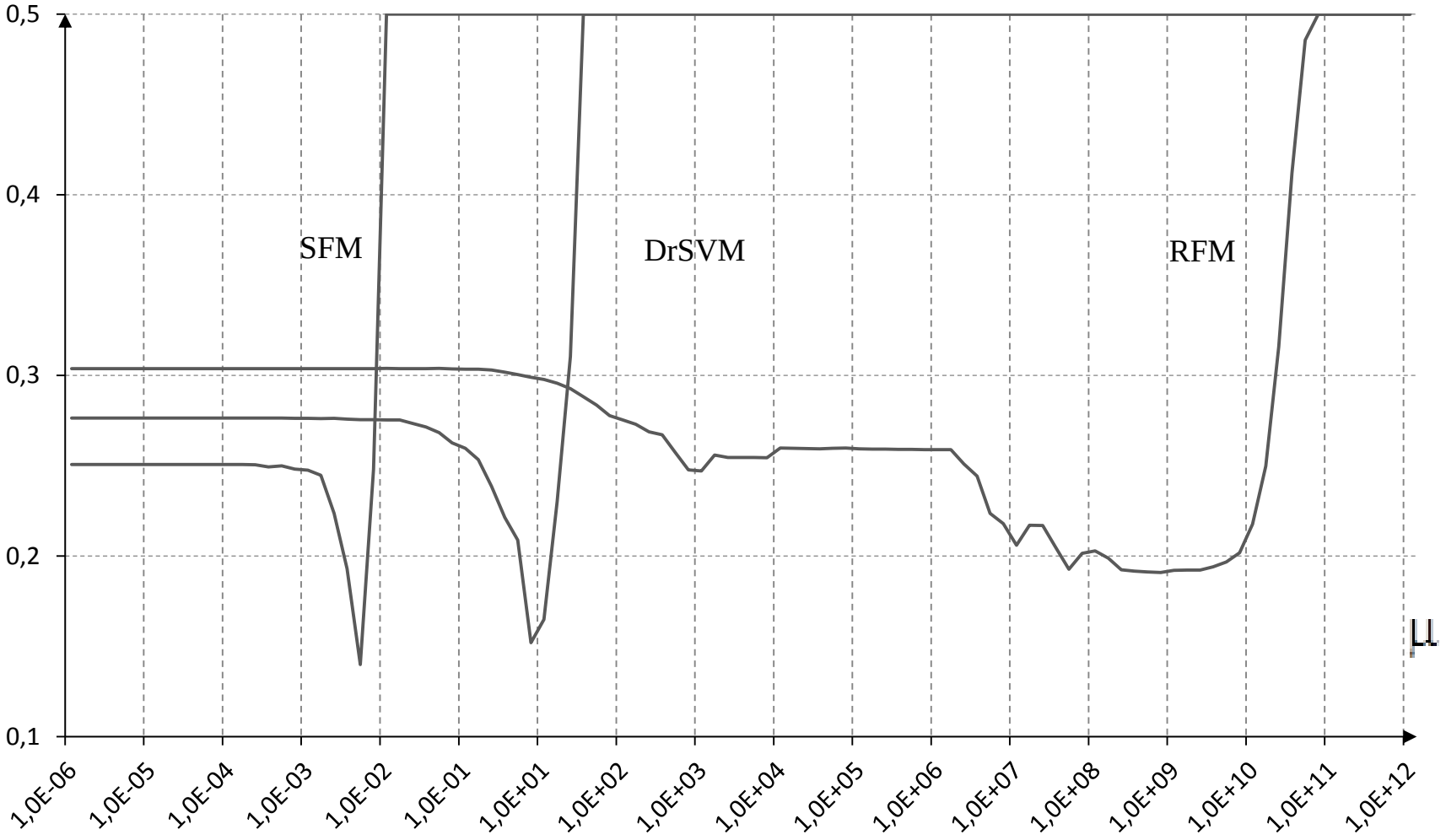
где коэффициенты $\hat{\eta}_i$ *и* $\hat{b}$ *определяются задачей линейного программирования.*

Результаты экспериментального исследования: Структура модельного эксперимента

Модель данных	100-мерное признаковое пространство $\mathbf{x} = (x_1, ..., x_{100})$. Наблюдения сгенерированы независимо по многомерному нормальному закону распределения с математическими ожиданиями *первый класс* *второй класс* $\mathbf{m}_+ = (\underbrace{0.5, ..., 0.5}_{5}, \underbrace{0, ..., 0}_{95})^T \in \mathbb{R}^{100}$ и $\mathbf{m}_- = (\underbrace{-0.5, ..., -0.5}_{5}, \underbrace{0, ..., 0}_{95})^T \in \mathbb{R}^{100}$, *информативные признаки* *шумовые признаки* *информативные признаки* *шумовые признаки* единичной ковариационной матрицей $\mathbf{\Sigma} = \mathbf{I}_{n \times n} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$. **Все признаки независимы. Информативные признаки равнозначны.**
Обучающая совокупность	$N = N_1 = N_{-1} = 50 + 50 = 100$ векторов $\mathbf{x}_j$, $j = 1, ..., N$ двух классов $y_j = \pm 1$, выбранных случайно и независимо.
Обучение классификатора	Оценивание параметров гиперплоскости $(a_1, ..., a_{100}, b)$ независимо по 100 обучающим совокупностям для каждого набора пар значений параметра селективности $10^{-6} \le \mu \le 10^{12}$ и разделимости $10^{-6} \le C \le 10^{12}$.
Оценивание качества	Тестовая совокупность $N_{test} = 50\,000 + 50\,000 = 100\,000$ случайных векторов.

Модельный эксперимент (продолжение)

Доля ошибочно классифицированных обектов тестовой выборки



Пример эффекта регуляризации методов на одной из модельных обучающих совокупностей для возрастающих значений параметра селективности μ при фиксированных оптимальных значениях параметра разделимости $\hat{C}$.

Итоги серии модельных экспериментов

Метод обучения		**Минимальная средняя ошибка на тестовой совокупности** по 100 обучающим совокупностям			
		Информативные признаки линейно независимы		Только совместно информативные признаки обеспечивают приемлемую точность	
		Шумовые признаки независимы	Шумовые признаки линейно зависимы	Шумовые признаки независимы	Шумовые признаки линейно зависимы
Предложенные	**SFM**	**0.1495**	**0.1429**	0.3140	0.2364
	RFM	0.1797	0.1611	**0.2325**	**0.2212**
Существующие	DrSVM	0.1523	0.1489	0.3140	0.3406
	SVM	0.2353	0.1620	0.4359	0.2364
	SVM на 5 заведомо информативных признаках	**0.1430**	**0.1397**	**0.2238**	**0.2177**
	«Оракульная» ошибка	**0.1320**	**0.132**	**0.2150**	**0.215**

Основные результаты работы

1. Общая математическая постановка задачи обучения двухклассовому распознаванию образов в линейном пространстве признаков на основе количественного измерения расстояния между вектором признаков объекта и разделяющей гиперплоскостью.

2. Вероятностная модель наблюдения объектов в пространстве векторов признаков относительно фиксированной разделяющей гиперплоскости.

3. Два семейства априорных вероятностных моделей направляющего вектора разделяющей гиперплоскости, отражающих стратегии отбора признаков на основе взвешивания всех исходно заданных признаков (feature weighting) и на основе жесткого выбора их подмножества (feature subset selection).

4. Комплекс байесовских методов и алгоритмов оценивания разделяющей гиперплоскости, реализующих принцип обучения с заданной селективностью отбора признаков объектов.