ПРИНЦИП ОККАМА. АППРОКСИМАЦИЯ ДИСКРИМИНАНТНОЙ ФУНКЦИИ В ОКРЕСТНОСТИ НУЛЕВЫХ ЗНАЧЕНИЙ

к. т. н. Зенков В. В.

Россия, Москва, Институт проблем управления им. В.А. Трапезникова РАН

Abstract. *The method for approximation of the discriminant function by the cylindrical surface in the classification tasks and pattern recognitions is described. The main idea of the method is connected with successive approximations to zero of the discriminant function (the guideline of cylinder) using the weighted least squares method. The method improves the efficiency of decision rules. Some model examples and one example with real data for the diagnosis of cervical cancer are presented.*

Keywords: *classification, pattern recognition, approximation of the discriminant function, weighted least squares method, diagnosis of cervical cancer.*

1. Введение. При решении задачи классификации или распознавания образов с учителем в стохастической постановке используются дискриминантные функции (ДФ), по знакам которых объект относится к одному из классов. В качестве критерия качества метода обычно используется минимум средней стоимости ошибки классификации. Для достижения минимума критерия используется косвенный путь – путь аппроксимации ДФ [1].

В работе обращается внимание на то, что ДФ со всеми ее изящными изгибами аппроксимировать не нужно. От аппроксимации ДФ (АДФ) требуется максимальное приближение к ДФ только в области её нулевых значений, на границе между двумя классами. С учетом этого обстоятельства в работе вместо обычного МНК для построения АДФ использована последовательность решений с помощью взвешенного МНК, обеспечивающая приближение АДФ к ДФ в окрестности нулевых значений последней. В результате повышается эффективность простых АДФ и реализуется принцип Оккама: "Не множить сущее без необходимости" и KISS-принцип: "Keep It Simple, Stupid" – "Не усложняй, глупый".

2. Постановка задачи. Ограничимся вероятностной постановкой задачи обучения [1,2,3,4,5].

Имеется обучающая выборка, состоящая из случайных и независимых точек (x_n, k_n), n = (1, … , N), N – количество строк выборки; k_n – номер класса в строке n, $k_n \in \{1,\dots,K\}$, K-количество классов, x_n – вектор действительных значений признаков, $x_n \in R^d$, d – размерность пространства признаков.

Априорные вероятности классов P_k и плотности условных распределений p(x|k) не известны.

Задана матрица стоимостей ошибок классификации с положительными недиагональными элементами Crs – стоимость ошибки, когда точка из класса s относится к классу r, и нулевыми диагональными. Индексы изменяются от 1 до K.

Необходимо в некотором множестве решающих правил найти по выборке такое решающее правило, зависящее от векторного параметра λ, $F(\lambda) : X \rightarrow \{1,\dots,K\}$, которое минимизирует среднюю по выборке стоимость потерь от ошибок классификации

$$G(\lambda) = 1/N \sum_r \sum_{s\#r} C_{rs} N_{rs}(\lambda), \quad (1)$$

где N_{rs} – количество точек выборки из класса s, ошибочно отнесенных решающим правилом F(λ) в класс r.

2.1Два свойства ДФ

1. ДФ, $f_{rs}(x)$, разделяющая классы r и s в пространстве признаков, по определению есть функция регрессии [2]

$$f_{rs}(x) = G_r(x) - G_s(x) = \sum_k C_{rk}\, p(k|x) - \sum_k C_{sk} p(k|x) = M_{k|x}(C_{rk} - C_{sk}), \quad (2)$$

где $G_r(x)$, $G_s(x)$ - средние потери, если точку x отнести к классу r или к классу s ; $p(k|x)$ – апостериорная вероятность класса k в точке x,

$p(k|x) = P_k\, p(x|k) / p(x)$, $p(x) = \sum_r P_r\, p(x|r)$, P_k – априорные вероятности классов, $p(x|k)$ - априорные условные распределения признаков классов. Таким образом, в точке x случайная по k дискретная величина

$$C_{rk|x} - C_{sk|x} = f_{rs}(x) + \varepsilon_{rs} \quad (3)$$

имеет распределение $p(k|x)$, $\sum_k p(k|x) = 1$, среднее $f_{rs}(x)$ и случайное отклонение от него ε_{rs}

Если $f_{rs}(x) \le 0$, то точку x нужно отнести в класс r, иначе – в класс s. Так обеспечивается минимум средних потерь от ошибок классификации.

Обучающая выборка состоящая из строк (k, x), где x – вектор-строка значений признаков; k –номер класса, затем заменяется разностью стоимостей ошибок (3), соответствующих разделяемой паре классов, чтобы получить типичную выборку задачи регрессионного анализа.

Для большого количества признаков в регрессионном анализе существуют методы выбора подходящего состава признаков по коэффициентам корреляции признаков с искомой величиной [6].

2. При K= 2 апостериорная вероятность и ДФ связаны тождеством

$$p(1|x) = (C_{12} - f_{12}(x)) / (C_{12} + C_{21}), \quad (4)$$

которое следует из определения ДФ (2) и тождества $p(1|x) + p(2|x) = 1$.

Из (4), в частности, следует, что в точках на границе классов, где ДФ $f_{12}(x) = 0$, апостериорная вероятность первого класса p* и стоимости ошибок, с учетом которых построена ДФ, связаны отношениями

$$p^* = C_{12} / (C_{12} + C_{21}), \quad 1 - p^* = C_{21} / (C_{12} + C_{21}), \quad C_{12} / C_{21} = p^* / (1 - p^*), \quad (5)$$

а в точках, относимых решающим правилом в первый класс, т.е. в точках, где $f_{12}(x) < 0$, из (4) следует, что $p(1|x) > p^*$.

Поскольку на величину p* влияет лишь отношение стоимостей ошибок, то при условии $C_{21} = 1$ тождество (4) упрощается

$$p(1|x) = p^* - (1 - p^*) f_{12}(x). \quad (6)$$

2.2 Методы решения задач классификации в вероятностной постановке.

По обучающей выборке отметим следующие методы решения задачи классификации в вероятностной постановке:

1) Параметрические и непараметрические методы построения по выборке распределений классов $p(x|k)$, по которым, пользуясь формулой Байеса, находятся АДФ. Результат решения задачи зависит от достоверности предположений о виде априорных распределений классов. Достоинство этих методов в простоте алгоритмов решения задачи и возможности не только строить границы классов, но и оценивать апостериорные вероятности $p(k|x)$.

2) Методы, при K=2 минимизирующие критерий качества (1). Точкам одного класса приписываются значения 1, а другого -1 или 0. Ищется функция от x, знаки которой в точках выборки максимально совпадают с этими значениями. Эти методы решают задачу классификации при равных стоимостях ошибок, аппроксимируя ДФ, но связь их с байесовской классификацией не отмечается за исключением [1], где по выборке для K≥2 и для разных стоимостей ошибок целенаправленно решается байесовская задача классификации путем аппроксимации ДФ.

При этом отсутствует стремление аппроксимировать ДФ в области её нулевых значений.

3) Логистическая регрессия [5,7] – аппроксимация апостериорной вероятности $p(r|x)$

сигмоидной функцией. Сигмоидная функция удовлетворяет естественным ограничениям вероятностей – её значения лежат в интервале (0,1). Применяется, как правило, для случая K=2. Позволяет оценивать вероятности классов в разделяемых точках. Параметры сигмоидной функции находятся по выборке по критерию максимального правдоподобия с применением градиентных методов.

Недостатки – зависимость от вида условных распределений классов и недостатки градиентных способов поиска экстремума.

4) Используемый в данной работе метод, основанный на аппроксимации ДФ в окрестности её нулевых значений [8]. Приведенные примеры показывают улучшение классификации этим методом. ДФ относится к более сложному классу, чем поверхность, полученная приравниванием ее к нулю. ДФ $f_{12}(x)$ при K=2 в случае нормальных априорных распределений содержит экспоненту, в то время как уравнение $f_{12}(x) = 0$ имеет эквивалентное уравнение не выше второго порядка.

3. Способ решения задачи.

В качестве АДФ используется функция вида $\lambda'\varphi(x)$, где λ – неизвестный вектор, а $\varphi(x)$ – заданная вектор – функция. Поскольку ДФ не известна, то используется итерационный процесс для приближения АДФ к области нулей ДФ, на первом шаге которого находится λ_1 обычным МНК, а на каждом последующем шаге i находится λ_i минимизацией критерия взвешенного МНК при известном, вычисленном на предыдущем шаге, векторе λ_{i-1}

$$Q(\lambda_i) = \frac{1}{N} \Sigma_n \; [\, C_{rk_n} - C_{sk_n} - \lambda_i' \, \varphi(x_n)]^2 \exp(-W \mid \lambda_{i-1}' \varphi(x_n) \mid), \qquad (7)$$

решением системы линейных уравнений относительно λ_i, где W – заданный коэффициент весовой функции, W>0, i – номер итерации, i = 2,…, I, I – заданное количество итераций.

Весовая функция придает бОльший вес тем точкам, которые ближе находятся к нулевым значениям предыдущей АДФ. Мерой близости являются сами значения предыдущей АДФ. Для большей корректности этой меры близости на каждой итерации выполняется нормализация вектора λ. Последнее, однако, не имеет отношения к регуляризации решения задачи. Регуляризация, как средство борьбы с переучиванием алгоритма, здесь не рассматривается.

Геометрически любая ДФ или АДФ - это цилиндр, где $\lambda'\varphi(x_n) = 0$ – это поверхности, играющие роль направляющих. Образующие цилиндров – это прямые, параллельные той оси, по которой откладываются значения ДФ или АДФ. Внутри цилиндров – точки одного класса, $f(x) \le 0$ или $\lambda'\varphi(x_n) \le 0$, вне- другого.

Вопросы выбора весового коэффициента W и вида весовой функции (в (7) –экспонента от модуля аргумента, в [8] – экспонента от квадрата аргумента) - являются отдельной проблемой.

В качестве лучшего значения λ выбирается тот вектор, которому соответствуют наименьшие потери (1) по итерациям.

В приведенных ниже примерах величина W не была постоянной. Начальная точка графика дает потери обычного МНК. Далее при удачном выборе W потери должны уменьшаться, затем с ростом W может начаться процесс резких выбросов потерь. На этом этапе процесс прекращается или изменяется количество итераций I и выполняется следующий вариант с другим значением W и другим приращением. Такую человеко – машинную процедуру проводят несколько раз.

Решение задачи зависит от выбора отношения стоимостей ошибок классификации. Удобно при K=2 выбирать отношение стоимостей ошибок, задаваясь величиной p^* (5). При выборе слишком большого значения p^* при определенных значениях вероятностных характеристик классов искомая область класса может оказаться пустой, см., например, (8).

4. Примеры. Пусть одномерная случайная величина x принадлежит одному из двух классов, имеющих нормальные распределения. Для первого класса, N(0,1), среднее нуль, стандартное отклонение - единица. Для второго класса, $N(m,k^2)$, среднее $m \ge 0$ и стандартное отклонение $k \ge 1$. P_1 – априорная вероятность первого класса. На рис.1 представлены плотности распределений, функции потерь и ДФ при $C_{12}=1$, $C_{21}=2$, $P_1=0.4$, m=2, k=1.5.

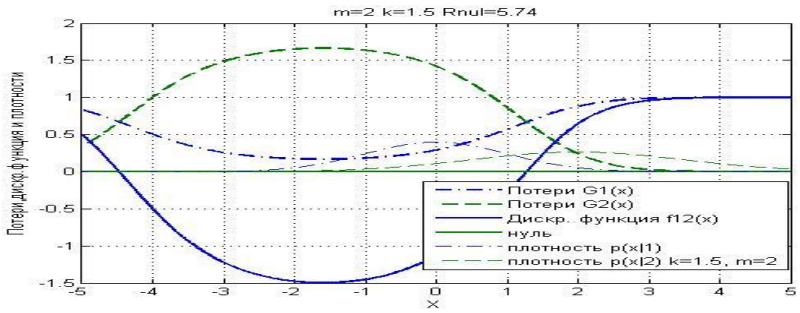

Рис.1

Точки, для которых $f_{12}(x) < 0$, следует отнести к классу 1, иначе – к классу 2. Величина Rnul равна расстоянию между корнями ДФ: $Rnul=|x_1 - x_2|$. Лишь эти точки важны для классификации, а не ДФ со всеми ее изгибами. Левый корень лежит в области малых вероятностей и им можно пренебречь.

Направляющей цилиндра, $f_{12}(x)=0$ является уравнение

$$ax^2 + bx + c = 0,$$

где $a = 1 - k^2$, $b = -2m$, $c = m^2 - 2k^2 \operatorname{Ln}(C_{12}(1-P_1)/(C_{21}P_1k))$.

При k=1 (равные дисперсии классов) уравнение имеет один корень, $x_1=-c/b$. Классы в этом случае различимы всегда, если m # 0.

При $k > 0$ наличие двух корней уравнения, когда $D= b^2 - 4ac > 0$, свидетельствует о различимости классов. Один корень, когда D=0, $x_1 = -b / 2a$, соответствует случаю, когда $\min f_{12}(x) = f_{12}(x_1)=0$, т.е. область класса 1 состоит из единственной точки x_1, где ДФ касается оси X.

Из условия $D<0$ при $k>1$ вытекают условия неразделимости классов:

(8) $m^2<2 (k^2-1) \operatorname{Ln}(C_{12}(1-P_1)/(C_{21}P_1 k))$, $1 < k < C_{12}(1-P_1)/(C_{21}P_1))$,

откуда следует и ограничение на отношение стоимостей потерь

$$C_{12}/ C_{21} > kP_1/(1-P_1), \qquad (9)$$

при котором классы не различаются – точки надо относить во второй класс.

На рис.2 представлены зависимости Rnul от параметров m и k при $P_1=0.5$ и $C_{12}= C_{21}$, т.е. классы различимы при $k>1$. Из рисунка видно, что при $m > 1$ для $k > 1.3$ расстояние между корнями Rnul > 3. Поэтому для АДФ можно использовать линейную функцию.

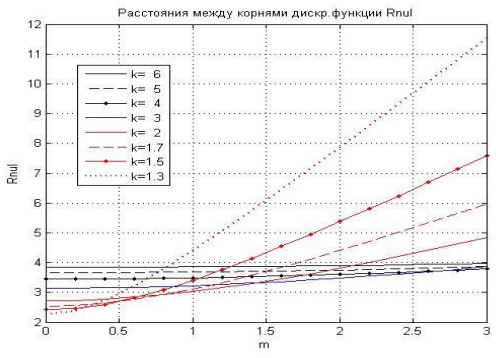

Рис. 2

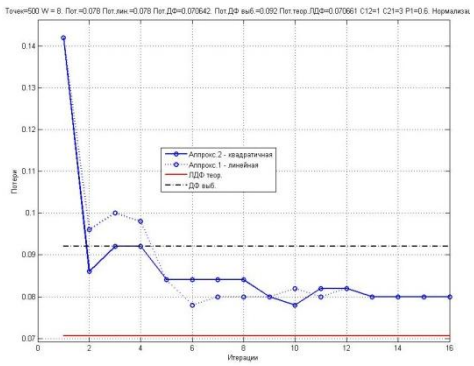

Рис.3

На рис.3 представлен пример изменения выборочных потерь классификации по итерациям (7) при аппроксимации ДФ линией $\lambda_1 + \lambda_2 x$ (точечная линия) и полиномом $\lambda_1 + \lambda_2 x + \lambda_3 x^2$ (сплошная) для обучающей выборки в 500 точек, сгенерированной при $P_1=0.6$, k=1.5, m=4, $C_{12}=1$, $C_{21}=3$.

Первое приближение получено обычным МНК. Далее 4 итерации при W=2, затем

каждые 4 итерации выполнялись сериями при увеличенном W на два по сравнению с предыдущей серией, достигнув W=8. Горизонтальная сплошная линия - теоретические минимальные средние потери при линейной аппроксимации ДФ (ЛДФ) в одном из корней $f_{12}(x)=0$. Горизонтальная штрихпунктирная линия – средние потери, если использовать ДФ по выборочным оценкам параметров распределений вероятностей. На рис.4 представлены аналогичные результаты по выборке в 50000 точек.

Из графиков рис.3, рис.4 и рис.5 (для m=1, 500 точек) видно, как по итерациям уменьшаются ошибки АДФ и разница между ними.

Данные для двумерного случая признаков x: для первого класса: m_1 = (0,0), ковариационная матрица по строкам S_1 = (1, 0; 0, 1); для второго: m_2=(0,4), ковариационная матрица S_2 = (6, 2; 2, 4). P_1 =0.6, C_{12} = 1, C_{21} = 3. На рис.6 изображен график изменения потерь в итерационном процессе. Точек в выборке 500. На рис.7 изображены результаты, аналогичные рис.6, для выборки в 50000 точек.

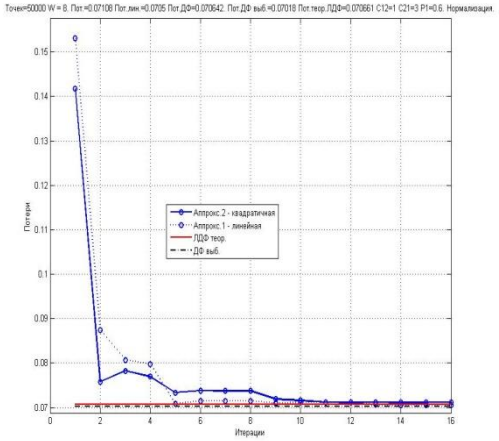

Рис. 4

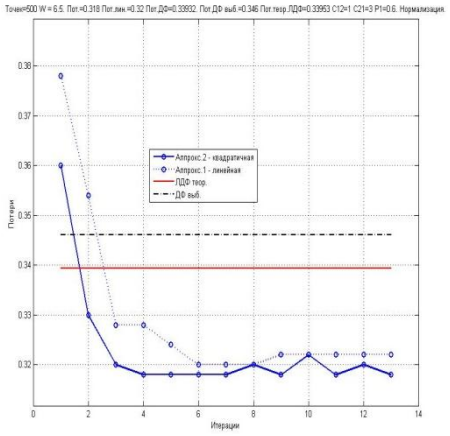

Рис.5

Выполнялись аппроксимации цилиндрами с направляющими: линейная, $\varphi= (1, x_1 , x_2)$ - точки; квадратичная (кв), $\varphi= (1, x_1^2 , x_2^2)$ – сплошная синяя; полином второго порядка, $\varphi= (1, x_1^2 , x_2^2, x_1 x_2, x_1 , x_2)$ – сплошная черная. Сплошной красной горизонтальной линией изображены потери по ДФ, вычисленной по вышеприведенным данным. Штрихпунктирной черной горизонтальной линией – потери, если ДФ строится по выборочным параметрам законов распределений.

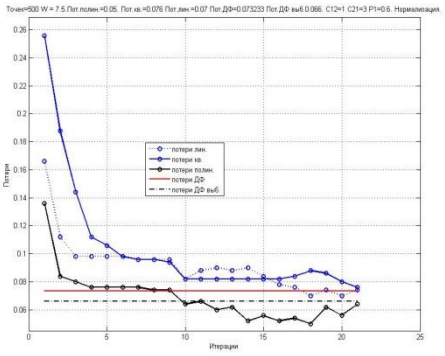

Рис. 6

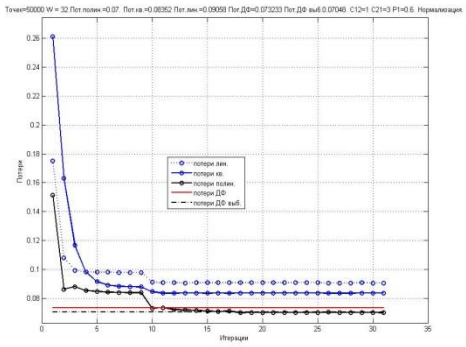

Рис.7

В качестве реального примера [9] решена задача диагностики рака шейки матки. Обучающая выборка - матрица 252x217. В 252 строках первые 99 строк соответствуют здоровым людям, остальные 153 строки – больным.

Первый столбец содержит номер класса: 0 - здоровые, 1 – больные. Остальные 216 столбцов соответствуют признакам.

Примитивно из набора 216 признаков отобраны наиболее коррелированные с искомой

величиной признаки: столбцы 22, 104, 115 с коэффициентами корреляции: 0.6037, 0.6372, 0.6361.

Для двух АДФ – полинома второго и первого порядков – вероятности ошибок диагностики равны 4.76% и 6.75%. Показатели AUC ROC-кривой [5] вычислялись по апостериорным вероятностям заболевания в точках выборки, вычисленным определенным образом при изменении p^* в диапазоне от 0.02 до 0.95, и равнялись 0.99 и 0.98 . Соответственно, две АДФ:

$$f_{12}(x) = -0.96932 - 0.00092584x_1^2 - 0.013056x_2^2 - 0.0045356x_3^2 + 0.0067098x_1x_2 + 0.061701x_1x_3 + 0.010848x_2x_3 + 0.053476x_1 + 0.19326x_2 + 0.12671x_3, \quad (10)$$

$$f_{12}(x) = -0.96892 + 0.16965x_1 + 0.10908x_2 + 0.14321x_3, \quad (11)$$

где x_1 , x_2 , x_3 – вышеуказанные столбцы – признаки.

Если $f_{12}(x) \le 0$, то точку следует отнести в класс 1 – здоровые, иначе – в класс 2 – больные.

4. Заключение.

1. Для решения задач классификации в стохастической постановке описан способ аппроксимации ДФ взвешенным методом наименьших квадратов с бОльшими весами в окрестности нулевых значений ДФ.

2. Способ повышает эффективность простых решающих правил. Более простой вид ДФ в окрестности нулевых значений позволяет использовать и более простые аппроксимации ДФ, что уменьшает риск переобучения.

3. Перспективность метода продемонстрирована на модельных примерах и реальном примере – диагностике рака шейки матки.

ЛИТЕРАТУРА

1. *Цыпкин Я.З., Кельманс Г.К.* Адаптивный байесов подход. // Проблемы передачи информации. 1970, том 6, выпуск 1. С. 52-59.
2. *Андерсон Т.* Введение в многомерный статистический анализ. М.: Физматгиз, 1963.
3. *Цыпкин Я.З.* Основы теории обучающихся систем. М.: Наука, 1970.
4. Vladimir N. Vapnik. An Overview of Statistical Learning Theory. http://www.recognition.mccme.ru/pub/papers/SLT/Vapnik99overview.pdf
5. Воронцов К. В.. Математические методы обучения по прецедентам (теория обучения машин) http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf
6. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. М.: Статистика, 1973.
7. *Chih-Jen Lin , Ruby C. Weng , and S. Sathiya Keerthi.* Trust Region Newton Method for Large-Scale Logistic Regression. http://www.recognition.mccme.ru/pub/papers/logistic/lin07trust.pdf
8. *Зенков В.В.* Аппроксимация дискриминантных функций в окрестности нулевых значений. // Известия Академии наук СССР. Техническая кибернетика. 1973. №2. С. 152-156.
9. Данные для задания на ТМШ 2014. http://www.machinelearning.ru/wiki/images/e/e1/School-VI-2014-task-3.rar