Министерство образования и науки Российской Федерации

Самарский государственный аэрокосмический университет
имени академика С.П. Королева

КЛАССИФИКАЦИЯ, ОСНОВАННАЯ НА НЕПАРАМЕТРИЧЕСКОМ ОЦЕНИВАНИИ ПЛОТНОСТИ ВЕРОЯТНОСТЕЙ

Методические указания к лабораторной работе № 5
по курсу «МЕТОДЫ РАСПОЗНАВАНИЯ ОБРАЗОВ»

САМАРА
2015

Составители: д.ф.-м..н. В.В.Мясников,

к.т.н. А.В.Кузнецов

УДК 681.3

Классификация, основанная на непараметрическом оценивании плотности вероятностей

Методические указания к лабораторной работе № 5
Самарский государственный аэрокосмический университет
имени академика С.П.Королева
Составители: В.В.Мясников, А.В.Кузнецов
Самара, 2015. 21 с.

В лабораторной работе № 5 по курсу «Методы распознавания образов» изучаются классификаторы, основанные на непараметрических методах оценивания плотности вероятностей.

Методические указания предназначены для студентов специальности 01.02.00 “Прикладная математика и информатика”, обучающихся по специализации «Математическое обеспечение обработки изображений».

Печатается по решению редакционно-издательского совета
Самарского государственного аэрокосмического университета
имени академика С.П.Королева

Рецензент: д.ф.-м.н., профессор А.И.Жданов

Цель работы - изучение теоретических основ и экспериментальное исследование классификаторов, основанных на непараметрических методах оценивания плотности вероятностей, для распознавания образов.

В лабораторной работе изучаются методы построения классификаторов, основанных на непараметрических методах оценивания плотности вероятностей, в частности: оценки Парзена, метода ближайшего соседа и K-ближайших соседей

1. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ ЛАБОРАТОРНОЙ РАБОТЫ

Если известны плотности вероятностей (далее - ПВ) случайного вектора признаков для каждого класса и априорные вероятности классов (при простейшей матрице штрафов), то оптимальной стратегией классификации является использование байесовского классификатора в виде:

$$\forall j \neq l \quad P(\Omega_l) f\left(\bar{x}/\Omega_l\right) > P(\Omega_j) f\left(\bar{x}/\Omega_j\right) \quad \Rightarrow \quad \bar{x} \in D_l . \tag{1}$$

Если соответствующие плотности и вероятности неизвестны, естественный подход состоит в нахождении их оценок $\hat{f}\left(\bar{x}/\Omega_l\right)$, $\hat{P}(\Omega_l)$ на основе имеющейся обучающей выборки объектов из каждого класса и использовании выражения (1) с заменой теоретических величин на эмпирические. Ниже представлены краткие сведения, необходимые для построения и использования классификаторов, использующих два метода непараметрического оценивания ПВ: метод Парзена и метод *K* ближайших соседей.

1.1. Метод Парзена оценивания плотности вероятностей

Пусть $\{\bar{x}_i\}_{i=0}^{N-1}$ - *N* независимых реализаций случайного вектора $\bar{X}$ размерности *n*, распределенного по закону, задаваемому неизвестной ПВ $f(x)$. Метод Парзена получения состоятельных оценок ПВ $f(\bar{x})$ состоит в использовании непрерывных функций вида $\frac{1}{h^n} k\left(\frac{\bar{x}-\bar{x}_i}{h}\right)$, соотнося их с каждым выборочным значением $\bar{x}_i$, при определенных требованиях к *h* и функции *k*. В соответствии с этой идеей приходим к оценке вида

$$\hat{f}_N(\bar{x}) = \frac{1}{N}\sum_{i=0}^{N-1}\frac{1}{h^n}k\left(\frac{\bar{x}-\bar{x}_i}{h}\right), \tag{2}$$

где функция $\frac{1}{h^n}k\left(\frac{\bar{y}}{h}\right)$ называется *ядром оценки*. Оценка (2) используется в случае "круговых" (изотропных) ядер. Для "некруговых" ядер можно использовать следующую формулу (когда $h_1 \neq h_2 \neq \ldots \neq h_N$):

$$\hat{f}_N(\bar{x}) = \frac{1}{N}\cdot\sum_{i=0}^{N-1}\left(\prod_{j=0}^{n-1}h_j\right)^{-1}\cdot k\left[\frac{x_0-\bar{x}_0^i}{h_0},\ldots,\frac{x_n-\bar{x}_{n-1}^i}{h_{n-1}}\right]. \tag{3}$$

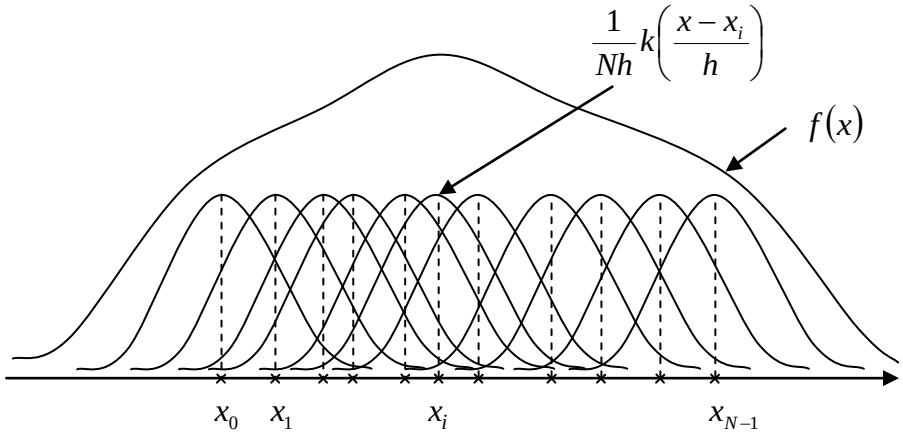


Рисунок 1 - Иллюстрация к оценке Парзена ПВ

Утверждение. Оценка (2) является асимптотически несмещенной и состоятельной оценкой функции ПВ $f(\bar{x})$ в точках ее непрерывности, если функция $k\left(\frac{\bar{y}}{h}\right)$ и число *h* удовлетворяют следующим условиям:

$$\int_{R^n} k\left(\frac{\bar{y}}{h}\right)d\frac{\bar{y}}{h^n} = \int_{R^n} k(z)dz = 1, \tag{4}$$

$$\int\left|k\left(\frac{\bar{y}}{h}\right)\right|d\frac{\bar{y}}{h^n} = \int|k(z)|dz < \infty, \tag{5}$$

$$\sup_{\frac{\bar{y}}{h}\in R^n}\left|k\left(\frac{\bar{y}}{h}\right)\right| = \sup_{z\in R^n}|k(z)| < \infty, \tag{6}$$

$$\lim_{\frac{y}{h}\to\infty}\left|\frac{y}{h}k\left(\frac{y}{h}\right)\right| = \lim_{|z|\to\infty}|zk(z)| = 0, \tag{7}$$

$$\lim_{n\to\infty} h^n(N) = 0 \ (\textit{условие асимптотической несмещенности}) \tag{8}$$

$$\lim_{n\to\infty} N\cdot h^n(N) = \infty \ (\textit{условие состоятельности}) \tag{9}$$

В случае, если дополнительно к условиям (4)-(9) выполняется следующее условие

$$\lim_{n\to\infty} N\cdot h^{2n}(N) = \infty \ (\textit{условия равномерной состоятельности}),$$

то при некоторых дополнительных ограничениях оценка (2) является *равномерно состоятельной:*

$$\forall\varepsilon>0 \quad \lim_{N\to\infty}\cdot P\left\{\sup_{-\infty<x<\infty}\left|\hat{f}_N(x) - f(x)\right| > \varepsilon\right\} = 0.$$

Примером многомерного ядра, имеющего вид нормальной ПВ. Это ядро определяется следующим образом:

$$\frac{1}{h^n}\cdot k\cdot\left(\frac{\bar{x}-\bar{x}_i}{h}\right) = (2\cdot\pi)^{-n/2}\cdot h^{-n}\cdot|B|^{-1/2}\cdot\exp\left[-\frac{1}{2}\cdot h^{-2}\cdot(\bar{x}-\bar{x}_i)'\cdot B^{-1}\cdot(\bar{x}-\bar{x}_i)\right], \tag{10}$$

здесь, очевидно:

$$(2\cdot\pi)^{-n/2}\cdot h^{-n}\cdot|B|^{-1/2} = \frac{1}{(2\cdot\pi)^{n/2}\cdot\sqrt{|B|\cdot h^n}}.$$

Примечание. Пусть в выражении (10) в качестве матрицы B использована выборочная ковариационная матрица $\hat{B}$, рассчитанная по выборке. Тогда ковариационная матрица оценки Парзена B_p и матрица $\hat{B}$ связаны соотношением:

$$B_p = \frac{1}{N}\cdot\sum_{i=0}^{N-1}\left(\bar{x}_i\cdot\bar{x}_i' + h^2\cdot\hat{B}\right) \approx \left(1+h^2\right)\cdot\hat{B}.$$

Параметр $h(N)$ удобно использовать в виде:

$$h(N) = N^{-\frac{k}{n}} = \left(N^{-\frac{k}{n}}\right). \tag{11}$$

Для выполнения условий (8) – (10) величина k должна удовлетворять следующим ограничениям:

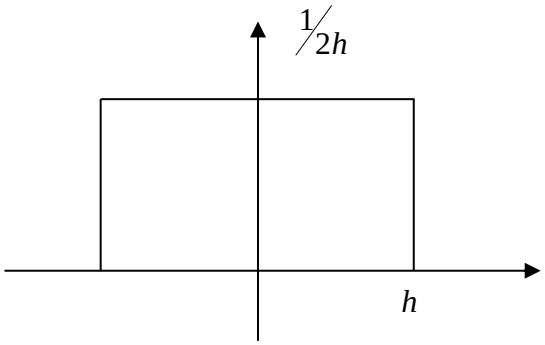
$1 > k > 0$ - условие состоятельности,

$\frac{1}{2} > k > 0$ - условие равномерной состоятельности.

Примеры одномерных ядер, используемых в оценке Парзена, приведены ниже.

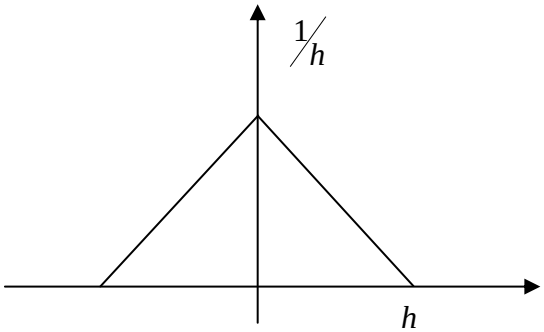
Прямоугольное ядро

$$\frac{1}{h}k\left(\frac{y}{h}\right) = \begin{cases} \frac{1}{2h}, & \left|\frac{y}{h}\right| \le 1, \\ 0, & > 1. \end{cases}$$



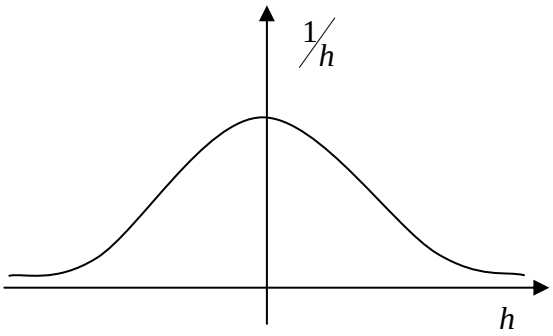
Треугольное ядро

$$\frac{1}{h}k\left(\frac{y}{h}\right) = \begin{cases} \frac{1}{h}\left(1 - \left|\frac{y}{h}\right|\right), & \left|\frac{y}{h}\right| \le 1, \\ 0, & > 1. \end{cases}$$



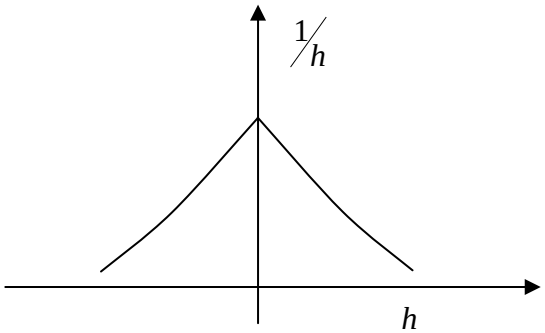
Нормальное ядро

$$\frac{1}{h}k\left(\frac{y}{h}\right)=\frac{1}{\sqrt{2\pi h}}\exp\left(-\frac{y^2}{2h^2}\right),$$



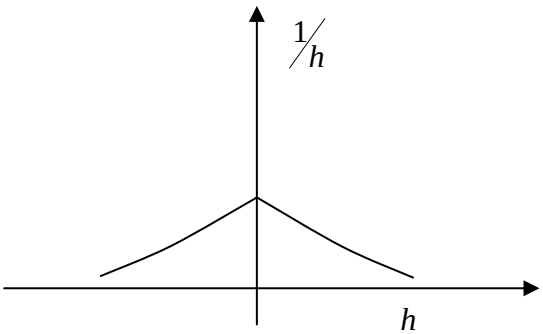
Экспоненциальное ядро:

$$\frac{1}{2h}\exp\left(-\frac{y}{h}\right),$$



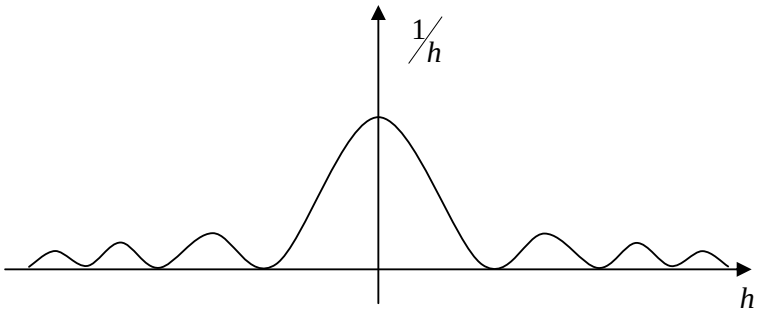
Рациональное ядро

$$\frac{1}{\pi h}\frac{1}{1+\left(\frac{y}{h}\right)^2},$$



Ядро "sinc"

$$\frac{1}{\pi h}\left(\frac{\sin\frac{y}{2h}}{\frac{y}{2h}}\right)^2,$$



Достоинства и недостатки оценки Парзена

К достоинствам оценки следует отнести ее универсальность. Недостатками оценки являются ее высокая вычислительная сложность (число слагаемых в выражении (2) соответствует объему обучающей выборки); необходимость хранения всей обучающей выборки; неоднозначность оценки, связанной с произвольностью выбора ядер и величины $h = h(N)$.

1.2. Метод К ближайших соседей

Используя выборку $\{\bar{x}_i\}_{i=0}^{N-1}$ объема N, найдем расстояние r от точки $\bar{x}$ до *K*-го ближайшего к $\bar{x}$ элемента выборки (*K-го ближайшего соседа*). Для измерения "близости" можно воспользоваться любой подходящей меркой. Тогда в качестве оценки ПВ в точке $\bar{x}$ можно принять:

$$\hat{f}_N(\bar{x}) = \frac{K}{N}\cdot\frac{1}{A(k,N,\bar{x})}, \qquad (12)$$

где $A(k,N,\bar{x})$ - объем множества всех точек, расстояние которых до $\bar{x}$ меньше, чем r. Величина A является случайной (и зависит фактически от $\bar{x}$), зависящей от выбранного множества N элементов выборки.

Утверждение. Если параметр $K(N)$, удовлетворяет условиям:

$$\lim_{W\to\infty} K(N) = \infty,$$

$$\lim_{N\to\infty} \frac{K(N)}{N} = 0,$$

то *оценка (12) является асимптотически несмещенной и состоятельной оценкой ПВ* $f(\bar{x})$ в точках непрерывности ПВ.

1.2.1 Решающее правило K ближайших соседей

Полученная оценка $\hat{f}_N(\bar{x})$ ПВ может использоваться следующим образом. Когда требуется классифицировать неизвестный объект (вектор признаков) $\bar{x}$, среди имеющихся N объектов (уже проклассифицированных), из которых N_l объектов из класса Ω_l $\left(\sum_{l=0}^{L-1} N_l = N\right)$, находят K ближайших к точке $\bar{x}$ объектов. Пусть k_l $\left(l = \overline{0, L-1}\right)$ - число объектов из класса Ω_l среди этих K ближайших соседей $\left(\sum_{l=0}^{L-1} K_l = K\right)$. Тогда оценка ПВ принимает вид (в *l*-ом классе):

$$\hat{f}_{N_i}^{\,i}\left(\bar{x}/\Omega_l\right) = \frac{K_l}{N_l} \cdot \frac{1}{V},\ l = \overline{0, L-1}. \tag{13}$$

Поскольку независимо от номера класса, K_l объектов извлечены из одной и той же области A, то величина объема V оказывается одинаковой для всех классов. Следовательно, байесовский классификатор (1), минимизирующий общий риск, будет иметь вид (оценку априорной вероятности берем в виде (15): $\hat{P}(\Omega_l) = \frac{N_l}{N}$):

$$\forall\, j \neq l \quad \frac{N_l}{N} \cdot \hat{f}_N\left(\bar{x}/\Omega_l\right) \geq \frac{N_j}{N} \cdot \hat{f}_N\left(\bar{x}/\Omega_j\right) \Rightarrow \bar{x} \in D_l.$$

Подставляя в это выражение значение из (13), получим:

$$\forall\, j \neq l \quad \frac{N_l}{N} \cdot \frac{K_l}{N_l \cdot V} \geq \frac{N_j}{N} \cdot \frac{K_j}{N_j \cdot V} \Rightarrow \bar{x} \in D_l$$

или

$$\forall\, j \neq l \quad K_l \geq K_j \Rightarrow \bar{x} \in D_l. \tag{14}$$

Окончательно, решающее правило можно записать в виде:

$$l = \arg \max_{j=0,L-1} K_l, \quad \bar{x} \in D_l.$$

1.2.2 Решающее правило K ближайших соседей для двух классов

Из соотношения (14) непосредственно следует:

$$K_0 \underset{<}{>} K_1 \Rightarrow \bar{x} \in \begin{cases} D_0 \\ D_1 \end{cases}$$

Т.о. решение о принадлежности вектора $\bar{x}$ принимается в тот класс, соседей которого больше. Для устранения возможных проблем с «равенством» числа соседей, число *K в случае двух классов следует брать нечетным*.

1.2.3 Решающее правило ближайшего соседа

В случае K=1 решающее правило называют *правилом ближайшего соседа*. В соответствие с этим правилом решение о принадлежности вектора $\bar{x}$ принимается в тот класс, у которого оказался ближайший к $\bar{x}$ сосед.

Достоинства и недостатки решающего правила К соседей

Достоинство: решающее правило, основанное на методе K ближайших соседей, является очень простым и не требует знания (явного построения) ПВ (или ее оценки).

Его *недостаток* заключается в необходимости хранить в памяти машины все объемы (всю обучающую выборку) и сравнивать каждый из них с неизвестным объектом.

1.3 Оценка априорных вероятностей

Оценка величин $P(\Omega_l)$ для больших объемов обучающей выборки может быть выполнена следующим образом:

$$\hat{P}(\Omega_l) = \frac{N}{N_l}, \; N = N_0 + \ldots + N_{L-1}, \qquad (15)$$

Здесь:

- N – общее количество элементов обучающей выборки во всех классах;
- N_l - количество элементов обучающей выборки в l-ом классе.

3. ПОРЯДОК ВЫПОЛНЕНИЯ ЛАБОРАТОРНОЙ РАБОТЫ

3.1. Исходные данные

- параметры нормальных распределений для двух классов с равными и неравными ковариационными матрицами и наборы реализаций двумерных случайных векторов для этих классов (из лаб. работы № 1);
- исполняемый в системе Python файл, необходимый для выполнения лабораторной работы (предоставляется преподавателем).

3.2. Общий план выполнения работы

1. Синтезировать дополнительные реализации для каждого из классов (использовать исполняемый файл из л.р. № 1).
2. Построить классификатор, основанный на непараметрической оценки Парзена, используя сгенерированные в п.1 данные как обучающие выборки, а данные из первой лабораторной работы - как тестовые. В качестве ядра взять гауссовское (10), величину h взять в виде (11). Оценить эмпирический риск - оценку суммарной вероятности ошибочной классификации.
3. Построить классификатор, основанный на методе К ближайших соседей (для K=1,3,5), используя сгенерированные в п.1 данные как обучающие выборки, а данные из первой лабораторной работы - как тестовые. Оценить эмпирический риск - оценку суммарной вероятности ошибочной классификации.
4. Сравнить полученные в пп.2-3 классификаторы и качество их работы с байесовским классификатором из л.р.№2.

3.3. Содержание отчета

Отчет по работе должен содержать:

1. Графическая иллюстрация данных, полученных в результате выполнения п.1 плана и результатов их классификации из пп.2-3.
2. Вероятности ошибочной классификации построенных в пп.2-3 плана классификаторов, найденные экспериментально. Результаты сравнения построенных классификаторов с байесовским классификатором.

4. КОНТРОЛЬНЫЕ ВОПРОСЫ

СОДЕРЖАНИЕ

Учебное издание

КЛАССИФИКАЦИЯ, ОСНОВАННАЯ НА НЕПАРАМЕТРИЧЕСКОМ ОЦЕНИВАНИИ ПЛОТНОСТИ ВЕРОЯТНОСТЕЙ

Методические указания к лабораторной работе № 5

по курсу «Методы распознавания образов»

Составители: Мясников Владислав Валерьевич

Кузнецов Андрей Владимирович

Самарский государственный аэрокосмический университет

имени академика С.П.Королева

443086, Самара, Московское шоссе, 34

Отпечатано на кафедре геоинформатики СГАУ

Тираж 20 экз.