Лекция 9. Нелинейные преобразования и спрямляющие пространства

Представленные выше простейшие алгоритмы распознавания с использованием функций расстояния основаны на применении линейных разделяющих функций и их комбинаций. Получаемые при этом границы областей решений являются гиперплоскостями (в случае двух классов), либо имеют кусочно-линейный вид и образуются из фрагментов гиперплоскостей (в случае многих классов), попарно разделяющих классы. Ранее при решении многих задач распознавания в рамках статистического подхода были получены аналогичные по смыслу результаты.

Удобство и эффективность использований линейных разделяющих функций привели к появлению целого направления в теории распознавания образов, основанного на построении оптимальных в классе линейных разделяющих функций с использованием различных методов и методик как в рамках статистического, так и в рамках детерминистского подходов. При этом строго оптимальная разделяющая функция может быть и нелинейной, например, как в задаче о распознавании образов, описываемых векторами с различными матрицами ковариации (матрицами рассеяния). Главное, чтобы данные, представленные в виде образов - прецедентов различных классов, были бы в этом случае **почти (с допустимой ошибкой) линейно разделимы** в своем исходном признаковом представлении.

1. Нелинейные преобразования с повышением размерности

В случае же, когда данные с приемлемой ошибкой принципиально линейно не разделимы в исходном признаковом пространстве R^n, используется общий для многих алгоритмов подход.

Он состоит в том, что выполняется преобразование образов обучающих выборок в образы другого пространства большей размерности на основе отображения

$$\varphi:\ R^n \to R^m,\ \ n < m,$$

в котором преобразованные образы будут уже линейно разделимы. После этого строится линейный алгоритм попарного распознавания классов, который легко преобразуется в многоальтернативное решающее правило.

Таким образом, в итоге получается алгоритм распознавания с линейной разделяющей функцией, которая относительно исходных признаков будет, тем не менее, существенно нелинейным.

Теоретические обоснования подобного подхода заложены классической **теоремой Ковера**, которая формулируется следующим образом: ***нелинейное преобразование сложной задачи классификации образов в пространство более высокой размерности повышает вероятность линейной разделимости образов***.

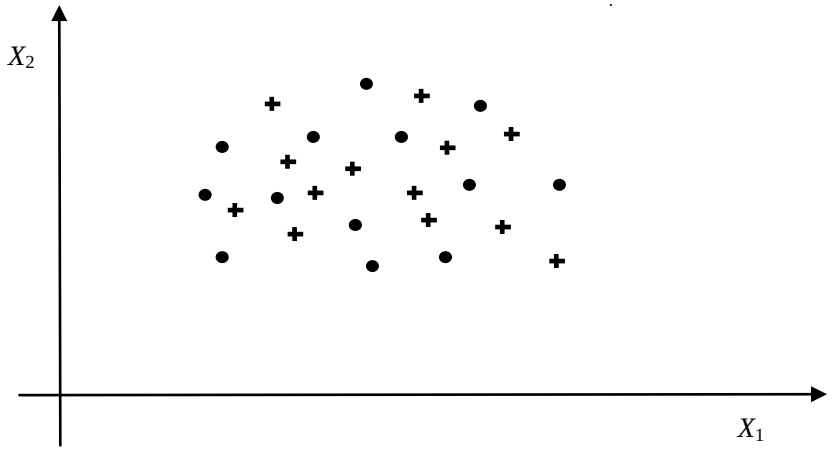


Рис.1. Пример линейно не разделимых данных

Можно доказать, что для любой локализации образов двух классов, входящих в состав обучающей выборки образов можно построить такое нелинейное преобразование, которое позволяет однозначно разделить эти образы в новом пространстве на основе линейной функции.

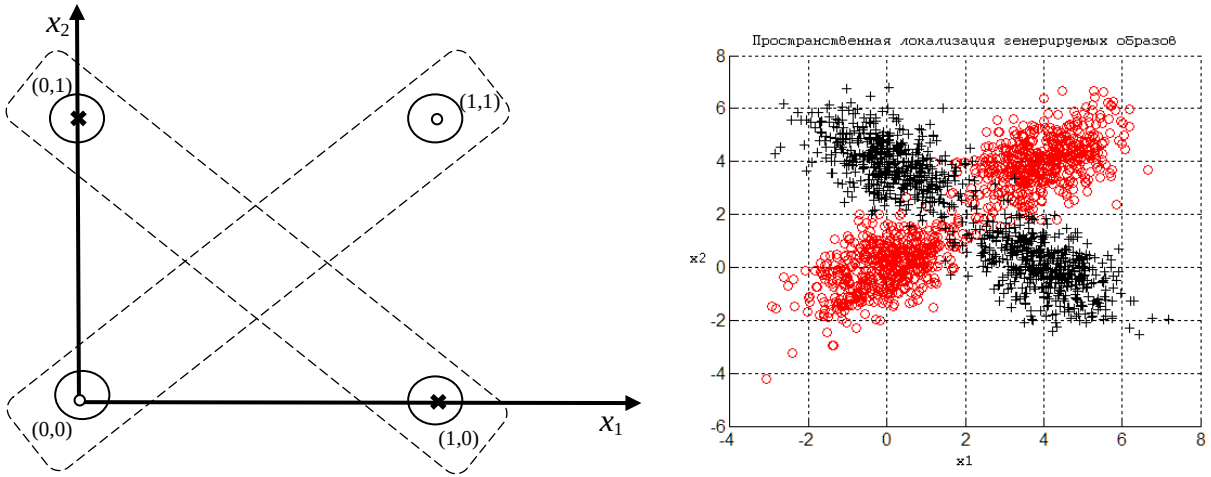
Рассмотрим классический пример с нелинейно неразделимыми образами: задачу «исключающего или». Как известно, логическая функция «исключающего или» относительно двух булевых переменных, объединенный в общий вектор входных данных $x = (x_1, x_2)^T$, формируется следующим образом:

Таблица 6.1 – Определение функции «исключающего или»

x_1	0	0	1	1
x_2	0	1	0	1
y	0	1	1	0

Получаемые таким образом значения y интерпретируются как индексы двух классов, каждый из которых представлен двумя образами $\omega_1: \left\{x^{(1,1)} = (0,0)^T, x^{(1,2)} = (1,1)^T\right\} \to y = 0$, $\omega_2: \left\{x^{(2,1)} = (1,0)^T, x^{(22)} = (0,1)^T\right\} \to y = 1$.

Конфигурация образов указанных классов показана на рис.2а.



а) б)

Рис.2. Иллюстрация к проблеме «исключающего или» (а) и конфигурация данных «заплетенные восьмерки» (б)

Таким образом, задача сводится к построению разделяющей функции, определяющей классы образов для выдачи правильной реакции $y = f(x)$ на входные воздействия. Визуальный анализ рис.2а показывает, что не существует линии, разделяющей выделенные прецеденты указанных классов (строгое доказательство можно получить и математически).

Другой пример линейно не разделимых данных двух классов может быть получен с помощью специальной функции генерации смесей ГСВ. Полученную на ее основе и представленную ниже на рис.1б конфигурацию образов двух классов мы называем «заплетенные восьмерки».

Степень «заплетенности» определяется параметром dm, задающим относительный сдвиг центров областей локализации данных в каждом классе (рис.1б). Если dm=0, то классы сливаются, если dm становится достаточно большим, то каждый класс состоит из двух не пересекающихся областей эллипсоидальной формы.

Вернемся к модели преобразования данных «исключающее или» и выполним на ее основе простое обоснование эффективности полхода, основанного на выполнении нелинейного преобразования с повышением размерности признакового пространства. Самым простым способом построить такое преобразование является использование так называемых радиальных базисных функции (РБФ), центрированных относительно точек, в которых находятся образы различных классов. Каждая такая функция радиально изменяет свои значения относительного выбранного центра – вектора c, принимает ненулевые значения в окрестности этого центра и стремится к нулю в области определения при удалении от центра. Подобные функции, определяемые в виде $\varphi(x)=\varphi(\|x-c\|)$, чаще всего задают в виде гауссианы

$$\varphi_j(x)=\varphi(\|x-c^{(j)}\|)=\exp\left(-\frac{\|x-c^{(j)}\|^2}{2\sigma_j^2}\right), \quad j=\overline{1,N},$$

где j – индексы центров РБФ.

Параметр σ_j отвечает за скорость стремления функций к нулю при $\|x-c^{(j)}\|\to\infty$. В нашем случае можно ограничиться заданием одинакового значения $\sigma=\sigma_j, j=\overline{1,N}$. Назначим в качестве центров РБФ точки, соответствующие исходным образам объединения множеств $\omega_1\cup\omega_2$, т.е. общего множества обучающих данных

$$c^{(1)}=x^{(1,1)}=(0,0)^T, c^{(2)}=x^{(1,2)}=(1,1)^T, c^{(3)}=x^{(2,1)}=(1,0)^T, c^{(4)}=x^{(2,2)}=(0,1)^T.$$

Тогда получим нелинейное преобразование исходного признакового пространства в четырехмерное

$$\varphi:\ R^2 \to R^4,\ y = \varphi(x) = \left(y_1, y_2, y_3, y_4\right)^T = \left(\varphi_1(x), \varphi_2(x), \varphi_3(x), \varphi_4(x)\right)^T.$$

При этом выберем параметр $\sigma = \sigma_j,\ j = \overline{1,4}$ таким образом, чтобы области локализации ненулевых значений, как это показано на рис.2а, у разных РБФ не пересекались. Для этого, например, достаточно выбрать σ так, что при $\left\|x - c^{(j)}\right\| > 0.25$ значения $\varphi_j(x) = \varphi\left(\left\|x - c^{(j)}\right\|\right) < \varepsilon$, где $\varepsilon > 0$ малая величина. Тогда для изначально линейно не разделимых данных в новом признаковом пространстве можно построить линейную разделяющую функцию. Для этого зададим в качестве весовых коэффициентов функции следующие значения

$$w_1 = \varphi_1(x^{(1,1)}) = 1,\ w_2 = \varphi_2(x^{(1,2)}) = 1,\ w_3 = -\varphi_3(x^{(2,1)}) = -1,\ w_4 = -\varphi_4(x^{(2,2)}) = -1.$$

В итоге получим, что линейная функция $g(y) = y^T w = \sum_{j=1}^{4} y_j w_j$ гарантированно имеет следующие значения при подстановке исходных точек (образов) первого и второго класса

$$\omega_1:\ g(y) = y^T w \geq 1 - 2\varepsilon > 0,\quad \omega_2:\ g(y) = y^T w \leq -1 + 2\varepsilon < 0.$$

При задании соответствующего значения ε это определяет линейную разделимость исходных данных в новом (четырехмерном) признаковом пространстве. И, хотя в новом признаковом пространстве разделяющая поверхность является линейной, после ее обратного проецирования на исходное пространство она окажется нелинейной.

Получаемые при применении такого подхода пространства называют спрямляющими пространствами. В качестве функций, используемых при выполнении нелинейного преобразования в спрямляющее пространство, могут применяться функции, обладающие необходимыми свойствами.

Очевидно, что прием, основанный на «окружении» каждого образа в исходном пространстве признаков с помощью РБФ-функции или ее аналогов, позволяет получить линейно разделимые совокупности данных для любой конфигурации обучающих данных. Т.е. в пространстве, имеющем, как максимум, размерность $N = N_1 + N_2$, проблема линейной неразделимости образов однозначно снимается. Очевидно также, что такое преобразование и реализация на его основе линейного классификатора является избыточным и ресурсоемким, поэтому количество нелинейных элементов, используемых при решении задачи желательно сократить.

2. Ядра скалярных произведений

Проблема заключается в том, что на практике заранее нельзя сказать, какие именно базисные функции нужно взять, чтобы добиться линейной разделимости, поэтому приходится брать сразу большой набор таких функций. В этом случае число признаков оказывается очень большим, из-за чего процесс обучения становится чрезмерно ресурсоемким. Однако в некоторых случаях оказывается, что достаточно уметь вычислять только скалярные произведения образов друг относительно друга. Действительно, мы видим, что практически все линейные правила принятия решений после перехода в спрямляющее пространство, имеют вид

$$\alpha = sign\left[y^T w + w_0\right] = sign\left[\varphi(x)^T w + w_0\right],$$

где вектор w после обучения и использования эталонов определяется образами обучающей выборки.

Например, если в исходном признаковом пространстве алгоритм распознавания по эталонам использует скалярные произведения вида $x^T z^{(*)}$, то в спрямляющем пространстве используется скалярные произведения $\varphi(x)^T \varphi(z^{(*)})$. Это означает, прежде всего, что **спрямляющее пространство должно быть наделено скалярным произведением**, как например, евклидово пространство.

Другая принципиальная идея состоит в том, чтобы использовать вместо прямого скалярного произведения образов – нелинейную функцию так называемого ядра скалярного произведения, в которой подобные скалярные произведения фигурировали бы в неявном виде, и настраивать классификатор в новом признаковом пространстве, подбирая весовые коэффициенты уже для функций ядра.

Данный прием, позволяющий строить нелинейные разделители, получил в англоязычной литературе название **kernel trick**. Результирующий алгоритм похож на алгоритм линейной классификации, с той лишь разницей, что каждое скалярное произведение заменяется нелинейной функцией ядра (реализующей скалярное произведением в пространстве с большей размерностью).

Ядром скалярного произведения здесь и далее мы будем называть функцию, представимую в виде скалярного произведения в некотором пространстве: $K(x,z)=\varphi(x)^T\varphi(z)$, где $\varphi\colon R^n\to R^m$ отображение из исходного признакового пространства в некоторое спрямляющее. Ядро содержит в себе много информации о спрямляющем пространстве и позволяет производить в нем различные операции, не задавая в явном виде самого отображения $\varphi(x)$.

Рассмотрим, например, задачу нахождения расстояния между векторами $y=\varphi(x)$ и $y'=\varphi(z)$, если известно только функция ядра $K(x,z)$

$$d_E(y,y')=\left[(y-y')^T(y-y')\right]^{\frac{1}{2}}=\left[\varphi(x)^T\varphi(x)-2\varphi(x)^T\varphi(z)+\varphi(z)^T\varphi(z)\right]^{\frac{1}{2}}=$$

$$=\left[K(x,x)-2K(x,z)+K(z,z)\right]^{\frac{1}{2}}.$$

Таким образом, ядра можно напрямую использовать в методах распознавания по мере близости.

Простейший способ задать ядро состоит в том, чтобы в явном виде построить отображение $\varphi(x)$ в спрямляющее признаковое пространство. Тогда ядро, как уже упоминалось, определяется как скалярное произведение образов в этом пространстве $K(x,z)=\varphi(x)^T\varphi(z)$. При таком способе, однако, опять возникают проблемы с ростом вычислительной сложности.

Ряд примеров показывает, что ядро выражается через скалярное произведение в исходном пространстве и для его вычисления необходимо порядка d операций, в то время как прямое вычисление ядра потребовало бы d^2 операций. При этом желательно определить ядро так, чтобы оно в явном виде не использовало отображение объектов в новое признаковое пространство.

Чтобы убедится, что используемая функция $K(x,z)$ определяет скалярное произведение в некотором пространстве, можно опереться на **теорему Мерсера**, которая формулируется следующим образом**: *функция* $K(x,z)$ *является ядром тогда и только тогда, когда она симметрична* $K(x,z)=K(z,x)$ *и неотрицательно определена, то есть для любой конечной выборки* $X^N=\{x^{(1)},...,x^{(N)}\}$ *матрица* $\left\|K(x^{(i)},x^{(j)}\right\|$ *неотрицательно определена.***

На практике для упрощения анализа вместо проверки на выполнение указанных условий используют типовые функции ядра, обладающие необходимыми свойствами. К ним относятся функции следующего вида:

$$K(x,z)=\exp\left(-\frac{\|x-z\|^2}{2\sigma^2}\right),\quad K(x,z)=(x^Tz+1)^p,\quad K(x,z)=\tanh(ax^Tz+b)$$

Последняя функция является функцией гиперболического тангенса, для нее свойства ядра выполняются не для всех значений коэффициентов.