3. Учет априорных предпочтений о классе решающих правил.

3.1. Обучение распознаванию сигналовс учетом критерия гладкости решающего правила

Гиперплоскость, оптимальная в смысле наилучшего разделения точек первого и второго класса в обучающей выборке по критерию В.Н. Вапника показана на рис. 3.1.1; выделены т.н. опорные точки, только на которые фактически и опирается оптимальная гиперплоскость

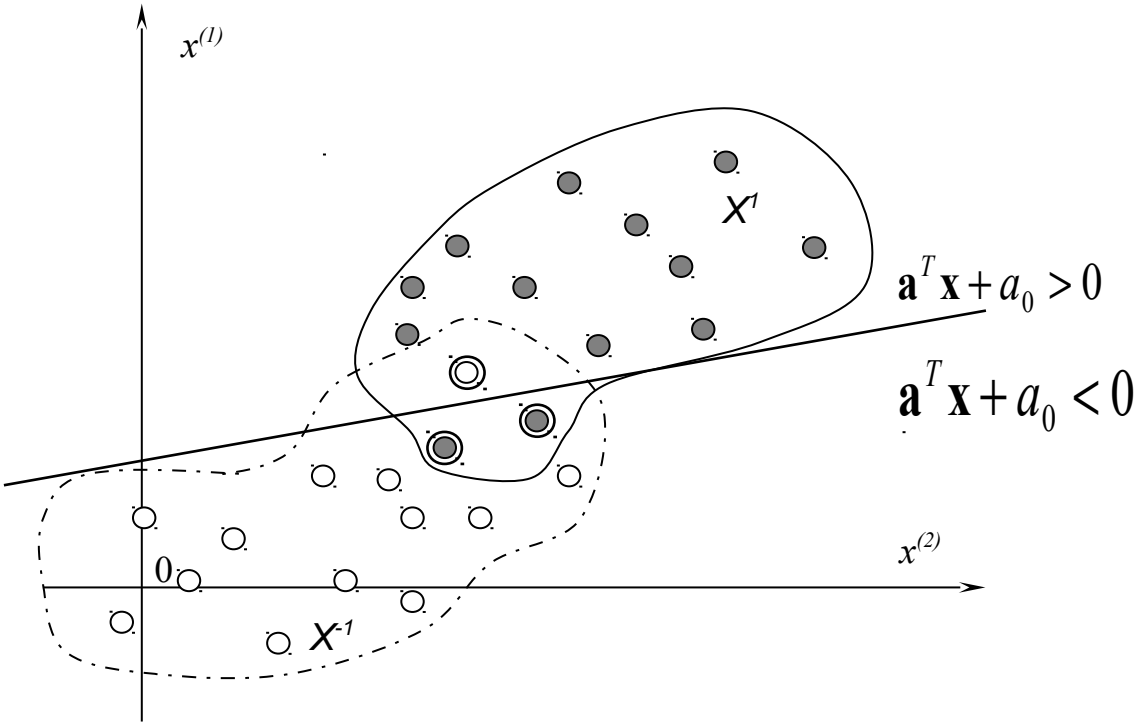


Рисунок 3.1.1 Гиперплоскость, оптимальная в смысле наилучшего разделения точек первого и второго класса в обучающей выборке по критерию В.Н. Вапника.

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

размерах выборки N обучающая совокупность, показывая, где в основном сосредоточены объекты разных классов, будет содержать весьма скудную информацию о форме границ областей. Поэтому доверять этой информации надо с осторожностью.

Кроме того, если число объектов N в обучающей совокупности недостаточно велико и сравнимо с размерностью пространства признаков n, то при любой комбинации объектов и их классов удается найти решающее правило, которое правильно классифицирует объекты данной выборки. Но в результате может оказаться, что оптимальное для данной выборки решающее правило будет очень плохо узнавать классы объектов в других выборках. Иначе говоря, при малых размерах обучающей выборки свобода выбора параметров решающего правила, в данном случае, оказывается слишком большой, и для устойчивости обучения ее надо как-то ограничить [17,18].

Предлагается новый подход к регуляризации решающего правила распознавания многомерных объектов, когда совокупность описывающих их признаков представляет собой результат упорядоченных вдоль оси некоторого аргумента измерений одной и той же характеристики. Типичным примером таких объектов являются некоторые виды сигналов, для которых естественно предположение о невозможности произвольно резких скачков значений соседних отсчетов. Требование гладкости, сужая область допустимых значений решающего правила, позволяет улучшить качество распознавания на генеральной совокупности, пусть даже ценой некоторого снижения его качества на обучающей выборке. Эффективность требования гладкости решающего правила для малонаполненных выборок подтверждена экспериментами .

Предположим, что отдельные признаки $(x_i,\ i=1,...,n)$ в составе вектора $\mathbf{x}$ представляют собой результат упорядоченного измерения некоторого свойства объекта вдоль координаты той или иной природы, причем есть основания полагать, что соседние признаки несут почти идентичную информацию о принадлежности объекта к определенному классу. Такое предположение эквивалентно принятию тезиса о существовании априорной информации о значениях коэффициентов $(a_i,\ i=1,...,n)$ в составе вектора параметров $\mathbf{a}$, заключающееся в том, что соседние коэффициенты, скорее всего, не слишком сильно отличаются друг от друга, т.е. плавно изменяются при увеличении индекса i .

Для того, чтобы в процессе обучения предпочтение отдавалось решающим правилам с плавным изменением коэффициентов линейной части, можно, например, внести в критерий дополнительную аддитивную составляющую

$$J'(\mathbf{a}) = \sum_{i=2}^{n} (a_i - a_{i-1})^2 . \qquad (2.4.1.)$$

Нетрудно убедиться, что такая квадратичная функция может быть записана в виде

$$J'(\mathbf{a}) = \mathbf{a}^T \widetilde{\mathbf{B}} \mathbf{a},$$

$$\widetilde{\mathbf{B}}(n \times n) = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 & 0 & 0 \\ -1 & 2 & -1 & \cdots & 0 & 0 & 0 \\ 0 & -1 & 2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & -1 & 0 \\ 0 & 0 & 0 & \cdots & -1 & 2 & -1 \\ 0 & 0 & 0 & \cdots & 0 & -1 & 1 \end{pmatrix}. \qquad (2.4.2)$$

Тогда целевая функция будет включать в себя еще одно слагаемое

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

или более компактно

$$\frac{1}{2}\mathbf{a}^T \mathbf{B} \mathbf{a} + C \sum_{j=1}^{N} \delta_j \to \min, \text{ где } \mathbf{B} = \mathbf{I} + \alpha \widetilde{\mathbf{B}}, \qquad (2.4.3)$$

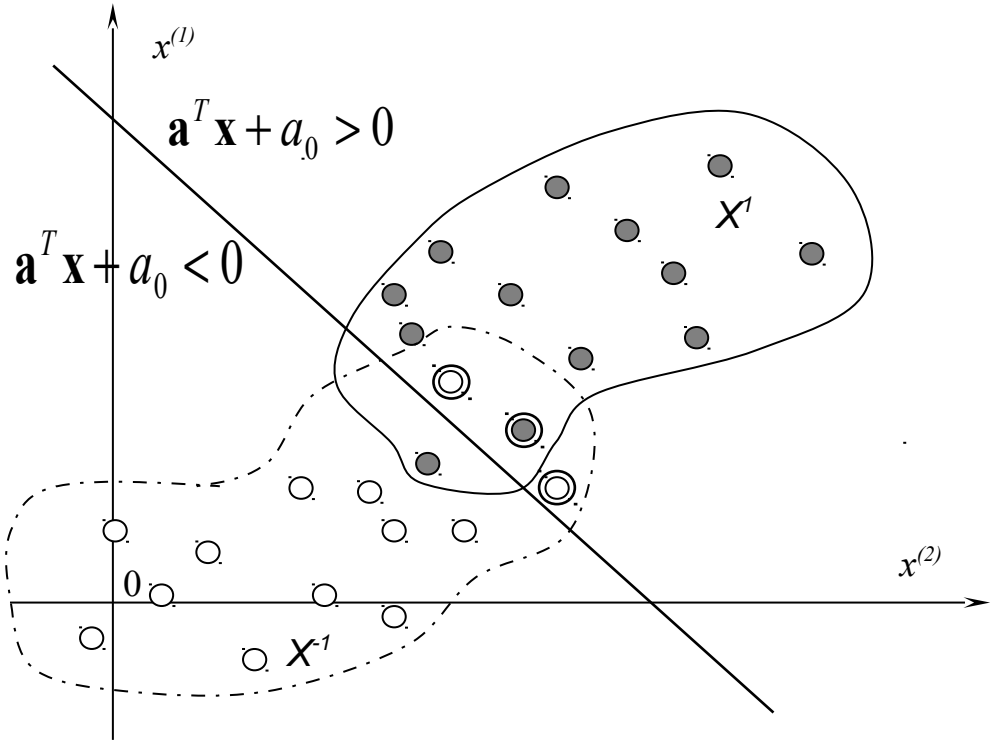
при ограничениях

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

Здесь коэффициент $\alpha \geq 0$ определяет степень, с которой штраф на негладкость последовательности коэффициентов участвует в процессе обучения. Очевидно, что матрица $\mathbf{B}$ положительно определена [30].

Гиперплоскость, оптимальная в смысле наилучшего разделения точек первого и второго класса в обучающей выборке по критерию В.Н. Вапника с учетом критерия гладкости решающего правила, показана на рисунок. 3.1.1.

Рисунок 3.1.2. Гиперплоскость, оптимальная в смысле наилучшего разделения точек первого и



второго класса в обучающей выборке по критерию В.Н. Вапника с учетом критерия гладкости решающего правила

3.2.Оптимальное линейное решающее правило в метрическом пространстве.

Метод опорных векторов, как следует из его названия, разработан для линейного пространства векторов признаков. не смотря на это существует широкий класс приложений, в которых трудно или невозможно выбрать фиксированное множество признаковых переменных, формирующих линейное пространство, в котором задача распознавания образов может быть решена как задача нахождения разделяющей гиперплоскости. Часто случается, что степень различия может быть измерена только для двух объектов, или, другими словами, может быть сформировано только метрическое пространство признаков. Эта ситуация и обсуждается в этой главе.

Пусть Ω, как и ранее, представляет собой гипотетическое множество объектов распознавания, элемент которого будем обозначать $\omega \in \Omega$. Определим на множестве Ω неотрицательную функцию $r(\omega', \omega'') \geq 0$, которая обладает свойствами метрики, т.е. $r(\omega, \omega) = 0$, $r(\omega', \omega'') = r(\omega'', \omega')$, $r(\omega', \omega''') = r(\omega'', \omega') + r(\omega'', \omega''')$. Как и ранее индикаторная функция $g(\omega)$, определенная на Ω, принимающая значения из двуэлементного множества $\{1, -1\}$ и определяющее конкретное действительное разбиение множества всех объектов на два класса.

Пусть $\left((\omega_j, g_j),\ j = 1,..., N\right)$ – обучающая выборка, в которой «учитель» указал класс $g_j = g(\omega_j)$ каждого объекта. Если появляется новый объект $\omega \in \Omega$, то принять решение о его классе $\hat{g}(\omega)$ можно лишь на основании измерения расстояний этого объекта до всех объектов обучающей выборки, классы которых известны. В этой ситуации задача обучения распознаванию образов сводится к формированию решающего правила вида $\hat{g}(\omega) = \hat{g}\left(r(\omega, \omega_1),..., r(\omega, \omega_N)\right)$ на основании гипотезы компактности, согласно которой два объекта, характеризующиеся малым расстоянием между ними $r(\omega', \omega'')$ скорее всего принадлежат к одному и тому же классу.

Например, естественно ввести в рассмотрение линейную дискриминантную функцию

$$d(\omega \mid a_1,..., a_N) = \sum_{k:\, g(\omega_j) = -1} a_k r(\omega, \omega_k) - \sum_{l:\, g(\omega_l) = 1} a_l r(\omega, \omega_l) = \sum_{j=1}^{N} \left(-g_j r(\omega, \omega_j)\right) a_j ,$$

(3.2.1)

построенную как разность средневзвешенных расстояний вновь поступившего объекта до объектов второго и первого класса, где $a_j \geq 0$ – некоторые весовые коэффициенты, и принять решающее правило распознавания в виде

$$\hat{g}(\omega \mid a_1,...,a_N) = \begin{cases} 1, & d(\omega \mid a_1,...,a_N) \geq 0, \\ -1, & d(\omega \mid a_1,...,a_N) < 0. \end{cases} \qquad (3.2.2)$$

Такая дискриминантная функция и, соответственно, решающее правило, полностью характеризуются значениями неотрицательных весовых коэффициентов при объектах обучающей выборки $a_j \geq 0$, $j = 1,..., N$, играющих роль параметров.

Будем рассматривать совокупность расстояний произвольного объекта ω до всех элементов обучающей выборки ω_j с учетом их классов g_j как N-мерный вектор метрических признаков $\mathbf{x} = (x_1 \cdots x_N)^T \in \mathsf{R}^N$, $x_j = -g_j r(\omega, \omega_j)$, число которых совпадает с числом элементов выборки. Точно так же удобно ввести в рассмотрение N-мерный вектор весовых коэффициентов $\mathbf{a} = (a_1 \cdots a_N)^T \in \mathsf{R}^N$. Тогда каждая линейная дискриминантная функция из параметрического семейства (3.2.1)

$$d(\omega \mid a_1,..., a_N) = d(\mathbf{x} \mid \mathbf{a}) = \mathbf{a}^T \mathbf{x} = \sum_{j=1}^{N} a_j x_j ,$$

$$\hat{g}(\omega \mid a_1,...,a_N) = \hat{g}(\mathbf{x} \mid \mathbf{a}) = \begin{cases} 1, & d(\mathbf{x} \mid \mathbf{a}) > 0, \\ -1, & d(\mathbf{x} \mid \mathbf{a}) > 0, \end{cases}$$

определяет дискриминантную гиперплоскость $d(\mathbf{x} \mid \mathbf{a}) = \mathbf{a}^T \mathbf{x} = 0$, и соответственно решающее правило распознавания $\hat{g}(\omega \mid a_1,..., a_N) = \hat{g}(\mathbf{x} \mid \mathbf{a})$, в пространстве метрических признаков объекта относительно фиксированной совокупности объектов аналогично классическому случаю, когда признаки представляли собой результаты измерений произвольных свойств объекта. Специфика линейного пространства метрических признаков, образованного самой обучающей выборкой, заключается в том, что компоненты вектора параметров гиперплоскости выбираются лишь из множества неотрицательных значений $a_j \geq 0$, $j = 1,..., N$.

На этапе обучения нет другой информации о связи класса объекта с его расстояниями до других объектов, кроме обучающей выборки, поэтому, в

качестве критерия обучения следует принять правильность определения классов объектов в составе обучающей выборки $i = 1,..., N$

$$d(\omega_i \mid \mathbf{a}) = d(\mathbf{x}_i \mid \mathbf{a}) = \mathbf{a}^T \mathbf{x}_i = \sum_{j=1}^{N} a_j x_{ij} = \sum_{j=1}^{N} a_j \left(-g_j r(\omega_i, \omega_j) \right) \begin{cases} > 0 \text{ if } g_i = 1, \\ < 0 \text{ if } g_i = -1, \end{cases}$$

Такой принцип обучения в пространстве метрических признаков полностью аналогичен принципу обучения в произвольном линейном пространстве с тем лишь отличием, что направляющего вектора разделяющей гиперплоскости ищется лишь среди векторов с неотрицательными компонентами и среди параметров разделяющей гиперплоскости нет константы b. Даже если предположить, что гиперплоскость, в точности удовлетворяющая критерию обучения, не существует, ничто не мешает использовать тот же прием, который был применен в задаче обучения распознаванию образов в линейном пространстве общего вида, заключающийся в поиске таких минимальных сдвигов «мешающих» точек выборки в направлении «своего» класса $\delta_i \geq 0$, которые обеспечат существование разделяющей гиперплоскости. Повторяя рассуждения, приведенные в главе 2, мы придем к общей математической постановке задачи построения оптимального решающего правила распознавания в пространстве метрических признаков, аналогичной (2.2.19) с дополнительными ограничениями на неотрицательность компонент направляющего вектора разделяющей гиперплоскости $a_1 \geq 0,..., a_N \geq 0$:

$$\begin{aligned} &\frac{1}{2}\sum_{j=1}^{N} a_j^2 + C\sum_{j=1}^{N} \delta_j \to \min, C > 0, \\ &\sum_{k=1}^{N} (-g_j g_k r_{jk}) a_k \geq 1 - \delta_j, \delta_j \geq 0, a_j \geq 0, j = 1,..., N. \end{aligned} \quad . \qquad (3.2.3)$$

Для решения этой задачи составим функцию Лагранжа:

$$\begin{aligned} &L(a_1,...,a_N;\delta_1,...,\delta_N;\nu_1,...,\nu_N;\mu_1,...,\mu_N;\lambda_1,...,\lambda_N) = \\ &\frac{1}{2}\sum_{j=1}^{N} a_j^2 + C\sum_{j=1}^{N} \delta_j - \sum_{j=1}^{N} \lambda_j \left[\sum_{k=1}^{N} (-g_j g_k r_{jk}) a_k - 1 + \delta_j \right] - \sum_{j=1}^{N} \mu_j \delta_j - \sum_{j=1}^{N} \nu_j a_j \end{aligned} . \qquad (3.2.4)$$

Преобразуем функцию Лагранжа к более удобному виду

$$L(a_1,\dots,a_N;\delta_1,\dots,\delta_N;\nu_1,\dots,\nu_N;\mu_1,\dots,\mu_N;\lambda_1,\dots,\lambda_N)=$$

$$\frac{1}{2}\sum_{j=1}^{N}a_j^2+C\sum_{j=1}^{N}\delta_j-\sum_{j=1}^{N}\lambda_j\left[\sum_{k=1}^{N}(-g_jg_kr_{jk})a_k-1+\delta_j\right]-\sum_{j=1}^{N}\mu_j\delta_j-\sum_{j=1}^{N}\nu_ja_j=$$

$$\frac{1}{2}\sum_{j=1}^{N}a_j^2+C\sum_{j=1}^{N}\delta_j-\sum_{j=1}^{N}\lambda_j\sum_{k=1}^{N}(-g_jg_kr_{jk})a_k+\sum_{j=1}^{N}\lambda_j-\sum_{j=1}^{N}\lambda_j\delta_j-\sum_{j=1}^{N}\mu_j\delta_j-\sum_{j=1}^{N}\nu_ja_j=$$

$$\frac{1}{2}\sum_{j=1}^{N}a_j^2+C\sum_{j=1}^{N}\delta_j-\sum_{k=1}^{N}\lambda_k\sum_{j=1}^{N}(-g_jg_kr_{jk})a_j+\sum_{j=1}^{N}\lambda_j-\sum_{j=1}^{N}\lambda_j\delta_j-\sum_{j=1}^{N}\mu_j\delta_j-\sum_{j=1}^{N}\nu_ja_j=$$

$$\frac{1}{2}\sum_{j=1}^{N}a_j^2+C\sum_{j=1}^{N}\delta_j-\sum_{j=1}^{N}\left(\sum_{k=1}^{N}\lambda_k(-g_jg_kr_{jk})\right)a_j+\sum_{j=1}^{N}\lambda_j-\sum_{j=1}^{N}\lambda_j\delta_j-\sum_{j=1}^{N}\mu_j\delta_j-\sum_{j=1}^{N}\nu_ja_j=$$

$$\sum_{j=1}^{N}\left(\frac{1}{2}a_j-\nu_j-\sum_{k=1}^{N}\lambda_k(-g_jg_kr_{jk})\right)a_j+\sum_{j=1}^{N}\lambda_j-\sum_{j=1}^{N}(\lambda_j+\mu_j-C)\delta_j$$

Решением является седловая точка функции Лагранжа

$$L(a_1,\dots,a_N;\delta_1,\dots,\delta_N;\nu_1,\dots,\nu_N;\mu_1,\dots,\mu_N;\lambda_1,\dots,\lambda_N)\to\min_{a_j,\delta_j,\ j=1,\dots,N}$$

$$L(a_1,\dots,a_N;\delta_1,\dots,\delta_N;\nu_1,\dots,\nu_N;\mu_1,\dots,\mu_N;\lambda_1,\dots,\lambda_N)\to\max_{\nu_j\ge0,\,\mu_j\ge0,\,\lambda_j\ge0,\ j=1,\dots,N}$$

Первое из этих условий дает

$$\frac{\partial}{\partial a_j}L(a_1,\dots,a_N;\delta_1,\dots,\delta_N;\nu_1,\dots,\nu_N;\mu_1,\dots,\mu_N;\lambda_1,\dots,\lambda_N)=a_j-\nu_j-\sum_{k=1}^{N}\lambda_k(-g_jg_kr_{jk})=0$$

$$\boxed{a_j=\nu_j+\sum_{k=1}^{N}\lambda_k(-g_jg_kr_{jk})} \tag{3.2.5}$$

$$\frac{\partial}{\partial\delta_j}L(a_1,\dots,a_N;\delta_1,\dots,\delta_N;\nu_1,\dots,\nu_N;\mu_1,\dots,\mu_N;\lambda_1,\dots,\lambda_N)=$$

$$-(\lambda_j+\mu_j-C)=0,\quad j=1,\dots,N$$

$$\boxed{\lambda_j+\mu_j=C,\quad j=1,\dots,N} \tag{3.2.6}$$

Подстановка (3.2.5) и (3.2.6) в (3.2.4) дает целевую функцию при условно оптимальных значениях $\hat{a}_j,\hat{\delta}_j,j=1,\dots,N$

$$W(\nu_1,\dots,\nu_N;\lambda_1,\dots,\lambda_N)=-\frac{1}{2}\sum_{j=1}^{N}\hat{a}_j^2(\nu_j,\lambda_1,\dots,\lambda_N)+\sum_{j=1}^{N}\lambda_j=$$

$$-\frac{1}{2}\sum_{j=1}^{N}\left(\nu_j+\sum_{k=1}^{N}\lambda_k(-g_jg_kr_{jk})\right)^2+\sum_{j=1}^{N}\lambda_j=$$

$$-\frac{1}{2}\sum_{j=1}^{N}\left[\nu_j^2+2\nu_j\sum_{k=1}^{N}\lambda_k(-g_jg_kr_{jk})+\left(\sum_{k=1}^{N}\lambda_k(-g_jg_kr_{jk})\right)^2\right]+\sum_{j=1}^{N}\lambda_j=$$

$$-\frac{1}{2}\sum_{j=1}^{N}\nu_j^2-\sum_{j=1}^{N}\sum_{k=1}^{N}(-g_jg_kr_{jk})\nu_j\lambda_k-\frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}\left(\sum_{l=1}^{N}(-g_jg_lr_{jl})(-g_kg_lr_{kl})\right)\lambda_j\lambda_k+\sum_{j=1}^{N}\lambda_j$$

Т.о. мы приходим к двойственной задаче квадратичного программирования

$$W(\nu_1,...,\nu_N;\lambda_1,...,\lambda_N) =$$
$$\sum_{j=1}^{N}\lambda_j - \frac{1}{2}\sum_{j=1}^{N}\nu_j^2 - \sum_{j=1}^{N}\sum_{k=1}^{N}(-g_j g_k r_{jk})\nu_j\lambda_k - \frac{1}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}\left(\sum_{l=1}^{N}(-g_j g_l r_{jl})(-g_k g_l r_{kl})\right)\lambda_j\lambda_k \to \max,$$
$$\nu_j \geq 0, 0 \leq \lambda_j \leq C$$

если $\nu_j > 0$, то $a_j = 0$ и $\nu_j = -\sum_{k=1}^{N}\lambda_k(-g_j g_k r_{jk}) > 0$,

если $\nu_j = 0$, то $a_j = \sum_{k=1}^{N}\lambda_k(-g_j g_k r_{jk}) > 0$.

Или другой вид двойственной задачи:

$$W(a_1,...,a_N;\lambda_1,...,\lambda_N) = \frac{1}{2}\sum_{j=1}^{N}a_j^2 - \sum_{j=1}^{N}\lambda_j \to \min,$$
$$a_j - \sum_{k=1}^{N}\lambda_k(-g_j g_k r_{jk}) \geq 0, a_j \geq 0, 0 \leq \lambda_j \leq C, j = 1,...,N$$

По-прежнему значения $\lambda_j > 0$ указывают опорные элементы обучающей выборки

Правда, в отличие от линейного пространства общего вида обучающая выборка в пространстве метрических признаков всегда линейно разделима в классе гиперплоскостей $\mathbf{a}^T\mathbf{x} = 0$ с неотрицательными компонентами направляющего вектора в силу того, что число элементов в обучающей выборке совпадает с размерностью пространства, и что по определению метрических признаков $x_{ij} = -g_j r(\omega_i, \omega_j) = -r(\omega_i, \omega_j) \leq 0$ при определении расстояния данного объекта ω_i до объектов первого класса $g_j = 1$ и $x_{ij} = -g_j r(\omega_i, \omega_j) = r(\omega_i, \omega_j) \geq 0$ при определении расстояния до объектов второго класса $g_j = -1$. В силу этого обстоятельства при решении задачи квадратичного программирования, всегда будут получаться нулевые значения оптимальных сдвигов $\delta_i = 0$ для всех объектов $i = 1,..., N$.

Однако линейная разделимость обучающей выборки в пространстве метрических признаков обманчива, и в реальных задачах так строить процесс обучения нельзя. Дело в том, что решающее правило распознавания, построенное по обучающей выборке, будет хорошо работать на новых объектах лишь при условии, что число элементов выборки по крайней мере на порядок превосходит размерность пространства признаков [58]. В то же время, при обучении в пространстве метрических признаков число элементов выборки совпадает с размерностью пространства, и разделяющая гиперплоскость, оптимальная в смысле обеспечения максимального «зазора» между классами при ограничениях (3.2.3), будет слишком чувствительна к случайным аспектам пространственной формы обучающей выборки, в которых «утонет» информация

о фактической связи класса объекта распознавания с его расстояниями до объектов выборки. Для повышения статистической стабильности процесса обучения, или, как говорят, для его регуляризации, необходимо привлечение некоторой дополнительной априорной информации об ожидаемом направлении разделяющей гиперплоскости [18].

Заметим, что требование минимизации квадрата нормы направляющего вектора искомой разделяющей гиперплоскости $\mathbf{a}^T\mathbf{a} \to \min$ выражается присутствием в целевой функции задачи квадратичного программирования квадратичной формы с единичной матрицей $\mathbf{a}^T\mathbf{a} = \sum_{j=1}^{N} a_j^2$. Часто оказывается, что априорная информация, существенно снижающая свободу выбора разделяющей гиперплоскости, может быть формализована путем использования квадратичной функции с некоторой неединичной положительно определенной матрицей $\mathbf{a}^T\mathbf{Q}\mathbf{a}$. Мы приходим к формальной постановке задачи обучения в пространстве метрических признаков как задачи квадратичного программирования

$$\begin{aligned} &\frac{1}{2}\mathbf{a}^T\mathbf{B}\mathbf{a} + C\sum_{i=1}^{N}\delta_i = \sum_{j=1}^{N}\sum_{k=1}^{N}\beta_{jk}a_j a_k + C\sum_{i=1}^{N}\delta_i \to \min, \\ &g_i(\mathbf{a}^T\mathbf{x}_i) = g_i\left(\sum_{j=1}^{N} a_j x_{ij}\right) \geq 1-\delta_i, \quad \delta_i \geq 0, \quad i = 1,\ldots,N, \quad a_j \geq 0, \quad j = 1,\ldots,N. \end{aligned} \tag{3.2.4}$$

При такой постановке задачи обучения гиперплоскость, в точности разделяющая объекты разных классов в обучающей выборке, может оказаться невыгодной с точки зрения априорных предпочтений, выражаемых матрицей $\mathbf{B}$. В результате оптимальная гиперплоскость пожертвует, если понадобится, правильной классификацией некоторых наиболее «диких» объектов выборки, что выразится в положительных значениях их сдвигов $\delta_i > 0$. При правильном выборе матрицы $\mathbf{B}$ в учетом имеющейся априорной информации доля таких объектов будет более лучше соответствовать фактической разрешимости задачи распознавания, и при применении решающего правила (3.2.2.) к новым объектам ω будет существенно повышена надежность распознавания.

ω_j

Специфической особенностью метрических признаков $x_j = r(\omega, \omega_j)$ всякого объекта ω является то обстоятельство что они образованы некоторой совокупностью реальных базовых объектов $\omega_1, \ldots, \omega_N$ которые сами характеризуются взаимными расстояниями согласно той же метрики. Если расстояние $r(\omega_j, \omega_k)$ мало, т.е. они близко расположены друг к другу в метрическом пространстве то соответствующие признаки x_j и x_k не несут

существенно разной информации об объекте распознавания ω. Линейная дискриминантная функция имеет вид линейной комбинации расстояний объекта до базовых объектов $\mathbf{a}^T\mathbf{x} = \quad = \sum_{j=1}^{N} a_j x_{ij} = \sum_{j=1}^{N} a_j \left(-g_j r(\omega_i, \omega_j)\right)$, и разные значения коэффициентов a_j и a_k имеют смысл только в том случае, если помогают учесть разную роль этих двух признаков в определении класса объекта. Матрица расстояний между базовыми признаками является, в сущности, матрицей расстояний между признаками, и при малом значении $r(\omega_j, \omega_k)$ для некоторой пары признаков соответствующие коэффициенты a_j и a_k должны мало отличаться друг от друга. В этом и заключается априорная информация о направляющем векторе разделяющей гиперплоскости, которую предлагается учитывать в процессе обучения в пространстве метрических признаков.

Эту дополнительную информацию можно внести в критерий обучения (3.2.3) в виде дополнительного квадратичного члена

$$J'(\mathbf{a}) = \frac{\alpha}{2}\sum_{j=1}^{N}\sum_{k=1}^{N}\tilde{\beta}_{jk}\left(a_j - a_k\right)^2 = \alpha\mathbf{a}^T\tilde{\mathbf{B}}\mathbf{a}, \quad \tilde{\mathbf{B}} = \begin{bmatrix} -\tilde{\beta}_{11} + \sum_{i=1}^{N}\tilde{\beta}_{1i} & \cdots & -\tilde{\beta}_{1N} \\ \vdots & \ddots & \vdots \\ -\tilde{\beta}_{N1} & \cdots & -\tilde{\beta}_{NN} + \sum_{i=1}^{N}\tilde{\beta}_{Ni} \end{bmatrix},$$

слагаемые которого имеют смысл штрафов на попарные различия коэффициентов a_j и a_k, общая величина которого регулируется выбором значения коэффициента $\alpha \geq 0$. Штрафные коэффициенты $\tilde{\beta}_{jk}$ должны возрастать с уменьшением расстояния $r(\omega_j, \omega_k)$ между признаками, т.е. объектами обучающей выборки, стремясь к бесконечно большим $\tilde{\beta}_{jk} \to \infty$ при $r(\omega_j, \omega_k) \to 0$. Например такому условию удовлетворяет соотношение $\tilde{\beta}_{jk} = 1/\left[r(\omega_j, \omega_k)\right]^{\gamma}$, где $\gamma > 0$ - параметр определяющий скорость возрастания штрафа при сближении признаков.

Тогда целевая функция (3.2.3.) будет включать в себя еще одно слагаемое

$$\frac{1}{2}\left(\mathbf{a}^T\mathbf{a} + \alpha\mathbf{a}^T\tilde{\mathbf{B}}\mathbf{a}\right) + C\sum_{j=1}^{N}\delta_j \to \min$$

или более компактно

$$\frac{1}{2}\mathbf{a}^T\mathbf{B}\mathbf{a} + C\sum_{j=1}^{N}\delta_j \to \min, \text{ где } \mathbf{B} = \mathbf{I} + \alpha\tilde{\mathbf{B}},$$

при ограничениях

$$g_i(\mathbf{a}^T\mathbf{x}_i) = g_i\left(\sum_{j=1}^{N} a_j x_{ij}\right) \geq 1 - \delta_i, \quad \delta_i \geq 0, \quad i = 1,...,N, \quad a_j \geq 0, \quad j = 1,...,N.$$

Служить мерой сходства и различия между объектами, а следовательно и между признаками, может не сама метрика $r(\omega', \omega'')$, а некоторая функция от нее $\mu(\omega', \omega'') = \varphi[r(\omega', \omega'')]$. Если используется убывающая функция, такая что $\varphi(r) \to 0$ при $r \to \infty$ и $\varphi(0) = \mu^0$ то характеристика симметричного парного отношения между объектами $\mu(\omega', \omega'') = \mu(\omega'', \omega')$ является мерой близости между ними принимающей минимальное значение $\mu(\omega, \omega) = \mu^0$ при совпадении двух объектов.

В частности в задачах молекулярной биологии сходства и различия между парами белков часто характеризуют близостью, определение которой строится на понятии гипотетической общей последовательности, из которой обе сравниваемые последовательности могут быть получены небольшим числом единичных пропусков вставок или замен аминокислот. При таком определении величина близости оказывается зависящей от длин сравниваемых последовательностей, а разные последовательности характеризуются разными значениями близости к самим себе.

В таком случае матрица близости между объектами обучающей выборки $\mu(\omega_j, \omega_k)$, $j, k = 1, \ldots, N$ должна быть подвергнута предварительному преобразованию так, что бы диагональные элементы имели одинаковое значение. Это можно сделать, например, разделив каждое значение $\mu(\omega_j, \omega_k)$ на величину $\sqrt{\mu(\omega_j, \omega_j)}\sqrt{\mu(\omega_k, \omega_k)}$. После такой операции матрица останется симметричной, а все ее диагональные элементы будут иметь единичное значение, соответствующее максимальному значению близости при полном совпадении объектов.