2. Обучение распознавания образов в линейных пространствах признаков.

2.1 Концепция оптимальной разделяющей гиперплоскости.

Рассмотрим теперь концепцию оптимальной разделяющей гиперплоскости. Пусть предъявлена обучающая выборка $(\mathbf{x}_j, g_j)$, $\mathbf{x}_j \in \mathsf{R}^{n+1}$, $g_j \in \{1,-1\}$ $j = 1,..., N$. По своей идее искомая разделяющая гиперплоскость $\mathbf{a}^T\mathbf{x} + a^{(0)} = 0$ призвана как можно лучше отделять друг от друга точки первого и второго класса. Поскольку единственное, что известно о границах областей классов в пространстве признаков, полностью содержится в обучающей выборке, то представляется естественным выбрать вектор $\mathbf{a} \in \mathsf{R}^n$ и скаляр $a^{(0)}$ так, чтобы

$$\mathbf{a}^T\mathbf{x}_j + a^{(0)} \begin{cases} > 0, & \text{если } g_j = 1, \\ < 0, & \text{если } g_j = -1. \end{cases}$$

Такая пара $(\mathbf{a}, a^{(0)})$ существует, если выпуклые оболочки подвыборок первого и второго класса в обучающей совокупности не пересекаются, причем в этом случае существует множество таких пар. Среди них естественно выбрать ту, которая определяет гиперплоскость, наиболее удаленную от краев подвыборок. Заметим, что нетривиальна лишь задача поиска направляющего вектора $\mathbf{a} \in \mathsf{R}^n$, качество которого естественно оценивать величиной остаточного "зазора":

$$J(\mathbf{a}) = \min_{j:\, g_j=1} \mathbf{a}^T\mathbf{x}_j - \max_{j:\, g_j=-1} \mathbf{a}^T\mathbf{x}_j . \qquad (2.1.1)$$

Нам будет удобно ввести специальное обозначение для скалярного произведения $\mathbf{a}^T\mathbf{x}_j$, рассматривая его как промежуточный критерий в составе полного критерия (1.2.16), количественно характеризующий определенную точку первого или второго класса относительно всей подвыборки этого класса. Пока мы примем оба критерия одинаковыми

$$Q^{(1)}(j, \mathbf{a}) = Q^{(-1)}(j, \mathbf{a}) = \mathbf{a}^T\mathbf{x}_j , \qquad (2.1.2)$$

однако ниже мы обобщим эти понятия.

С учетом принятых обозначений критерий качества направляющего вектора примет вид

$$J(\mathbf{a}) = \min_{j:g_j=1} Q^{(1)}(j, \mathbf{a}) - \max_{j:g_j=-1} Q^{(-1)}(j, \mathbf{a}) . \qquad (2.1.3)$$

После того, как вектор $\mathbf{a}$ выбран, оптимальное значение скаляра $a^{(0)}$ определяется как среднее значение

$$a^{(0)} = \frac{1}{2}\left(\min_{j:g_j=1} Q^{(1)}(j,\mathbf{a}) + \max_{j:g_j=-1} Q^{(-1)}(j,\mathbf{a}) \right). \qquad (2.1.4)$$

Заметим, что существенно только соотношение между значениями элементов вектора $\mathbf{a}$, но не их величины, поэтому направляющий вектор разделяющей гиперплоскости достаточно выбирать среди векторов единичной нормы $\|\mathbf{a}\| = (\mathbf{a}^T\mathbf{a})^{1/2} = 1$. Таким образом, обучение сводится к решению задачи

$$\hat{\mathbf{a}} = \underset{\mathbf{a}\in R^n, \|\mathbf{a}\|=1}{\operatorname{argmax}} J(\mathbf{a}) \qquad (2.1.5).$$

Такая задача остается полностью корректной с точки зрения конечной цели обучения и в случае пересекающихся выпуклых оболочек подвыборок первого и второго класса. При этом, наибольшее возможное значение критерия будет отрицательным $J(\hat{\mathbf{a}}) < 0$, и направляющий вектор, найденный согласно (2.1.2), (2.1.3) и (2.1.1), будет определять гиперплоскость, обеспечивающую наименьший по абсолютной величине остаточный пространственный дефицит разделения подвыборок (рис. 2.1.1). Такая гиперплоскость называется оптимальной. В.Н. Вапник, опираясь на выпуклость критерия $J(\mathbf{a})$, показывает, что оптимальная гиперплоскость единственна.

Заметим, что задача (2.1.5) представляет собой задачу максимизации кусочно-линейной функции при квадратичном ограничении типа равенства. Впрочем, возможны и другие эквивалентные представления этой задачи. В частности, она может быть переформулирована как задача минимизации квадратичной функции при совокупности ограничений в виде линейных неравенств [55].

Можно показать, что каковы бы ни были обучающие подвыборки первого и второго класса, из них всегда можно удалить часть точек так, что оптимальной решение $\hat{\mathbf{a}}$ для оставшихся точек будет в точности таким же, как и для выборки в целом. Минимальное число точек, которое надо оставить, чтобы получающаяся гиперплоскость не изменилась, зависит от конкретной конфигурации подвыборок, но оно всегда не меньше двух, по одной точке первого и второго класса, и не больше $n+1$, т.е. на единицу больше исходной размерности n пространства признаков. Именно эти точки подвыборок и определяют оптимальную разделяющую гиперплоскость, она как бы “опирается” на них, в силу чего такие точки называют опорными. Это самые крайние точки подвыборок с тех сторон, которыми они обращены друг к другу в R^n.

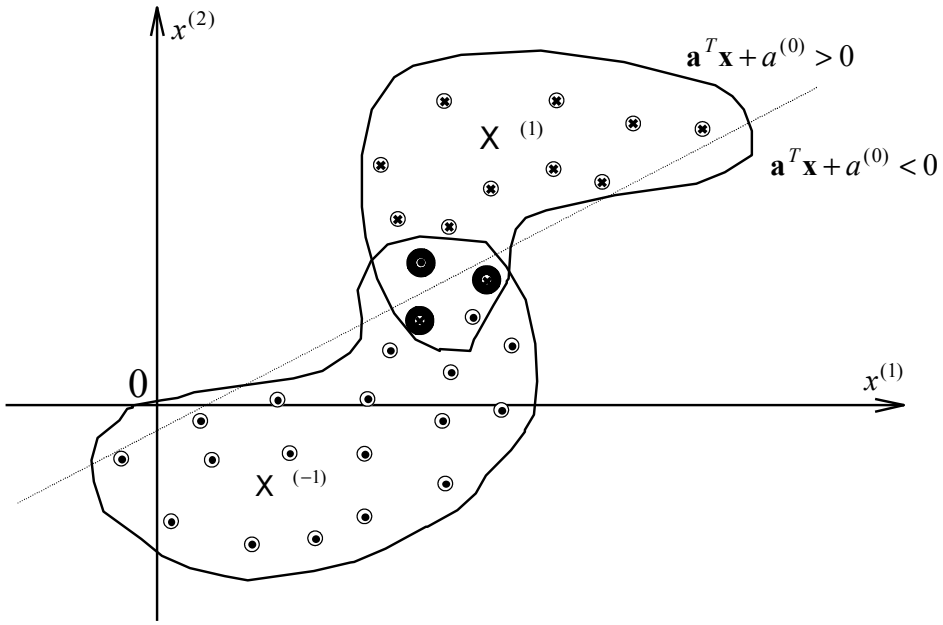


Рисунок. 2.1.1. Гиперплоскость, оптимальная в смысле наилучшего разделения точек первого и второго класса в обучающей выборке; выделены т.н. опорные точки, только на которые фактически и опирается оптимальная гиперплоскость.

Таким образом, неотъемлемой особенностью концепции оптимальной разделяющей гиперплоскости как общей стратегии обучения является то обстоятельство, что обучение опирается только на очень малую часть выборки. Если повторить обучение еще раз, то при малом размере подвыборок первого и второго класса их крайние точки могут “лечь” существенно по-другому, и оптимальная разделяющая гиперплоскость приобретет другой “наклон”. Эта особенность оптимальной разделяющей гиперплоскости наглядно иллюстрируется примером на рис. 2.1.1.

Можно дать и другую интерпретацию эффекту высокой чувствительности оптимальной разделяющей гиперплоскости к конфигурации точек в обучающей выборке. В конечном итоге, мы хотели бы построить гиперплоскость, оптимальную по отношению к истинной форме областей классов $\mathsf{X}^{(1)}$ и $\mathsf{X}^{(-1)}$. Абсолютно вся информация, которая для этого нужна, содержится в функции

$$J^*(\mathbf{a}) = \min_{\mathbf{x}\in\mathsf{X}^{(1)}} \mathbf{a}^T\mathbf{x} - \max_{\mathbf{x}\in\mathsf{X}^{(-1)}} \mathbf{a}^T\mathbf{x},$$

которую естественно назвать функцией линейной разделимости классов. Эта функция нам недоступна, и мы пользуемся ее кусочно-линейной оценкой $J(\mathbf{a})$ (2.1.2), (2.1.3), измеренной только в очень редких точках в пространстве направляющих векторов $\mathbf{a}\in\mathsf{R}^n$, число и расположение которых определяются конфигурацией опорных векторов выборки в пространстве признаков. Максимизацию этой функции мы также проводим только среди этих самых опорных точек, как это схематически иллюстрирует рис. 2.1.2. Как следствие, оцененная точка максимума в

пространстве направляющих векторов, как правило, будет существенно отличаться от искомой “истинной” точки $\mathbf{a}^*$.

В следующем разделе мы рассмотрим способ компенсации высокой чувствительности оптимальной разделяющей гиперплоскости к вариабельности конфигурации опорных точек обучающей выборки, основанный на более общей версии понятий оптимальной разделяющей гиперплоскости и множества опорных точек, позволяющей вовлечь в процесс формирования разделяющей гиперплоскости, значительно большую часть точек выборки.

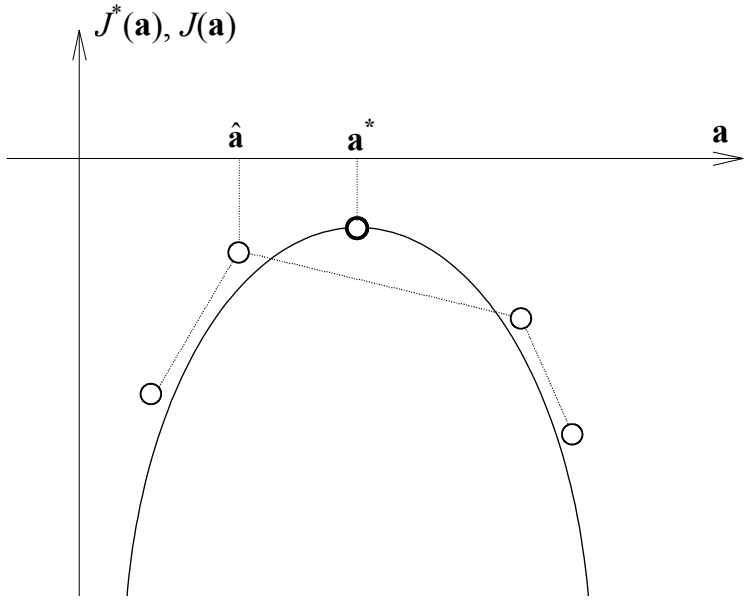


Рисунок. 2.1.2. Схематическое изображение функции линейной разделимости классов $J^*(\mathbf{a})$ и ее кусочно-линейной оценки $J(\mathbf{a})$.

Следует отметить, что алгоритм построения оптимальной разделяющей гиперплоскости был предложен еще Б.Н. Козинцом [3]. Однако, его процедура работает лишь с непересекающимися выпуклыми оболочками двух множеств. Причем итерационный процесс поиска такой гиперплоскости останавливается на основании достаточно эвристического критерия, определяющего расстояние (зазор) между точками, определяющими искомую границу. Хотя надо добавить, что в случае непересекающихся множеств не составляет труда на основании результатов работы процедуры Б.Н. Козинца явно указать опорные векторы.

2.2 Метод опорных векторов и алгоритм обучения распознаванию для двух классов

Пусть обучающая совокупность содержит N объектов двух классов, представленных векторами их действительнозначных признаков $\mathbf{x}_j \in \mathbb{R}^n$ и индексами классов $g_j \in \{1,-1\}$, $j = 1,...,N$.

1.2.3 Первая форма задачи построения оптимальной разделяющей гиперплоскости (общая для разделимых и неразделимых объектов двух классов)

Предположим, что объекты классов 1 и -1 линейно разделимы. Тогда существует гиперплоскость $\mathbf{a}^T\mathbf{x} + b = 0$, такая, что

$$\mathbf{a}^T\mathbf{x}_j + b \geq \xi \text{ при } g_j = 1 \text{ и } \mathbf{a}^T\mathbf{x} + b \leq -\xi \text{ при } g_j = -1,\ j = 1,...,N, \qquad (2.2.1)$$

где $\xi > 0$.

Оптимальной называется такая гиперплоскость, для которой зазор ξ является наибольшим среди всех гиперплоскостей с направляющими векторами единичной нормы:

$$\xi \to \max \text{ при ограничениях (1) и } \mathbf{a}^T\mathbf{a} = 1$$

или, в более конструктивном виде,

$$J(\mathbf{a}) = \min_{j:g_j=1} \mathbf{a}^T\mathbf{x}_j - \max_{j:g_j=-1} \mathbf{a}^T\mathbf{x}_j \to \max \text{ при ограничении } \mathbf{a}^T\mathbf{a} = 1. \qquad (2.2.2)$$

Предположим теперь, что объекты классов 1 и -1 линейно неразделимы. В этом случае для любой гиперплоскости $\mathbf{a}^T\mathbf{x} + b = 0$ существуют точки класса 1, в которых $\mathbf{a}^T\mathbf{x}_j + b < 0$, и точки класса -1, в которых $\mathbf{a}^T\mathbf{x}_j + b > 0$, так что не существует гиперплоскости, удовлетворяющей условиям (2.2.1) с каким бы то ни было положительным зазором ξ. Какую бы гиперплоскость мы ни выбрали, условия (2.2.1) будут выполняться лишь для некоторой величины $\xi < 0$.

Естественно оптимальной называть такую гиперплоскость, которая обеспечивает выполнение таких неравенств с наименьшим по абсолютной величине “обратным” зазором, т.е. с наибольшим значением ξ, только это наибольшее значение уже не сможет стать положительным. Таким образом, задача поиска оптимальной разделяющей гиперплоскости по-прежнему формально выражается условием (.2.2.2).

Задача (2.2.2) представляет собой задачу минимизации кусочно линейной функции $J(\mathbf{a})$ на сфере $\mathbf{a}^T\mathbf{a} = 1$. Хотя сама целевая функция выпукла, область поиска выпуклой не является, поэтому задача неудобна для численного решения.

2.2.2. Вторая форма задачи построения оптимальной разделяющей гиперплоскости (разная для разделимых и неразделимых объектов двух классов)

Пусть объекты классов 1 и -1 линейно разделимы. Очевидно, что изменением масштаба оси $\mathbf{a}$ всегда можно сделать правые части неравенств (2.2.1) равными, соответственно, 1 и -1. Для этого достаточно разделить оба неравенства на ξ

$$\frac{1}{\xi}\mathbf{a}^T\mathbf{x}_j + \frac{1}{\xi}b \geq 1, \quad \frac{1}{\xi}\mathbf{a}^T\mathbf{x}_j + \frac{1}{\xi}b \leq -1$$

и принять $\frac{1}{\xi}\mathbf{a}$ и $\frac{1}{\xi}b$ в качестве новых вектора $\mathbf{a}$ и порога b:

$$\mathbf{a}^T\mathbf{x}_j + b \geq 1 \text{ при } g_j = 1 \text{ и } \mathbf{a}^T\mathbf{x} + b \leq -1 \text{ при } g_j = -1. \qquad (2.2.3)$$

Эти неравенства можно записать в более компактной форме, умножив обе части второго неравенства на -1, поменяв при этом, соответственно, его знак:

$-\mathbf{a}^T\mathbf{x} - b \geq 1$ при $g_j = -1$. Тогда оба неравенства можно заменить одним общим неравенством:

$$g_j(\mathbf{a}^T\mathbf{x}_j + b) \geq 1, \quad j = 1,\ldots,N.$$

После деления на ξ новый вектор $\mathbf{a}$ будет в ξ раз короче прежнего, норма которого была равна 1. Чем больше ξ, т.е. чем лучше была прежняя гиперплоскость, тем меньше будет норма нового вектора $\mathbf{a}$. Таким образом, задача построения оптимальной разделяющей гиперплоскости для линейно разделимых объектов классов 1 и -1 принимает следующий вид:

$$\mathbf{a}^T\mathbf{a} \to \min \quad \text{при ограничениях} \quad g_j(\mathbf{a}^T\mathbf{x}_j + b) \geq 1, \quad j = 1,\ldots,N. \qquad (2.2.5)$$

Это задача минимизации квадратичной функции при линейных ограничениях типа неравенств, т.е. классическая задача квадратичного программирования. Минимальное значение $1/\xi^2 = \mathbf{a}^T\mathbf{a}$ указывает максимальное возможную величину ξ зазора между гиперплоскостью и векторами первого и второго классов, безошибочно разделяемыми этой гиперплоскостью, если вернуться к исходным параметрам $\mathbf{a} = \xi\mathbf{a}$ и $b = \xi b$, удовлетворяющим условию $\mathbf{a}^T\mathbf{a} = 1$.

При неразделимых совокупностях объектов классов 1 и -1 множество, определяемое ограничениями (2.2.4), будет пустым, т.е. задача (2.2.5) не будет иметь решения. В этом случае неравенства (1) могут быть выполнены, как мы уже говорили, только при отрицательном значении ξ. Если по-прежнему считать ξ положительной величиной, то их надо заменить неравенствами

$$\mathbf{a}^T\mathbf{x}_j + b \geq -\xi \text{ при } g_j = 1 \text{ и } \mathbf{a}^T\mathbf{x} + b \leq \xi \text{ при } g_j = -1, \; j = 1,\ldots,N. \qquad (2.2.6)$$

Гиперплоскость тем лучше, чем меньше значение ξ. Разделив оба неравенства на ξ и изменив тем самым масштаб оси $\mathbf{a}$, всегда можно сделать правые части неравенств (6) равными, соответственно, -1 и 1

$$\frac{1}{\xi}\mathbf{a}^T\mathbf{x}_j + \frac{1}{\xi}b \geq -1, \quad \frac{1}{\xi}\mathbf{a}^T\mathbf{x}_j + \frac{1}{\xi}b \leq 1,$$

или, приняв $\frac{1}{\xi}\mathbf{a}$ и $\frac{1}{\xi}b$ в качестве новых вектора $\mathbf{a}$ и порога b,

$\mathbf{a}^T\mathbf{x}_j + b \geq -1$ при $g_j = 1$ и $\mathbf{a}^T\mathbf{x} + b \leq 1$ при $g_j = -1$.

Эти два неравенства эквивалентны одному неравенству

$$g_j(\mathbf{a}^T\mathbf{x}_j + b) \geq -1, \quad j = 1,..., N.$$

Чем меньше значение ξ и, соответственно, больше значение $\mathbf{a}$, тем лучше гиперплоскость, поэтому задача построения оптимальной разделяющей гиперплоскости для линейно неразделимых объектов классов 1 и -1 может быть записана в виде:

$$\mathbf{a}^T\mathbf{a} \to \max \text{ при ограничениях } g_j(\mathbf{a}^T\mathbf{x}_j + b) \geq -1, \quad j = 1,..., N. \qquad (2.2.8)$$

Максимальное значение $1/\xi^2 = \mathbf{a}^T\mathbf{a}$ дает минимальную величину остаточного дефекта ξ, с которым гиперплоскость с исходными параметрами $\mathbf{a} = \xi\mathbf{a}$ и $b = \xi b$, удовлетворяющими условию $\mathbf{a}^T\mathbf{a} = 1$, разделяет векторы первого и второго классов, не разделимые никакой гиперплоскостью без ошибки.

2.2.3. Третья форма задачи построения оптимальной разделяющей гиперплоскости (разная для разделимых и неразделимых объектов двух классов)

Пусть объекты двух классов линейно разделимы. Задача (2.2.5) есть задача минимизации квадратичной функции при линейных ограничениях типа неравенств, т.е. классическая задача квадратичного программирования. Ей соответствует функция Лагранжа

$$L(\mathbf{a}, b, \lambda_1,..., \lambda_N) = \frac{1}{2}\mathbf{a}^T\mathbf{a} - \sum_{j=1}^{N} \lambda_j \left[g_j(\mathbf{a}^T\mathbf{x}_j + b) - 1 \right], \qquad (2.2.9)$$

где для удобства дальнейших выкладок принят коэффициент $1/2$ перед целевой функцией. $\lambda_j \geq 0$, $j = 1,..., N$ – неотрицательные множители Лагранжа. Решением задачи является седловая точка функции Лагранжа:

$$L(\mathbf{a}, b, \lambda_1,..., \lambda_N) \to \min \text{ по } \mathbf{a}, b,$$

$$L(\mathbf{a}, b, \lambda_1,..., \lambda_N) \to \max \text{ по } \lambda_1,..., \lambda_N,$$

при ограничениях

$$\lambda_j \geq 0, \ j = 1,..., N.$$

Первое из этих условий (2.2.10) дает

$$\nabla_{\mathbf{a}} \left\{ \frac{1}{2}\mathbf{a}^T\mathbf{a} - \sum_{j=1}^{N} \lambda_j \left[g_j(\mathbf{a}^T\mathbf{x}_j + b) - 1 \right] \right\} = 0 \text{ и } \frac{\partial}{\partial b} \left\{ \frac{1}{2}\mathbf{a}^T\mathbf{a} - \sum_{j=1}^{N} \lambda_j \left[g_j(\mathbf{a}^T\mathbf{x}_j + b) - 1 \right] \right\} = 0,$$

откуда получим

$$\mathbf{a} = \sum_{j=1}^{N} \lambda_j g_j \mathbf{x}_j,$$

$$\sum_{j=1}^{N}\lambda_j g_j = 0.$$

Подстановка (2.2.13) во второе условие (2.2.11) превращает его в целевую функцию, вообще говоря, относительно множителей Лагранжа $\lambda_1,...,\lambda_N$ и порога b

$$W(b,\lambda_1,...,\lambda_N) = \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}(g_j g_k \mathbf{x}_j^T \mathbf{x}_k)\lambda_j\lambda_k - \sum_{j=1}^{N}\sum_{k=1}^{N}(g_j g_k \mathbf{x}_j^T \mathbf{x}_k)\lambda_j\lambda_k - \left(\sum_{j=1}^{N}\lambda_j g_j\right)b + \sum_{j=1}^{N}\lambda_j,$$

однако в силу равенства (2.2.14) активными аргументами являются только множители Лагранжа. Таким образом, мы приходим к следующей формулировке задачи построения оптимальной разделяющей гиперплоскости в виде задачи квадратичного программирования, называемой в литературе двойственной по Вульфу [R. Fletcher. Practical Methods of Optimizations. John Wiley and Sons Inc., 2nd edition, 1987.] или по Лагранжу [Базара М., Шетти К. Нелинейное программирование. Теория и алгоритмы. М.: Мир, 1982] по отношению к задаче (2.2.5):

$$\begin{aligned} &W(\lambda_1,...,\lambda_N) = \sum_{j=1}^{N}\lambda_j - \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}(g_j g_k \mathbf{x}_j^T \mathbf{x}_k)\lambda_j\lambda_k \to \max, \\ &\sum_{j=1}^{N}\lambda_j g_j = 0, \quad \lambda_j \geq 0,\ j = 1,...,N. \end{aligned} \tag{2.2.15}$$

$W(\lambda_1,...,\lambda_N)$ - двойственная функция Лагранжа - является вогнутой, следовательно, всякий ее локальный максимум является и глобальным [Базара М., Шетти К. Нелинейное программирование. Теория и алгоритмы. М.: Мир, 1982].

После того, как множители Лагранжа найдены, направляющий вектор оптимальной разделяющей гиперплоскости определяется по формуле (2.2.13) как линейная комбинация векторов обучающей совокупности. Те векторы, для которых $\lambda_j \neq 0$, т.е. $\lambda_j > 0$ с учетом ограничения (2.2.12), называются опорными векторами оптимальной разделяющей гиперплоскости.

Для определения значения порога b используем тот факт, что в оптимальной точке из совокупности ограничений $g_j(\mathbf{a}^T\mathbf{x}_j + b) \geq 1$ (2.2.5) активными, т.е. превращающимися в равенства $g_j(\mathbf{a}^T\mathbf{x}_j + b) = 1$, являются те, для которых множители Лагранжа положительны $\lambda_j > 0$. Для этих ограничений имеем

$$(\lambda_j g_j)g_j(\mathbf{a}^T\mathbf{x}_j + b) = \lambda_j(\mathbf{a}^T\mathbf{x}_j + b) = \lambda_j g_j$$

Но эти же равенства выполняются и для всех остальных j в силу того, что для них $\lambda_j = 0$. Сложив все эти равенства, получим

$$\sum_{j=1}^{N}\lambda_j(\mathbf{a}^T\mathbf{x}_j + b) = \sum_{j=1}^{N}\lambda_j g_j = 0, \text{ т.е. } \sum_{j=1}^{N}\lambda_j\mathbf{a}^T\mathbf{x}_j + \left(\sum_{j=1}^{N}\lambda_j\right)b = 0, \text{ откуда следует}$$

$$b = -\frac{\sum_{j=1}^{N} \lambda_j \mathbf{a}^T \mathbf{x}_j}{\sum_{j=1}^{N} \lambda_j}.$$

Заметим, что мы по-прежнему решаем задачу (2.2.5), и значение $1/\xi^2 = \mathbf{a}^T\mathbf{a}$ для направляющего вектора (2.2.12), вычисленного при оптимальных значениях весов, дает максимальную величину ξ остаточного зазора, с которым гиперплоскость разделяет точки первого и второго классов. С учетом (2.2.13) эта величина выражается непосредственно через значения весов

$$1/\xi^2 = \sum_{j=1}^{N}\sum_{k=1}^{N}(g_j g_k \mathbf{x}_j^T \mathbf{x}_k)\lambda_j\lambda_k .$$

К тому же, аналогично (16), $\lambda_j g_j(\mathbf{a}^T\mathbf{x}_j + b) = \lambda_j$, что при суммировании по всем j дает

$$\sum_{j=1}^{N}\lambda_j = \sum_{j=1}^{N}\lambda_j g_j(\mathbf{a}^T\mathbf{x}_j + b) = \sum_{j=1}^{N}\lambda_j g_j \mathbf{a}^T\mathbf{x}_j + \left(\sum_{j=1}^{N}\lambda_j g_j\right)b,$$

откуда с учетом (13) и ограничения в виде равенства в (2.2.15) получим

$$\sum_{j=1}^{N}\lambda_j = \sum_{j=1}^{N}\lambda_j g_j\left(\sum_{k=1}^{N}\lambda_k g_k \mathbf{x}_k\right)^T \mathbf{x}_j = \sum_{j=1}^{N}\sum_{k=1}^{N}(g_j g_k \mathbf{x}_j^T \mathbf{x}_k)\lambda_j\lambda_k = 1/\xi^2 .$$

Таким образом, $\max_{\lambda_1,\ldots,\lambda_N} W(\lambda_1,\ldots,\lambda_N) = 1/2\xi^2$, и при любых других значениях весов выполняется неравенство $W(\lambda_1,\ldots,\lambda_N) \le 1/2\xi^2$.

Пусть $\varepsilon > 0$ – некоторое достаточно малое число, такое, что множества точек первого и второго класса можно считать линейно неразделимыми, если никакая гиперплоскость не может обеспечить величину остаточного зазора, большую ε. Если в процессе решения задачи квадратичного программирования (2.2.15) наступит ситуация $W(\lambda_1,\ldots,\lambda_N) > 1/2\varepsilon^2$, то точки первого и второго класса линейно неразделимы, и процесс должен быть остановлен.

Рассмотрим теперь третью постановку задачи для случая линейной неразделимости объектов классов 1 и –1. Следует отметить тот факт, что для задачи (8) система ограничений образует замкнутую область в пространстве варьируемых параметров, а критерий представляет собой ***максимизацию выпуклой*** квадратичной целевой функции. Это приводит к наличию нескольких локальных экстремумов, и как следствие - трудности в построении эффективной процедуры направляющего вектора оптимальной разделяющей гиперплоскости. Этот факт наглядно демонстрируется на примере, приведенном на рис. 2.2.1.

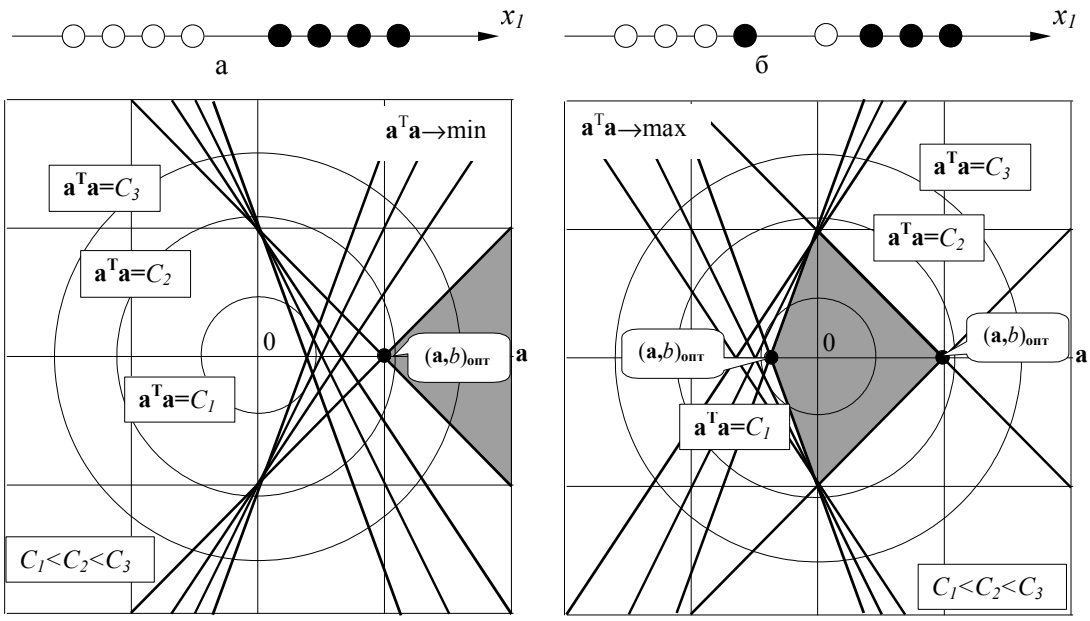
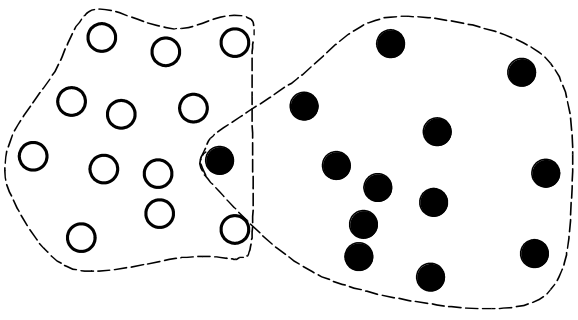


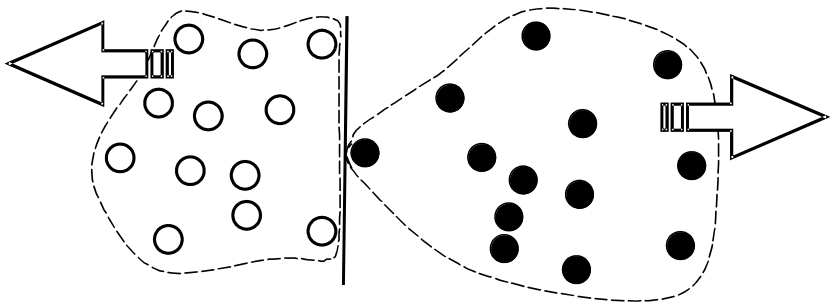
Рисунок. 2.2.1 Максимизация вогнутой (а) и выпуклой (б) функции при системе ограничений.

Это обстоятельство заставило Вапника и Кортес отказаться от записи задачи (2.2.8) в виде, эквивалентном задаче (2.2.9). В работе [C.Cortes and V.Vapnik, 1995] они предложили отличную схему, суть которой заключается в следующем.

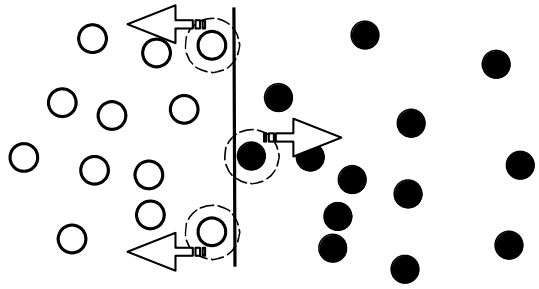
Вернемся ко второй форме задачи построения оптимальной разделяющей гиперплоскости. Как очевидно то, что для случая линейно разделимых классов изменением масштаба оси $\mathbf{a}$, всегда можно сделать правые части неравенств (2.2.1) равными соответственно 1 и -1, очевидно и то, что смещением объектов 1-го класса в положительном направлении вектора $\mathbf{a}$ и -1-го класса в отрицательном направлении удастся линейно неразделимые классы представить как разделимые. Предлагается применить такое "центробежное" преобразование по-разному для различных объектов выборки, а именно, для объектов, попавших в область чужого класса, необходимо ощутимо отодвинуть их в "свою" сторону, в то время как объекты из задних областей вообще не нуждаются в подобном смещении. Понятно преимущество такого подхода (рис.2.2.2 *в*) по сравнению с добавлением одинаковой константы ко всем объектам выборки (рис.2.2.2 *б*).



а



б



в

Рисунок 2.2.2 Смещение объектов для случая линейно неразделимых классов.

Он позволяет оценить число ошибок в обучающей выборке $R = \frac{n_1 + n_{-1}}{N_1 + N_{-1}}$, где n_1 и n_{-1} - минимальное число объектов в классе 1 и -1, соответственно, сместив которые, удастся добиться линейной разделимости классов.

Таким образом, для неразделимого случая перепишем неравенства (2.2.3) в виде

$$\mathbf{a}^T\mathbf{x}_j + b \geq +1 - \delta_j \text{ при } g_j = 1 \text{ и } \mathbf{a}^T\mathbf{x} + b \leq -1 + \delta_j \text{ при } g_j = -1,$$

где $\delta_j \geq 0, \quad j = 1,...,N$ - неотрицательные константы, на которые необходимо сместить объекты обучающей выборки, чтобы добиться линейной разделимости. В более компактной форме это может быть записано следующим образом

$$g_j(\mathbf{a}^T\mathbf{x}_j + b) \geq +1 - \delta_j .$$

Тогда ошибка распознавания на обучающей выборке может быть представлена в следующем виде

$$R = \frac{\sum_{j=1}^{N} \theta(\delta_j)}{N}, \text{ где } \theta(\delta) = \begin{cases} 1, \textit{если } \delta > 0 \\ 0, \textit{если } \delta = 0 \end{cases}.$$

Задача заключается в выборе $\delta_j, \quad j = 1,...,N$ таким образом, чтобы $\sum_{j=1}^{N} \delta_j \to \min$. В таком случае общий критерий поиска можно представить в одном из следующих видов:

$$\frac{1}{2}(\mathbf{a}^T\mathbf{a} + C\sum_{j=1}^{N} \delta_j) \to \min,$$

$$\frac{1}{2}[\mathbf{a}^T\mathbf{a} + C\sum_{j=1}^{N} (\delta_j)^2] \to \min$$

$$\frac{1}{2}[\mathbf{a}^T\mathbf{a} + C(\sum_{j=1}^{N}\delta_j)^k] \to \min, \quad k > 1,$$

где C - положительная константа - параметр пользователя, при ограничениях

$$g_j(\mathbf{a}^T\mathbf{x}_j + b) \geq +1 - \delta_j,$$
$$\delta_j \geq 0,\ j = 1,..., N.$$

Таким образом, получена задача минимизации квадратичной функции при линейных ограничениях типа неравенств. Осталось получить двойственную по отношению к ней.

Функция Лагранжа, соответствующая такой задаче имеет вид

$$L(\mathbf{a}, b, \lambda_1,..., \lambda_N, \delta_1,..., \delta_N, \mu_1,..., \mu_N) = \frac{1}{2}\mathbf{a}^T\mathbf{a} + \frac{C}{2}\sum_{j=1}^{N}\delta_j - \\ - \sum_{j=1}^{N}\lambda_j\left[g_j(\mathbf{a}^T\mathbf{x}_j + b) - 1 + \delta_j\right] - \sum_{j=1}^{N}\mu_j\delta_j, \tag{2.2.20}$$

где $\lambda_j \geq 0$, $\mu_j \geq 0$, $j = 1,..., N$ - неотрицательные множители Лагранжа.

Решением задачи является седловая точка функции Лагранжа:

$$L(\mathbf{a}, b, \lambda_1,..., \lambda_N, \delta_1,..., \delta_N, \mu_1,..., \mu_N) \to \min \text{ по } \mathbf{a}, b, \delta_1,..., \delta_N, \tag{2.2.21}$$

$$L(\mathbf{a}, b, \lambda_1,..., \lambda_N, \delta_1,..., \delta_N, \mu_1,..., \mu_N) \to \max \text{ по } \lambda_1,...\lambda_N, \mu_1,..., \mu_N, \tag{2.2.22}$$

при ограничениях

$$\lambda_j \geq 0, \mu_j \geq 0, \quad j = 1,..., N.$$

Первое из этих условий дает

$$\nabla_{\mathbf{a}} L(\mathbf{a}, b, \lambda_1,..., \lambda_N, \delta_1,..., \delta_N, \mu_1,..., \mu_N) = 0,$$
$$\frac{\partial L(\mathbf{a}, b, \lambda_1,..., \lambda_N, \delta_1,..., \delta_N, \mu_1,..., \mu_N)}{\partial b} = 0,$$
$$\frac{\partial L(\mathbf{a}, b, \lambda_1,..., \lambda_N, \delta_1,..., \delta_N, \mu_1,..., \mu_N)}{\partial \delta_j} = 0, \quad j = 1,..., N.$$

откуда получим

$$\mathbf{a} = \sum_{j=1}^{N}\lambda_j g_j \mathbf{x}_j, \quad j = 1,..., N,$$

$$\sum_{j=1}^{N}\lambda_j g_j = 0, \quad j = 1,..., N,$$

$$\lambda_j + \mu_j = \frac{C}{2}, \quad j = 1,..., N.$$

Отметим что, так как $\lambda_j \geq 0, \mu_j \geq 0$ и $\lambda_j + \mu_j = \frac{C}{2}$, то $0 \leq \lambda_j \leq \frac{C}{2}$ и $0 \leq \mu_j \leq \frac{C}{2}$.

Подстановка (2.2.23) в условие (2.2.22) превращает его в целевую функцию

$$W(b,\lambda_1,...,\lambda_N,\delta_1,...,\delta_N,\mu_1,...,\mu_N)=\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}(g_j g_k \mathbf{x}_j^T\mathbf{x}_k)\lambda_j\lambda_k+\frac{C}{2}\sum_{j=1}^{N}\delta_j-$$

$$-\sum_{j=1}^{N}\sum_{k=1}^{N}(g_j g_k \mathbf{x}_j^T\mathbf{x}_k)\lambda_j\lambda_k-\left(\sum_{j=1}^{N}\lambda_j g_j\right)b+\sum_{j=1}^{N}\lambda_j-\sum_{j=1}^{N}\lambda_j\delta_j-\sum_{j=1}^{N}\mu_j\delta_j$$

однако в силу равенств (2.2.24) и (2.2.25) активными аргументами являются только множители Лагранжа $\lambda_j \ge 0, \quad j=1,...,N$.

Таким образом, мы приходим к следующей формулировке задачи построения оптимальной разделяющей гиперплоскости в виде задачи квадратичного программирования

$$W(\lambda_1,...,\lambda_N)=\sum_{j=1}^{N}\lambda_j-\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}(g_j g_k \mathbf{x}_j^T\mathbf{x}_k)\lambda_j\lambda_k\to\max,$$
$$\sum_{j=1}^{N}\lambda_j g_j=0,\quad 0\le\lambda_j\le\frac{1}{2}C,\ j=1,...,N. \qquad (2.2.26)$$

По-прежнему значения $\lambda_j>0$ указывают опорные элементы обучающей выборки, определяющие параметры оптимальной разделяющей гиперплоскости, вычисляемые в данном случае по формуле (2.2.26). Максимальное значение критерия (26) $\max\limits_{\lambda_1,...,\lambda_N} W(\lambda_1,...,\lambda_N)=-1/2\xi^2$ укажет минимальную величину ξ дефицита, с которым точки первого и второго классов могут быть разделены гиперплоскостью, а направляющий вектор разделяющей гиперплоскость определяется как $\mathbf{a}=\sum\limits_{j+1}^{N}\lambda_j g_j \mathbf{x}_j$, то есть так же как и для случая линейно разделимых выборок. Константа же b в этом случае будет другой. Определим ее.

Смещения δ_j равны нулю тогда т только тогда, кода соответствующие им суть множителей Лагранжа $\mu_j>0$, то есть, когда $\lambda_j<\frac{1}{2}C$, $j=1,...,N.$ В других терминах: $\delta_j>0$ тогда и только тогда, когда $\lambda_j=\frac{1}{2}C$.

Для положительных λ_j выполняется условие $g_j(\mathbf{a}^T\mathbf{x}_j+b)=1-\delta_j$. Если $0<\lambda_j<\frac{1}{2}C$, то $\delta_j=0$, и, следовательно, $g_j(\mathbf{a}^T\mathbf{x}_j+b)=1$. Учитывая, тот факт что $g_j^2=1$, перепишем последнее уравнение в виде $\lambda_j(\mathbf{a}^T\mathbf{x}_j+b)=\lambda_j g_j$, откуда следует $\sum\limits_{j:0<\lambda_j<\frac{1}{2}}\lambda_j(\mathbf{a}^T\mathbf{x}_j+b)=\sum\limits_{j:0<\lambda_j<\frac{1}{2}}\lambda_j g_j$. Выразив отсюда b получим:

$$b = -\frac{\sum\limits_{j:\,0\le\lambda_j<\frac{1}{2}C}\lambda_j \mathbf{a}^T\mathbf{x}_j + \frac{1}{2}C\sum\limits_{j:\,\lambda_j=\frac{1}{2}C} g_j}{\sum\limits_{j:\,0\le\lambda_j<\frac{1}{2}C}\lambda_j}.$$

Для задачи (2.2.19b) аналогичным образом можно получить соответствующую ей двойственную задачу вида

$$\begin{aligned} &W(\lambda_1,...,\lambda_N) = \sum_{j=1}^{N}\lambda_j - \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}(g_j g_k \mathbf{x}_j^T\mathbf{x}_k)\lambda_j\lambda_k - \frac{1}{2C}\sum_{j=1}^{N}\lambda_j^2 \to \max, \\ &\sum_{j=1}^{N}\lambda_j g_j = 0, \quad \lambda_j > 0,\ j = 1,...,N. \end{aligned} \tag{2.2.27a}$$

Или введя новые обозначения $\mathbf{D} = \{(g_g g_g \mathbf{x}_j \mathbf{x}_k)\}$, $\mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ и $\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_N \end{pmatrix}$ в виде

$$\begin{aligned} &W(\mathbf{\Lambda}) = \mathbf{1\Lambda} - \frac{1}{2}\mathbf{\Lambda}^T(\mathbf{D} + \frac{1}{C}\mathbf{I})\mathbf{\Lambda} \to \max, \\ &\sum_{j=1}^{N}\lambda_j g_j = 0, \quad \lambda_j > 0,\ j = 1,...,N. \end{aligned}$$

Направляющий вектор разделяющей гиперплоскости будет определяться следующим образом.

$$\mathbf{a} = \sum_{j+1}^{N}\lambda_j g_j \mathbf{x}_j\ , b = -\frac{\sum\limits_{j=1}^{N}\lambda_j \mathbf{a}^T\mathbf{x}_j + \frac{1}{C}\sum\limits_{j=1}^{N}\lambda_j^2 g_j}{\sum\limits_{j=1}^{N}\lambda_j}.$$

Для задачи выпуклого программирования (19c) получим двойственную задачу

$$\begin{aligned} &W(\lambda_1,...,\lambda_N,d) = \sum_{j=1}^{N}\lambda_j - \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}(g_j g_k \mathbf{x}_j^T\mathbf{x}_k)\lambda_j\lambda_k - \frac{d^{k/k-1}}{(kC)^{k/k-1}}\left(1 - \frac{1}{k}\right) \to \max, \\ &\sum_{j=1}^{N}\lambda_j g_j = 0, \quad 0 \le \lambda_j \le d,\ j = 1,...,N, \quad k > 1. \end{aligned} \tag{2.2.28}$$

где произведена замена $\sum\limits_{j=1}^{N}\delta_j = \left(\frac{2d}{Ck}\right)^{\frac{1}{k-1}}$. В частности при $k = 2$

$$\begin{aligned} &W(\lambda_1,...,\lambda_N,d) = \sum_{j=1}^{N}\lambda_j - \frac{1}{2}\left[\sum_{j=1}^{N}\sum_{k=1}^{N}(g_j g_k \mathbf{x}_j^T\mathbf{x}_k)\lambda_j\lambda_k + \frac{d^2}{C}\right] \to \max, \\ &\sum_{j=1}^{N}\lambda_j g_j = 0, \quad 0 \le \lambda_j \le d,\ j = 1,...,N. \end{aligned} \tag{2.2.28a}$$

Направляющий вектор разделяющей гиперплоскости, определяемый таким набором опорных векторов, будет иметь вид.

$$\mathbf{a} = \sum_{j+1}^{N} \lambda_j g_j \mathbf{x}_j \,, b = -\frac{\sum\limits_{j:\,0\le\lambda_j\le d} \lambda_j a^T x_j + d \sum\limits_{j:\,\lambda_j = d} g_j}{\sum\limits_{j:\,0\le\lambda_j < d} \lambda_j} \,.$$

2.3 Метод опорных векторов и алгоритм обучения распознаванию для случая многих классов

Пусть обучающая совокупность содержит N объектов k классов, представленных векторами их действительнозначных признаков $\mathbf{x}_j \in \mathbf{R}^n$ и индексами классов $g_j \in \{1,..,k\}$, $j = 1,...,N$.

Для случая двух классов решение задачи распознавания образов хорошо известно []. Главная идея обучения состоит в построении разделяющей гиперплоскости таким образом, чтобы максимизировать расстояние между гиперплоскостью и крайними точками обучающих подвыборок. Это дает следующую задачу оптимизации:

$$\phi(\mathbf{a},\delta) = \frac{1}{2}\left(\mathbf{a}^T\mathbf{a} + C\sum_{j=1}^{N} \delta_j \right) \to \min \,, \qquad (2.4.1)$$

при ограничениях

$$\begin{gathered} g_j(\mathbf{a}^T\mathbf{x}_j + b) \ge +1 - \delta_j \,, \\ \delta_j \ge 0,\ j = 1,...,N, \\ g_j \in \{-1,1\} \,. \end{gathered} \qquad (2.4.2)$$

Двойственная к ней задача, имеющая вид

$$\begin{gathered} W(\lambda_1,...,\lambda_N) = \sum_{j=1}^{N} \lambda_j - \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N} (g_j g_k \mathbf{x}_j^T \mathbf{x}_k)\lambda_j\lambda_k \to \max, \\ \sum_{j=1}^{N} \lambda_j g_j = 0, \quad 0 \le \lambda_j \le \frac{1}{2}C,\ j = 1,...,N. \end{gathered} \qquad (2.4.3)$$

дает следующее решающее правило

$$f(x) = sign\left[\sum_{j=1}^{N} \lambda_j g_j \mathbf{x}_j + b \right].$$

Обучение распознаванию образов в случае многих классов может производиться несколькими способами. Во-первых, для каждой пары классов может строиться своя разделяющая гиперплоскость. Т.о. общее количество гиперплоскостей будет $\frac{k(k-1)}{2}$. Во-вторых, можно построить k гиперплоскостей, каждая из которых отделяет класс m от $k-1$ других классов. В обоих этих случаях общий алгоритм обучения строится как совокупность задач обучения для двух классов.

Более естественным кажется решать задачу обучения для многих классов, рассматривая все k классов одновременно. Отвлечемся от понятия оптимальной разделяющей гиперплоскости. Для каждого класса будем строить свою гиперплоскость, минимизирующую функционал, аналогичный (2.4.1):

$$\phi(\mathbf{a}\ ,\delta) = \frac{1}{2}\left(\sum_{m=1}^{k} \mathbf{a}_m{}^T \mathbf{a}_m + C \sum_{j=1}^{N} \sum_{m \neq g_j} \delta_j^m \right) \qquad (2.4.4)$$

при ограничениях

$$\mathbf{a}_{g_j}{}^T \mathbf{x}_j + b_{g_j} \geq \mathbf{a}_m{}^T \mathbf{x}_j + b_m + 2 - \delta_j^m, \qquad (2.\ 4.5)$$

$$\delta_j^m \geq 0,\ j = 1,...,N,\ m \in \{1,...,k\} \setminus g_j$$

Расстояния до этой гиперплоскости от всех точек родного класса больше, чем от всех остальных точек обучающей совоккупности. Еще раз отметим, что это есть не разделяющая, а неким хитрым образом организованная гиперплоскость, хотя для случая двух классов она совпадает с оптимальной разделяющей гиперплоскостью.

Эта задача дает решающее правило

$$f(\mathbf{x}) = \arg\max_m \left[\mathbf{a}_m{}^T \mathbf{x} + b_m \right],\ m = 1,..,k \qquad (2.4.6)$$

Таким образом, получена задача минимизации квадратичной функции при ограничениях типа неравенств. Осталось получить двойственную к ней.

Функция Лагранжа, соответствующая этой задаче имеет вид

$$L(\mathbf{a}, b, \delta, \lambda, \mu) = \frac{1}{2} \sum_{m=1}^{k} \mathbf{a}_m{}^T \mathbf{a}_m + \frac{C}{2} \sum_{j=1}^{N} \sum_{m=1}^{k} \delta_j^m -$$
$$- \sum_{j=1}^{N} \sum_{m=1}^{k} \lambda_j^m \left[\left(\mathbf{a}_{g_j} - \mathbf{a}_m \right)^T \mathbf{x}_j + b_{g_j} - b_m - 2 + \delta_j^m \right] - \sum_{j=1}^{N} \sum_{m=1}^{k} \mu_j^m \delta_j^m \qquad (2.4.7)$$

с фиктивными переменными

$$\lambda_j^{g_j} = 0,\ \delta_j^{g_j} = 2, \mu_j^{g_j} = 0,\ \ j = 1,..., N$$

где $\lambda_j^m \geq 0,\ \mu_j^m \geq 0,\ \ j = 1,..., N$, $m = 1,..,k$ - неотрицательные множители Лагранжа.

Решением задачи является седловая точка функции Лагранжа:

$$L(\mathbf{a}, b, \delta, \lambda, \mu) \to \min \text{ по } \mathbf{a}, b, \delta$$
$$L(\mathbf{a}, b, \delta, \lambda, \mu) \to \max \text{ по } \lambda, \mu$$

Введем следующее обозначения

$$c_j^n = \begin{cases} 1, если\ g_j = n \\ 0, если\ g_j \neq n \end{cases}$$

и

$$\Lambda_j = \sum_{m=1}^{k} \lambda_j^m$$

Дифференцируя (7) по $\mathbf{a_n}$, b_n и δ_j^n, получим

$$\frac{\partial L}{\partial \mathbf{a}_n} = \mathbf{a}_n + \sum_{j=1}^{N} \lambda_j^n \mathbf{x}_j - \sum_{j=1}^{N} \Lambda_j c_j^n \mathbf{x}_j$$

$$\frac{\partial L}{\partial b_n} = \sum_{j=1}^{N} \lambda_j^n - \sum_{j=1}^{N} \Lambda_j c_j^n$$

$$\frac{\partial L}{\partial \delta_j^n} = -\lambda_j^n + \frac{C}{2} - \mu_j^n$$

Седловая точка функции Лагранжа удовлетворяет следующим условиям

$$\frac{\partial L}{\partial \mathbf{a}_n} = 0 \Rightarrow \mathbf{a}_n = \sum_{j=1}^{N} (\lambda_j^n - \Lambda_j c_j^n)\mathbf{x}_j \qquad (2.4.8)$$

$$\frac{\partial L}{\partial b_n} = 0 \Rightarrow \sum_{j=1}^{N} \lambda_j^n = \sum_{j=1}^{N} \Lambda_j c_j^n \qquad (2.4.9)$$

$$\frac{\partial L}{\partial \delta_j^n} = 0 \Rightarrow \lambda_j^n + \mu_j^n = \frac{C}{2} \text{ или } 0 \le \lambda_j^n \le \frac{C}{2}. \qquad (2.4.10)$$

Подставляя (2.4.8) в (2.4.7), получим

$$W(b, \delta, \lambda, \mu) = \frac{1}{2} \sum_{m=1}^{k} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(c_i^m \Lambda_i - \lambda_i^m\right)\left(c_j^m \Lambda_j - \lambda_j^m\right) \mathbf{x}_i{}^T \mathbf{x}_j -$$

$$- \sum_{m=1}^{k} \sum_{i=1}^{N} \lambda_i^m \left[\sum_{j=1}^{N} \left(c_j^{g_i} \Lambda_j - \lambda_j^{y_i}\right) \mathbf{x}_i{}^T \mathbf{x}_j - \sum_{j=1}^{N} \left(c_j^m \Lambda_j - \lambda_j^m\right) \mathbf{x}_i{}^T \mathbf{x}_j + b_{y_i} - b_m - 2 \right] - \qquad (2.4.11)$$

$$- \sum_{m=1}^{k} \sum_{j=1}^{N} \lambda_j^m \delta_j^m + \frac{C}{2} \sum_{j=1}^{N} \sum_{m=1}^{k} \delta_j^m - \sum_{j=1}^{N} \sum_{m=1}^{k} \mu_j{}^m \delta_j^m$$

Добавив ограничение (2.4.10), исключим в (11) δ.

Заметим, что

$$B_1 = \sum_{i,m} \lambda_i^m b_{y_i} = \sum_m b_m \left(\sum_i c_i^m \Lambda_i \right),$$

$$B_2 = \sum_{i,m} \lambda_i^m b_m = \sum_m b_m \left(\sum_i \lambda_i^m \right),$$

но, принимая во внимание (2.4.9), $B_1 = B_2$, что дает

$$W(\lambda) = 2 \sum_{i,m} \lambda_i^m + \sum_{i,j,m} \left(\frac{1}{2} c_i^m c_j^m \Lambda_i \Lambda_j - \frac{1}{2} c_i^m \Lambda_i \lambda_j^m - \frac{1}{2} c_j^m \Lambda_i \lambda_i^m + \frac{1}{2} \lambda_j^m \lambda_i^m - c_j^{g_i} \Lambda_j \lambda_i^m + \right.$$

$$\left. + \lambda_j^m \lambda_i^{g_i} + c_j^m \Lambda_j \lambda_i^m - \lambda_j^m \lambda_i^m \right) \cdot \mathbf{x}_i{}^T \mathbf{x}_j$$

но $\sum_{i,j,m} c_i^m \Lambda_i \lambda_j^m = \sum_{i,j,m} c_j^m \Lambda_j \lambda_i^m$, поэтому

$$W(\lambda) = 2\sum_{i,m} \lambda_i^m + \sum_{i,j,m} (\frac{1}{2} c_i^m c_j^m \Lambda_i \Lambda_j - c_j^{g_i} \Lambda_i \lambda_i^m - \frac{1}{2} \lambda_j^m \lambda_i^m + + \lambda_j^m \lambda_i^{g_i}) \cdot \mathbf{x}_i^T \mathbf{x}_j .$$

Т.к. $\sum_m c_i^m c_j^m = c_i^{g_i} = c_j^{g_i}$, окончательно получим двойственную задачу

$$W(\lambda) = 2\sum_{i,m} \lambda_i^m + \sum_{i,j,m} (-\frac{1}{2} c_i^{g_i} \Lambda_i \Lambda_j - \frac{1}{2} \lambda_j^m \lambda_i^m + + \lambda_j^m \lambda_i^{g_i}) \cdot \mathbf{x}_i^T \mathbf{x}_j \to \max , \qquad (2.4.11a)$$

которая является квадратичной функцией в терминах множителей Лагранжа с линейными ограничениями

$$\sum_{j=1}^{N} \lambda_j^n = \sum_{j=1}^{N} c_j^n \Lambda_j , \ n = 1,\ldots,k$$

и

$$0 \le \lambda_j^m \le \frac{C}{2}, \ \lambda_j^{g_j} = 0 , \quad j = 1,\ldots,N, m \in \{1,\ldots,k\} \setminus g_j$$

Решив эту задачу, получим решающее правило

$$f(\mathbf{x}, \lambda) = \arg\max_m \left[\sum_{j=1}^{N} \left(c_j^m \Lambda_j - \lambda_i^m \right) \mathbf{x}_j^T \mathbf{x} + b_m \right].$$

По-прежнему значения $\lambda_j > 0$ указывают опорные элементы обучающей выборки, определяющие параметры оптимальной разделяющей гиперплоскости, вычисляемые в данном случае по формуле