Методическая разработка

для проведения лекции

Занятие 13. Базисные функции при непараметрическом распознавании образов

Учебные вопросы занятия:

1. Моделирование работы байесовского классификатора на основе аппроксимации плотностей распределения функциями (многочлены Лагерра)
2. Моделирование работы байесовского классификатора на основе аппроксимации плотностей распределения полиномиальными функциями Радемахера-Уолша

Заключительная часть

1. Моделирование работы байесовского классификатора на основе аппроксимации плотностей распределения функциями (многочлены Лагерра)

В настоящее время исследователи не ограничиваются применением только тригонометрической системы функций для разложения сигналов. Часто в качестве базисных функций используются многочлены Эрмита, Лагерра, Чебышева. Среди всего многообразия используемых систем ортогональных функций заметное место занимает система функций Лагерра. Это объясняется тем, что функции Лагерра обладают рядом достоинств.

Функции Лагерра получают с помощью ортогональных полиномов. Полиномы Лагерра ортогональны на полуоси $0 < \tau < \infty$ с весом $\rho(\tau) = \exp(-\tau)$, то есть они удовлетворяют условию

$$\int e^{-\tau} L_n(\tau) L_m(\tau) d\tau = \begin{cases} r_n = (n!)^2, & \text{если } n = m; \\ 0, & \text{если } n \neq m. \end{cases} \qquad (29)$$

После замены $\tau = 2\alpha t$ и умножения на нормирующий коэффициент $\sqrt{2\alpha}$ первые пять функций Лагерра принимают вид:

$$\varphi_0(t) = \sqrt{2\alpha} \cdot e^{-\alpha t};$$

$$\varphi_1(t) = \sqrt{2\alpha} \cdot e^{-\alpha t}(1 - 2\alpha t);$$

$$\varphi_2(t) = \sqrt{2\alpha} \cdot e^{-\alpha t}(1 - 4\alpha t + 2\alpha^2 t^2);$$

$$\varphi_3(t) = \sqrt{2\alpha} \cdot e^{-\alpha t}(1 - 6\alpha t + 6\alpha^2 t^2 - 4\alpha^3 t^3 / 3);$$

$$\varphi_4(t) = \sqrt{2\alpha} \cdot e^{-\alpha t}(1 - 8\alpha t + 12\alpha^2 t^2 - 16\alpha^3 t^3 / 3 + 2\alpha^4 t^4 / 3),$$

где α – масштабный коэффициент.

В общем виде функции описываются формулой

$$\varphi_n(t) = \sqrt{2\alpha} \cdot e^{-\alpha t} \sum_{j=0}^{n} (-1)^j \frac{C_n^j}{j!} (2\alpha t)^j,$$

где C_n^j – число сочетаний из n по j.

Функции Лагерра образуют полную и ортогональную систему на одностороннем интервале $[0,\infty)$, то есть они удовлетворяют соотношению

$$\int \varphi_n(t)\varphi_m(t)dt = \begin{cases} 1, & \text{если } n = m; \\ 0, & \text{если } n \neq m. \end{cases}$$

Важным пунктом спектрального анализа с использованием функций Лагерра является выбор значения масштабного коэффициента α. Его начальное значение рекомендуется выбирать так, чтобы длительности исследуемого сигнала и функции Лагерра с номером $i \approx \frac{N}{2}$ были примерно равны. В последующем значение коэффициента α уточняется. Функции Лагерра получили широкое распространение в системах обработки сигналов различного назначения. Это в значительной степени объясняется простотой их генерирования. Оказывается, что функция Лагерра $\varphi_i(t)$ по форме совпадает с импульсной характеристикой системы, состоящей из последовательно соединенных простых электрических цепей.

Пример 1.

Пусть существует некоторая случайная величина *t*, представленная выборкой значений $\{t_i\} = \{40,35,37,20,38,42\}$.

Необходимо оценить закон распределения методом линейной комбинации базисных функций, используя в качестве базисных функций многочлены Лагерра.

В качестве базисных функций воспользуемся следующими многочленами

$$\varphi_0(t) = \sqrt{2\alpha}\cdot e^{-\alpha t};\ \varphi_1(t) = \sqrt{2\alpha}\cdot e^{-\alpha t}(1-2\alpha t);\ \varphi_2(t) = \sqrt{2\alpha}\cdot e^{-\alpha t}(1-4\alpha t+2\alpha^2 t^2).$$

Вычислим коэффициенты при данных функциях

$$c_0 = \frac{1}{6}\left[\varphi_0(t_1)+\varphi_0(t_2)+\ldots+\varphi_0(t_6)\right] = \frac{1}{6}\left(0,043+0,055+0,05+0,116+0047+0,039\right) = 0,058$$

$$c_1 = \frac{1}{6}\left(-0,128-0,137-0,134-0,116-0,132-0,124\right) = -0,129,$$

$$c_2 = \frac{1}{6}\left(0,043+0,007+0,022-0,116+),029+0,055\right) = 0,007.$$

Следовательно, закон распределения можно аппроксимировать следующим образом

$$\tilde{f}(x) = c_0\varphi_0(t)+c_1\varphi_1(t)+c_2\varphi_2(t) = 0,058\sqrt{2\alpha}\cdot e^{-\alpha t} - 0,129\sqrt{2\alpha}\cdot e^{-\alpha t}(1-2\alpha t) + \\ +0,007\sqrt{2\alpha}\cdot e^{-\alpha t}(1-4\alpha t+2\alpha^2 t^2).$$

Пример 2.

Необходимо построить модель байесовского классификатора текстовых сообщений на русском языке в соответствии с тематическими направленностями (специальная военная операция, новогодние праздники). Для обучения использовать предварительно сформированные текстовые корпуса (обучающие выборки) по каждой тематике. Обосновать выбор непараметрических методов обучения на основе проверки закона

распределения для соответствующих выборок с использованием критерия согласия Смирнова (ω^2) (достоверность критерия d принять равной 0,95).

Решение.

На первом этапе необходимо обработать обучающие текстовые сообщения на предмет оценки частоты появления информативных признаков и формирования выборок в цифровой форме. С учетом заданных тематик сообщений в качестве информативного признака предлагается использовать морфему «укр».

После обработки текстового корпуса по теме «специальная военная операция» получена следующая выборка значений информативного признака *t*

$\{t_i\} = \{17,11,18,16,36,19,11\}$.

После обработки текстового корпуса по теме «новогодние праздники» получена следующая выборка значений информативного признака *t*

$\{t_i\} = \{3,4,4,3,2,5,4\}$.

На втором этапе необходимо проверить сформированные обучающие выборки на соответствие известным законам распределения (в частности, нормальному закону). Принимая во внимание малый объем выборки для решения этой задачи предлагается использовать критерий согласия Смирнова (ω^2) с достоверностью равной 0,95. Для данного значения достоверности критическое значения показателя согласованности, взятое из таблицы, принимает значение $\omega_d^2 = 0{,}126$. С учетом того, что объем выборки $n = 7 < 40$, пересчитываем критическое значение по формуле $\left(\omega_d^2\right)' = \left(\omega_d^2 - \frac{0{,}4}{n} + \frac{0{,}6}{n^2}\right)\left(1 + \frac{1}{n}\right)$ и получаем новое критическое значение $\left(\omega_d^2\right)' = 0{,}093$.

Проверим на соответствие нормальному закону распределения выборку по теме «новогодние праздники».

Упорядочим выборку по возрастанию элементов:

$\{t_i\} = \{2,3,3,4,4,4,5\}$.

Оценим математическое ожидание и дисперсию выборки:

$\tilde{m} = \frac{1}{7}\sum_{i=1}^{7} t_i = 3{,}57$, $\tilde{\sigma}^2 = \frac{1}{6}\sum_{i=1}^{7}(t_i - \tilde{m})^2 = 0{,}952$.

Определим значения $F(t_i, \tilde{m}, \tilde{\sigma}^2)$:

$$F(t_1, \tilde{m}, \tilde{\sigma}^2) = \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \int_{-\infty}^{t_1} e^{-\frac{(t-\tilde{m})^2}{2\tilde{\sigma}^2}} dt = 0{,}054; \quad F(t_2, \tilde{m}, \tilde{\sigma}^2) = 0{,}279; \quad F(t_3, \tilde{m}, \tilde{\sigma}^2) = 0{,}279;$$

$F(t_4, \tilde{m}, \tilde{\sigma}^2) = 0{,}67$; $F(t_5, \tilde{m}, \tilde{\sigma}^2) = 0{,}67$; $F(t_6, \tilde{m}, \tilde{\sigma}^2) = 0{,}67$; $F(t_7, \tilde{m}, \tilde{\sigma}^2) = 0{,}928$.

Определим значение показателя согласованности:

$$\omega^2 = \frac{1}{12 \cdot 7} + \sum_{i=1}^{7}\left[F\left(t_i, \tilde{m}, \tilde{\sigma}^2\right) - \frac{2i-1}{2 \cdot 7} \right]^2 = 0{,}065.$$

Проверяем условие: $\omega^2 < \left(\omega_d^2\right)'$.

Так как данное условие выполняется, гипотеза о нормальном законе распределения с достоверностью d = 0,95 для данной выборки принимается.

Проверим на соответствие нормальному закону распределения выборку по теме «специальная военная операция».

Упорядочим выборку по возрастанию элементов:

$\{t_i\} = \{11,11,16,17,18,19,36\}$.

Оценим математическое ожидание и дисперсию выборки:

$\tilde{m} = \frac{1}{7}\sum_{i=1}^{7} t_i = 18,286$, $\tilde{\sigma}^2 = \frac{1}{6}\sum_{i=1}^{7}(t_i - \tilde{m})^2 = 71,238$.

Определим значения $F(t_i, \tilde{m}, \tilde{\sigma}^2)$:

$$F(t_1, \tilde{m}, \tilde{\sigma}^2) = \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \int_{-\infty}^{t_1} e^{-\frac{(t-\tilde{m})^2}{2\tilde{\sigma}^2}} dt = 0,194; \quad F(t_2, \tilde{m}, \tilde{\sigma}^2) = 0,194; \quad F(t_3, \tilde{m}, \tilde{\sigma}^2) = 0,393;$$

$$F(t_4, \tilde{m}, \tilde{\sigma}^2) = 0,439;\ F(t_5, \tilde{m}, \tilde{\sigma}^2) = 0,486;\ F(t_6, \tilde{m}, \tilde{\sigma}^2) = 0,534;\ F(t_7, \tilde{m}, \tilde{\sigma}^2) = 0,982.$$

Определим значение показателя согласованности:

$$\omega^2 = \frac{1}{12 \cdot 7} + \sum_{i=1}^{7}\left[F\left(t_i, \tilde{m}, \tilde{\sigma}^2\right) - \frac{2i-1}{2 \cdot 7} \right]^2 = 0,123.$$

Проверяем условие: $\omega^2 < \left(\omega_d^2\right)'$.

Так как данное условие не выполняется, гипотеза о нормальном законе распределения с достоверностью d = 0,95 для данной выборки отвергается.

Этот результат обусловливает целесообразность аппроксимации законов распределения анализируемых выборок с использованием непараметрических методов обучения. В качестве таковых будем использовать метод линейной комбинации базисных функций на основе многочленов Лагерра.

На третьем этапе необходимо обучить модель классификатора методом линейной комбинации базисных функций.

Базисные функции предварительно сформируем на основе следующего математического выражения

$$\varphi_k(t) = \sqrt{2\alpha} \cdot e^{-\alpha t} \sum_{j=0}^{k} (-1)^j \frac{C_k^j}{j!} (2\alpha t)^j,$$

где C_k^j – число сочетаний из k по j.

Получим следующие многочлены:

$$\varphi_0(t) = \sqrt{2\alpha} \cdot e^{-\alpha t};$$

$$\varphi_1(t) = \sqrt{2\alpha} \cdot e^{-\alpha t}(1 - 2\alpha t);$$

$$\varphi_2(t) = \sqrt{2\alpha} \cdot e^{-\alpha t}(1 - 4\alpha t + 2\alpha^2 t^2);$$

$$\varphi_3(t) = \sqrt{2\alpha} \cdot e^{-\alpha t}(1 - 6\alpha t + 6\alpha^2 t^2 - 4\alpha^3 t^3 / 3);$$

$$\varphi_4(t) = \sqrt{2\alpha} \cdot e^{-\alpha t}(1 - 8\alpha t + 12\alpha^2 t^2 - 16\alpha^3 t^3 / 3 + 2\alpha^4 t^4 / 3),$$

где масштабному коэффициенту α присвоим значение 0,1.

Вычислим коэффициенты при данных функциях. Эти коэффициенты можно вычислить, воспользовавшись выражением

$$c_k^{(i)} = \frac{1}{n_i}\sum_{j=1}^{n_i}\varphi_k\left(t_{ij}\right),$$

где n_i – число образов, входящих в класс S_i, j изменяется в диапазоне от 1 до n_i (в нашем случае 7).

Для образов по теме «специальная военная операция» применение данной процедуры дает

$$c_0^{(1)} = \frac{1}{7}\left[\varphi_0\left(t_1^{(1)}\right)+\varphi_0\left(t_2^{(1)}\right)+...+\varphi_0\left(t_7^{(1)}\right)\right]=0{,}089;$$

$$c_1^{(1)} = -0{,}173;\ c_2^{(1)} = -0{,}015;\ c_3^{(1)} = 0{,}061;\ c_4^{(1)} = 0{,}071.$$

Применение этой же процедуры к образам по теме «новогодние праздники» приводит к следующим коэффициентам

$$c_0^{(2)} = 0{,}314;\ c_1^{(2)} = 0{,}095;\ c_2^{(2)} = -0{,}043;\ c_3^{(2)} = -0{,}12;\ c_4^{(2)} = -0{,}154.$$

Плотности распределения для каждого класса можно аппроксимировать, применив выражения вида

$$\omega\left(t/s_i\right) = \sum_{j=0}^{m-1} c_j^{(i)}\varphi_j\left(t\right)\ \text{(m – число базисных функций).}$$

Следовательно, плотности распределения вероятностей можно аппроксимировать следующим образом

$$\tilde{\omega}\left(t/S_1\right) = c_0^1\varphi_0\left(t\right)+c_1^1\varphi_1\left(t\right)+c_2^1\varphi_2\left(t\right)+c_3^1\varphi_3\left(t\right)+c_4^1\varphi_4\left(t\right)=$$

$$=0{,}089\sqrt{2\alpha}\cdot e^{-\alpha t}-0{,}173\sqrt{2\alpha}\cdot e^{-\alpha t}(1-2\alpha t)-0{,}015\sqrt{2\alpha}\cdot e^{-\alpha t}(1-4\alpha t+2\alpha^2t^2)+$$

$$+0{,}061\sqrt{2\alpha}\cdot e^{-\alpha t}(1-6\alpha t+6\alpha^2t^2-4\alpha^3t^3/3)+$$

$$+0{,}071\sqrt{2\alpha}\cdot e^{-\alpha t}(1-8\alpha t+12\alpha^2t^2-16\alpha^3t^3/3+2\alpha^4t^4/3).$$

$$\tilde{\omega}\left(t/S_2\right) = c_0^2\varphi_0\left(t\right)+c_1^2\varphi_1\left(t\right)+c_2^2\varphi_2\left(t\right)+c_3^2\varphi_3\left(t\right)+c_4^2\varphi_4\left(t\right)=$$

$$=0{,}314\sqrt{2\alpha}\cdot e^{-\alpha t}+0{,}095\sqrt{2\alpha}\cdot e^{-\alpha t}(1-2\alpha t)-0{,}043\sqrt{2\alpha}\cdot e^{-\alpha t}(1-4\alpha t+2\alpha^2t^2)-$$

$$-0{,}12\sqrt{2\alpha}\cdot e^{-\alpha t}(1-6\alpha t+6\alpha^2t^2-4\alpha^3t^3/3)-$$

$$-0{,}154\sqrt{2\alpha}\cdot e^{-\alpha t}(1-8\alpha t+12\alpha^2t^2-16\alpha^3t^3/3+2\alpha^4t^4/3).$$

На четвертом этапе необходимо проверить адекватность модели классификатора на основе текстовых файлов из контрольной выборки. Для этого воспользуемся байесовским правилом классификации, в соответствии с которым должно выполняться условие

$$\frac{\omega\left(t/s_1\right)}{\omega\left(t/s_2\right)} > \frac{P\left(s_2\right)\left(L_{12}-L_{11}\right)}{P\left(s_1\right)\left(L_{21}-L_{22}\right)},$$

для отнесения образа к классу s_1.

Так как в нашем случае исходные данные представляют собой только аппроксимации плотностей распределения вероятностей, предположим (зададимся), что

$$p(s_1) = p(s_2) = \frac{1}{2}, \quad L_{12} = L_{21} = 1, \quad L_{11} = L_{22} = 0.$$

В этом случае решающее правило в силу того, что аппроксимации плотностей распределения вероятностей могут удовлетворять не всем свойствам, предъявляемым к реальным плотностям распределения вероятностей, трансформируется в критерий максимального правдоподобия, который будем использовать в модифицированном виде

$$\omega(t/s_2) \underset{\leq}{\overset{>}{}}^{S_2}_{S_1} \omega(t/s_1).$$

Проверка адекватности осуществляется на основе контрольной выборки путем отнесения файлов из заданного каталога к одной либо другой тематике и последующего подсчета числа правильных и ошибочных решений.

2. Моделирование работы байесовского классификатора на основе аппроксимации плотностей распределения полиномиальными функциями Радемахера-Уолша

Важный частный случай аппроксимации посредством функций возникает при оценке плотностей распределения для двоичных образов.

Пусть имеется образ $\bar{x} = (x_1, ..., x_n)$, где x_k принимает значения 1 или 0. Общее число возможных образов равно $A_k = 2^n$.

В этом случае нет необходимости определять плотность непрерывного распределения. Требуется лишь определить вероятность появления каждого из 2^n возможных векторов образов.

Для оценки дискретного разделения в описанном случае часто используют *полиномиальные функции Радемахера-Уолша*.

Искомое множество базисных функций содержит 2^n членов, полученных перемножением различных членов вида $(2x_k - 1)$ в количестве нуля, одного, двух, трех и т.д. до n, где n – число признаков.

Эти дискретные полиномиальные функции ортогональны относительно весовой функции $u(x) = 1$

$$\sum_x \varphi_j(\bar{x})\varphi_k(\bar{x}) = \begin{cases} 2^n, & \text{если } j = k, \\ 0, & \text{если } j \neq k. \end{cases}$$

Например

j	$\varphi_j(\bar{x})$
1	1
2	$2x_1-1$
3	$2x_2-1$
$\dots$ (n)	$\dots$
$n+1$	$2x_n-1$
$n+2$	$(2x_1-1)(2x_2-1)$
$\dots$	$\dots$
$n+2+n(n-1)/2$	$(2x_{n-1}-1)(2x_n-1)$
$\dots$	$\dots$
2^n	$(2x_1-1)(2x_2-1)\dots(2x_n-1)$

Если в разложении используется только m базисных функций, то приближенное значение плотности дискретного распределения $\omega(\bar{x})$ имеет вид

$$\hat{\omega}(\bar{x}) = \sum_{j=1}^{m} c_j \varphi_j(\bar{x}),$$

где m – количество функций Радемахера-Уолша, участвующих в формировании оценки ω,

c_j – коэффициенты, определяющиеся для каждой функции

$$c_j = \frac{1}{2^n \cdot N} \sum_{i=1}^{N} \varphi_j(\bar{x}_i),$$

где N – количество образов, входящих в класс.

Пример 3.

Пусть существует некоторая случайная величина $\vec{x} = \{x_1, x_2, x_3\}$, представленная выборкой значений

$$\{\vec{x}\} = \begin{matrix} x_1 & x_2 & x_3 \\ \left\{\begin{matrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{matrix}\right. & & \end{matrix} \quad N.$$

$\longleftrightarrow$ p

Необходимо оценить дискретное распределение методом линейной комбинации базисных функций, используя в качестве базисных функций полиномиальные функции Радемахера-Уолша.

В качестве базисных функций воспользуемся следующими многочленами $\varphi_1(\vec{x})=1,\ \varphi_2(\vec{x})=2x_1-1,\ \varphi_3(\vec{x})=2x_2-1,\ \varphi_4(\vec{x})=2x_3-1$.

Вычислим коэффициенты при данных функциях

$$c_1=\frac{1}{2^p\cdot 4}\left[\varphi_1(\vec{x}_1)+\varphi_1(\vec{x}_2)+\varphi_1(\vec{x}_3)+\varphi_1(\vec{x}_4)\right]=\frac{1}{32}(1+1+1+1)=\frac{1}{8},$$

$$c_2=\frac{1}{2^p\cdot 4}\left[\varphi_2(\vec{x}_1)+\varphi_2(\vec{x}_2)+\varphi_2(\vec{x}_3)+\varphi_2(\vec{x}_4)\right]=\frac{1}{32}(-1-1-1-1)=-\frac{1}{8},$$

$$c_3=\frac{1}{2^p\cdot 4}\left[\varphi_3(\vec{x}_1)+\varphi_3(\vec{x}_2)+\varphi_3(\vec{x}_3)+\varphi_3(\vec{x}_4)\right]=\frac{1}{32}(-1-1+1+1)=0,$$

$$c_4=\frac{1}{2^p\cdot 4}\left[\varphi_4(\vec{x}_1)+\varphi_4(\vec{x}_2)+\varphi_4(\vec{x}_3)+\varphi_4(\vec{x}_4)\right]=\frac{1}{32}(-1+1-1+1)=0.$$

Следовательно, дискретное распределение можно аппроксимировать следующим образом

$$\widetilde{f}(\vec{x})=c_1\varphi_1(\vec{x})+c_2\varphi_2(\vec{x})+c_3\varphi_3(\vec{x})+c_4\varphi_4(\vec{x})=\frac{1}{8}-\frac{1}{8}(2x_1-1).$$

Пример 4.

Пусть существует некоторая совокупность комбинаций кода источника $\vec{x}=\{x_1,x_2,x_3,x_4\}$, содержащих искажения и представленная выборкой значений

$$\{\vec{x}\}=\begin{matrix}x_1 & x_2 & x_3 & x_4\end{matrix}\\ \left\{\begin{matrix}0&0&0&0\\0&0&0&1\\0&0&1&0\\0&0&1&1\\0&1&0&0\\0&1&0&1\\0&1&1&0\\1&0&0&0\end{matrix}\right\}\ N.\\ \underset{p}{\longleftrightarrow}$$

Необходимо оценить ПРВ методом линейной комбинации базисных функций, используя в качестве базисных функций полиномиальные функции Радемахера-Уолша.

Решение.

В качестве базисных функций воспользуемся следующими многочленами $\varphi_1(\vec{x})=1,\ \varphi_2(\vec{x})=2x_1-1,\ \varphi_3(\vec{x})=2x_2-1,\ \varphi_4(\vec{x})=2x_3-1,\ \varphi_5(\vec{x})=2x_4-1$.

Вычислим коэффициенты при данных функциях

$$c_1 = \frac{1}{2^p \cdot 8}\left[\varphi_1(\vec{x}_1) + \varphi_1(\vec{x}_2) + \varphi_1(\vec{x}_3) + \ldots + \varphi_1(\vec{x}_8)\right] = \frac{1}{128}(1+1+1+\ldots+1) = \frac{1}{16},$$

$$c_2 = \frac{1}{2^p \cdot 8}\left[\varphi_2(\vec{x}_1) + \varphi_2(\vec{x}_2) + \ldots + \varphi_2(\vec{x}_8)\right] = -\frac{3}{64},$$

$$c_3 = \frac{1}{2^p \cdot 8}\left[\varphi_3(\vec{x}_1) + \varphi_3(\vec{x}_2) + \ldots + \varphi_3(\vec{x}_8)\right] = -\frac{1}{64},$$

$$c_4 = \frac{1}{2^p \cdot 8}\left[\varphi_4(\vec{x}_1) + \varphi_4(\vec{x}_2) + \ldots + \varphi_4(\vec{x}_8)\right] = -\frac{1}{64},$$

$$c_5 = \frac{1}{2^p \cdot 8}\left[\varphi_5(\vec{x}_1) + \varphi_5(\vec{x}_2) + \ldots + \varphi_5(\vec{x}_8)\right] = -\frac{1}{64}.$$

Следовательно, плотность распределения вероятностей можно аппроксимировать следующим образом

$$\tilde{\omega}(\vec{x}) = c_1\varphi_1(\vec{x}) + c_2\varphi_2(\vec{x}) + c_3\varphi_3(\vec{x}) + c_4\varphi_4(\vec{x}) + c_5\varphi_5(\vec{x}) =$$

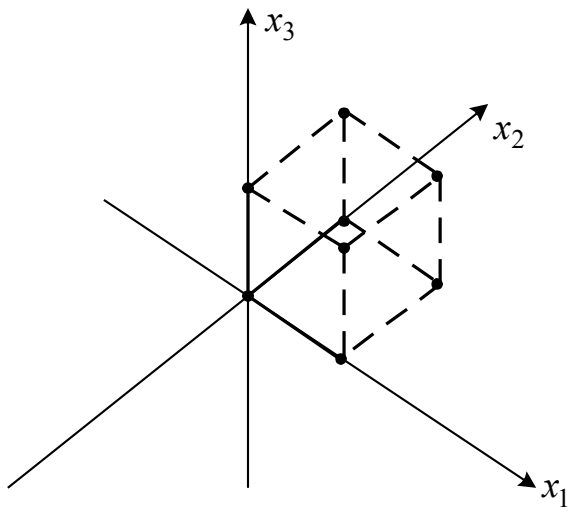
$$= \frac{1}{16} - \frac{3}{64}(2x_1 - 1) - \frac{1}{64}(2x_2 - 1) - \frac{1}{64}(2x_3 - 1) - \frac{1}{64}(2x_4 - 1).$$

Пример 4.

Сформировать байесовский классификатор для двух кодов источника

$$s_1 \begin{Bmatrix} x_1 & x_2 & x_3 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \end{Bmatrix} \quad N \qquad n$$

$$s_2 \begin{Bmatrix} x_1 & x_2 & x_3 \\ 0 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{Bmatrix}$$



Решение.

Решив воспользоваться линейной аппроксимацией плотности распределения, получаем для полиномиальных функций Радемахера-Уолша

$$\varphi_1(\vec{x})=1,\ \varphi_2(\vec{x})=2x_1-1,\ \varphi_3(\vec{x})=2x_2-1,\ \varphi_4(\vec{x})=2x_3-1,$$

где x_i – номер параметра (элемента) в векторе (все x_i принимают значения 0 или 1).

Коэффициенты для класса s_1 выражаются

$$c_{1k}=\frac{1}{2^n\cdot N_1}\sum_{i=1}^{N_1}\varphi_k(\vec{x}_{1i}),$$

где N_1 – количество образов, входящих в класс s_1, $n=3$ – размер вектора.

Проведя суммирование по образам класса ω_1, получим

$$c_{11}=\frac{1}{32}\sum_{i=1}^{4}\varphi_1(x_{1i})=\frac{1}{32}(1+1+1+1)=\frac{1}{8},$$

$$c_{12}=\frac{1}{32}\sum_{i=1}^{4}\varphi_2(x_{1i})=\frac{1}{32}(-1+1+1+1)=\frac{1}{16},$$

$$c_{13}=\frac{1}{32}\sum_{i=1}^{4}\varphi_3(x_{1i})=\frac{1}{32}(-1-1-1+1)=-\frac{1}{16},$$

$$c_{14}=\frac{1}{32}\sum_{i=1}^{4}\varphi_4(x_{1i})=\frac{1}{32}(-1+1-1-1)=-\frac{1}{16}.$$

Применение этой процедуры к классу s_2 дает

$$c_{21}=\frac{1}{8},\ c_{22}=-\frac{1}{16},\ c_{23}=\frac{1}{16},\ c_{24}=\frac{1}{16}.$$

В этом случае аппроксимация плотностей распределения выглядит следующим образом.

$$\hat{\omega}(\bar{x}/s_1)=\sum_{j=1}^{4}c_{1j}\varphi_j(\bar{x})=\frac{1}{8}+\frac{1}{16}(2x_1-1)-\frac{1}{16}(2x_2-1)-\frac{1}{16}(2x_3-1),$$

$$\hat{\omega}(\bar{x}/s_2)=\sum_{j=1}^{4}c_{2j}\varphi_j(\bar{x}).$$

Приняв, что $p(s_1)=p(s_2)=\frac{1}{2}$, $L_{11}=L_{22}=0$ и $L_{12}=L_{21}$, получим решающую функцию

$$\frac{1}{16}(2x_1-1)-\frac{1}{16}(2x_2-1)-\frac{1}{16}(2x_3-1)\geq 0,$$

после умножения на 16 функция принимает вид

$$d(x)=(2x_1-1)-(2x_2-1)-(2x_3-1)\underset{s_2}{\overset{s_1}{\geq}}0.$$

Функция $d(x)$ может иметь только 8 значений, следовательно, в данном случае понятие разделяющей поверхности в том виде, как оно вводилось раньше, не работает.

Задание 1 (факультативно).

Сформировать байесовский классификатор для кодов источника, представленных выборками значений

$$A_1=\begin{matrix}x_1&x_2&x_3&x_4\end{matrix}\begin{Bmatrix}0&0&0&0\\0&0&0&1\\0&0&1&0\\0&0&1&1\\0&1&0&0\\0&1&0&1\\0&1&1&0\\1&0&0&0\end{Bmatrix}\begin{matrix}\\\\\\\\\\\\\\\\\end{matrix}N \qquad A_2=\begin{matrix}x_1&x_2&x_3&x_4\end{matrix}\begin{Bmatrix}1&1&1&1\\1&1&1&0\\1&1&0&1\\1&1&0&0\\1&0&1&1\\1&0&1&0\\1&0&0&1\\0&1&1&1\end{Bmatrix}\begin{matrix}\\\\\\\\\\\\\\\\\end{matrix}N$$

$\longleftrightarrow$ p $\qquad\qquad$ $\longleftrightarrow$ p

Необходимо оценить ПРВ методом линейной комбинации базисных функций, используя в качестве базисных функций различное число полиномиальных функций Радемахера-Уолша (первые четыре и первые пять), оценить влияние числа базисных функций на точность классификации.

Заключительная часть.

Подвожу итоги занятия, анализирую степень достижения цели.

Рекомендованная литература:

1. Хемминг Р.В. Численные методы. – М.: Наука, 1972.

2. Пиотровский Р.Г. Информационные измерения языка. – Л.: Наука, 1968.

3. Левин Б.Р., Шварц В. Вероятностные модели и методы в системах связи и управления. – М.: Радио и связь, 1985.