§ 2. ОЦЕНКА «СКОЛЬЗЯЩИЙ КОНТРОЛЬ»

Оценим качество решающего правила $F(x, a_э)$ минимизирующего на заданной последовательности

$$x_1, y_1; ...; x_l, y_l$$

эмпирический риск

$$I_э(a) = \frac{1}{l-1} \sum_{i=1}^{l-1} (y_i - F(x_i, a))^2$$

с помощью. следующего приема. Исключим из обучающей последовательности первую пару x_1, y_1 и найдем функцию, минимизирующую эмпирический риск на оставшихся $l-1$ элементах обучающей последовательности. Пусть эта функция есть $F(x, a(x_1', y_1'; ...; x_l, y_l))$. Здесь знак x_1', y_1' указывает на то, что из обучающей последовательности была исключена пара x_i, y_i. Вычислим величину уклонения на исключенной паре:

$$(y_1 - F(x_1; a(x_1', y_1'; ...; x_i, y_i)))^2 .$$

Затем из обучающей последовательности исключим вторую пару (первая пара остается в последовательности) и вычислим уклонение

$$(y_2 - F(x_2; a(x_1, y_1; x_2', y_2'; ...; x_l, y_l)))^2 .$$

Так вычислим уклонение для всех l пар. Образуем величину

$$T(x_1, y_1; ..; x_l, y_l) = \\ = \frac{1}{l} \sum_{i=1}^{l} (y_i - F(x_i; a(x_1, y_1; ...; x_i', y_i'; ...; x_l, y_l)))^2$$

и примем ее за оценку качества функции $F(x, a_э^{l-1})$, минимизирующей эмпирический риск на выборке объема $l-1$:

$$I(a_э^{l-1}) = \int (y - F(x, a_э))^2 P(x, y,) dxdy \infty T(x_1, y_1; ...; x_l, y_l) .$$

Такую процедуру оценивания назовем: - *«скользящий контроль»*.

Справедлива теорема:

Теорема 4.I. *Оценка «скользящий контроль» является несмещенной, т. е.*

$$\mathrm{M} I(a_э(x_1, y_1; ...; x_{l-1}, y_{l-1})) = \mathrm{M} T(x_1, y_1; ...; x_l, y_l) .$$

З а м е ч а н и е. Процедура «скользящий контроль» определяет несмещенные оценки качества как при восстановлении индикаторных функций, так и при восстановлении произвольных функциональных зависимостей.

Свойство несмещенности, однако, недостаточно характеризует оценку. Необходимо знать еще и дисперсию $D(T)$ оценки T. Если бы дисперсия оценки T была известна, то можно было бы оценить качество решающего правила, минимизирующего эмпирический риск на выборках длины l. А именно, с вероятностью $1-h$ справедливо неравенство

$$\mathrm{M}I\left(a_{э}\left(x_1,y_1;\ldots;x_{l-1},y_{l-1}\right)\right)\leq T\left(x_1,y_1;\ldots;x_l,y_l\right)+\sqrt{\frac{D(T)}{h}} \qquad (4.11)$$

Где $1-h$ - надежность, с которой требуется выполнение неравенства (4.11). (Оценка (4.11) следует из утверждения теоремы 4.1 и неравенства Чебышева $P\left\{\left|\mathrm{M}T-T\right|>\sqrt{\frac{D(T)}{h}}\right\}<h$.

Однако вычислить дисперсию оценки «скользящий контроль» в достаточно общей постановке не удается. Для частного случая, когда $F(x,a)=\sum_{i=1}^{n}a_i x^i+a_0$, вектор x распределен но нормальному закону, $y=F(x,a)+x$ и x распределено по нормальному закону, можно доказать, что дисперсия оценки мала, имеет порядок $\frac{1}{l}$, а не $\frac{n}{l}$, и при $j>2n$ слабо зависит от размерности вектора x. Этот результат легко будет следовать из теоремы 4.3, приведенной в следующем параграфе.

Мы предполагаем, что такая же особенность оценки дисперсии сохранится и в общем случае: для задачи распознавания образов — когда класс $F(x,a)$ имеет ограниченную емкость; а для восстановления регрессии — когда класс $F(x,a)$ имеет ограниченную емкость и выполняется неравенство

$$\sup_{a}\frac{\mathrm{M}\left(y-F(x,a)\right)^4}{\left(\mathrm{M}\left(y-F(x,a)\right)^2\right)^2}=t^2<\infty .$$

§ 3. ЭКВИВАЛЕНТНОЕ ПРЕДСТАВЛЕНИЕ ОЦЕНКИ «СКОЛЬЗЯЩИЙ КОНТРОЛЬ». $J_{э}\left(x_1,y_1;\ldots;x_l,y_l\right)$-СТАТИСТИКИ

Рассмотрим теперь задачу восстановления функции регрессии в классе линейных по параметрам зависимостей

$$F(x,a)=\sum_{i=1}^{n}a_i j_i(x).$$

Оказывается, что в этом классе функций оценка «скользящий контроль» допускает следующее эквивалентное представление:

$$T\left(x_1,y_1;\ldots;x_l,y_l\right)=\frac{1}{l}\sum_{i=1}^{l}\frac{\left(y_i-f_i^T\left(\Phi_l^T\Phi_l\right)^{-1}\Phi_l^T Y_l\right)^2}{\left(1-f_i^T\left(\Phi_l^T\Phi_l\right)^{-1}f_i\right)^2}, \qquad (4.12)$$

где обозначено

$$\Phi_l = \begin{Vmatrix} j_1(x_1) & \dots & j_n(x_1) \\ \dots & \dots & \dots \\ j_1(x_l) & \dots & j_n(x_l) \end{Vmatrix},$$

f_i — i-я строка матрицы Φ_l, а Y_l есть l- мерный вектор-столбец значений y:

$$Y_l = (y_1, \dots, y_l)^T .$$

Выражение $\left(\Phi_l^T \Phi_l\right)^{-1} \Phi_l^T Y$ в числителе (4.12) есть оценка вектора параметров a, полученная методом наименьших квадратов по всей обучающей последовательности. Числитель в (4.12) определяет квадрат уклонения в точке x_i, а знаменатель определяет ту мультипликативную поправку, которая возникает, если оценку параметров a получать не по всей обучающей выборке, а по выборке, из которой исключена i-я пара x_i, y_i.

Представление (4 12) замечательно тем, что в нем используется лишь одно обращение матрицы, а не l, как в общей процедуре, описанной в предыдущем параграфе. Это обстоятельство делает процедуру «скользящий контроль» в вычислительном отношении немногим более трудоемкой, чем вычисление невязки в методе наименьших квадратов.

При построении алгоритмов восстановления зависимостей часто имеет смысл искать решение, доставляющее не только безусловный минимум функционалу $I_э(a)$, но и условный минимум при ограничении на решение [1)]:

$$|a|^2 = \sum a_i^2 \le C .$$

Отыскание такого условного минимума — задача, эквивалентная поиску. безусловного минимума функционала

$$I_э(a, g) = \frac{1}{l} \sum_{i=1}^{l} \left(y_i - F(x_i, a)\right)^2 + g|a|^2 , \qquad (4.13)$$

где g — положительная константа, зависящая от C (множитель Лагранжа).

Оценку качества решения $a = a_э$, минимизирующего функционал (4.13), будем проводить также, с помощью процедуры «скользящий контроль».

Найдем решения $a_g(x_i, y_i; \dots; x_i', y_i'; \dots; x_l, y_l)$, минимизирующие функционал (4.13), заданный на $l-1$ паре (пара x_i, y_i выпущена, g фиксировано), и образуем величину

$$T_g(x_1, y_1; \dots; x_l, y_l) =$$
$$= \frac{1}{l} \sum_{i=1}^{l} \left(y_i - F\left(x, a_g(x_1, y_1; \dots; x_i', y_i'; \dots; x_l, y_l)\right)\right)^2 .$$

Величина T_g, и будет оценкой качества функции $F(x, a_g)$, минимизирующей функционал (4.13).

Эквивалентное представление T_g, задается с помощью матрицы

[1)] Такое решение иногда называют ридж-оценкой.

$$A_g = \left(\Phi_l^T \Phi_l + g\mathrm{I}\right),\ \mathrm{I} = \begin{vmatrix} 1 & & & 0 \\ & 1 & & \\ & & \mathbf{O} & \\ 0 & & & 1 \end{vmatrix}.$$

А именно,

$$T_g\left(x_1, y_1; \ldots; x_l, y_l\right) = \frac{1}{l}\sum_{i=1}^{l} \frac{\left(y_i - f_i^T A_g^{-1} \Phi_l^T Y_l\right)^2}{\left(1 - f_i^T A_g^{-1} f_i\right)^2} \tag{4.14}$$

при. $g = 0$ выражение (4.14) совпадает с (4.12).

Введем теперь параметрическое (по параметру p) семейство статистик

$$J_p\left(x_1, y_1; \ldots; x_l, y_l\right) = \\ = \frac{\left(n + \dfrac{1}{l+p} + \dfrac{n}{l+p-n-2}\right)}{\left(1 + \dfrac{1}{l-1} + \dfrac{n}{l-n-3}\right)} \frac{1}{l}\sum_{i=1}^{l} \frac{\left(y_i - f_i^T\left(\Phi_l^T \Phi_l\right)^{-1} \Phi_l^T Y\right)^2}{\left(1 - f_i^T\left(\Phi_l^T \Phi_l\right)^{-1} f_i\right)^2}$$

$$l > n+7, \quad p > -l+n+2$$

содержащее (при $p = -1$) статистику «скользящий контроль». С помощью этого семейства статистик для задачи восстановления регрессии

$$y = \sum_{i=1}^{n} a_i^0 j_i(x) + a_0 ,$$

которая решается с помощью метода наименьших квадратов, будем по выборке $x_1, y_1; \ldots; x_l, y_l$ объема l оценивать величины I_{l+p} среднего риска для решений, полученных по выборкам объема $l+p$:

$$I_{l+p} = \mathrm{M}\left(y - f^T\left(\Phi_{l+p}^T \Phi_{l+p}\right)^{-1} \Phi_{l+p}^T Y_{l+p}\right)^2.$$

Здесь математическое ожидание вычисляется по матрицам Φ_{l+p} (размера $(l+p)\times n$), вектору Y_{l+p} (размерности $1+p$), величине y и вектору f (размерности n).

Итак, примем

$$J_p\left(x_1, y_1; \ldots; x_l, y_l\right) \infty I_{l+p}.$$

Оказывается, что в условиях, когда вектор f распределен по нормальному закону, а величина y задается моделью

$$y_k = \sum_{i=1}^{n} a_i^0 j_i\left(x_k\right) + a_0 + x_k$$

где помеха x не зависит от x и распределена по нормальному закону $N(0,s^2)$ (параметр s^2 нам не известен), справедливы следующие три теоремы [1)],

Теорема 4.2. *Статистика* $J_p(x_1,y_1;...;x_l,y_l)$ *является несмещенной оценкой риска* I_{l+p}.

Теорема 4.3. *Дисперсия оценки* $J_p(x_{1,y_1;...;x_l,y_l})$ *заключена в пределах*

$$\frac{2s^4}{l}\left(\frac{1+\frac{1}{l+p}+\frac{n}{l+p-n-2}}{1+\frac{1}{l-1}+\frac{n}{l-n-3}}\right)^2 A \le D(J_p(.)) \le$$

$$\le \frac{2s^4}{l}\left(\frac{1+\frac{1}{l+p}+\frac{n}{l+p-n-2}}{1+\frac{1}{l-1}+\frac{n}{l-n-3}}\right)^2 B,$$

где

$$A=\left(1+\frac{1}{l-1}+\frac{n}{l-n-3}\right)^3-\frac{nl^2}{(l-n-5)^3},$$

$$B=\left(1+\frac{1}{l-1}+\frac{n+4}{l-n-7}\right)^3+\frac{nl^2}{(l-n-5)^3}.$$

Теорема 4.4. *Среднеквадратичная относительная ошибка оценивания риска* I_{l+p}, *не зависит от* p *и лежит в пределах:*

$$\frac{\sqrt{\frac{2A}{l}}}{1+\frac{1}{l-1}+\frac{n}{l-n-3}} \le \frac{\sqrt{D(J_p(.))}}{I_{l+p}} \le \frac{\sqrt{\frac{2B}{l}}}{1+\frac{1}{l-1}+\frac{n}{l-n-3}}$$

З а м е т и м, что при $l>2n$ оценки, полученные в теоремах 4.2 и 4.3, слабо зависят от n - размерности вектора f.

Помимо оценки ожидаемого риска, на выборках объема $l+p$, статистики $J_p(x_1,y_1;...;x_l,y_l)$ позволяют:

1. Оценить минимальную возможную величину риска. Минимально возможный риск оценивается величиной $J_\infty(x_1,y_1;...;x_l,y_l)$.

2. Оценить, на сколько следует ожидать уменьшения риска, если выборка будет увеличена на p элементов. Оценка уменьшения риска задается величиной

$$\Delta_p = J_0(x_1,y_1;...;x_l,y_l)-J_p(x_1,y_1;...;x_l,y_l).$$

3. Оценить, на сколько элементов должна быть увеличена выборка, чтобы риск уменьшился на величину Δ. Величина выборки определяется минимальным решением неравенства

$$J_0(x_1,y_1;...;x_l,y_l)-J_p(x_1,y_1;...;x_l,y_l)\ge\Delta.$$

[1)] Доказательства этих теорем даны в английском издании [3]: V. N. Vapnik. Estimation of Dependences Based on Emyirieal Data.- Springer, New York, 1982.

4. Наконец, можно попытаться использовать статистику $J_p(x_1, y_1;...; x_l, y_l)$ для определения подпространства векторного пространства X , на. котором регрессии, построенные на выборках объема $l+p$, будут иметь минимальный средний риск. Для этого надо найти подпространство $\tilde{X}$ пространства X , на котором достигается минимум $J_p(\tilde{x}_1, y_1;...;\tilde{x}_l, y_l)$ (для различных p получаются разные оптимальные подпространства).