УДК 004.9

ФУНДАМЕНТАЛЬНЫЕ АЛГЕБРАИЧЕСКИЕ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ ДЛЯ КЛАССИФИКАЦИИ ЧИСЛОВЫХ ПРЕЦЕДЕНТОВ

Лобода Л.Д., студент группы ИТб-212, III курс,
Гутова Е.В., старший преподаватель, кафедра математики
Кузбасский государственный технический университет имени Т.Ф. Горбачева, г. Кемерово

Классификация – это задача, в рамках которой существует конечное множество объектов, которые по тем или иным признакам сепарированы на различающиеся классы. Каждый объект конечного множества *правильно* может быть отнесен только к какому-то единственному (конкретному) классу. Поскольку в предыдущем пункте настоящей главы было определено, что большинство признаков набора данных являются номинативными, то этот факт означает, что данные *линейно разделимы* друг относительно друга [1]. Поэтому далее рассмотрим задачу дихотомической классификации для линейно разделимой выборки объектов данных. Эти классы будут обозначаться как -1 и +1 соответственно. Определяемые объекты классификации – это векторы из множества R^n. То есть, считаем, что объекты описываются n числовыми признаками. Необходимо построить модель линейной классификации.

Линейный классификатор — это математический метод (аппроксимирующая функция), позволяющий вычислять скалярное произведение между вектором признакового описания объекта и направляющим вектором разделяющей гиперплоскости.

При этом принимаются следующие ограничения: если объект находится[1] по одну сторону от разделяющей гиперплоскости, то класс объекта будет определён как «+1». Иначе – класс определяется значением «-1». То есть, если скалярное произведение положительно, то и сам класс положителен, если отрицательно, – то и класс также будет отрицателен.

В общем случае, математический вид модели линейного классификатора представляется в виде формулы (1).

$$a(x;\ \omega;\ \omega_0) = sign(\langle x, \omega\rangle - \omega_0) \quad (1)$$

Параметры в формуле **(3)** $\omega \in R^n, \omega_0 \in R$. Параметр ω_0 именуется свободным членом. В некоторых типах классификаторов его исключают из рассмотрения [3 – 5]. Какую роль он играет в методе опорных векторов станет

[1] Под понятием «находится» понимается геометрическое трактование задачи классификации для разделения объектов выборки данных на разные классы.

понятно после аналитического решения задачи в математических терминах [2, 3].

Поскольку мы решаем задачу классификации для линейно разделимой выборки, то необходимо учитывать, что требуется определить некий принцип оптимизации для того, чтобы улучшать значения весовых коэффициентов ω и ω_0.

Таким принципом оптимизации является число ошибок на обучающей выборке. Но критерий «число ошибок» не удобен в алгоритмах классификации, так как он предоставляет для исследования кусочно-постоянную функцию от параметров модели и оптимизировать такую функцию алгоритмически сложно [3, 4]. Было бы лучше, чтобы оптимизируемый функционал по параметрам имел бы непрерывный характер или был гладким.

Для достижения данного приближения вводится понятие «отступ». В эквивалентном виде условие того, что алгоритм a ошибается на объекте x_i можно переписать как условие того, что отступ объекта отрицателен (2).

$$\sum_{i=1}^{l}[a(x;\ \omega;\ \omega_0) \neq y_i] = \sum_{i=1}^{l}[M_i(\omega;\ \omega_0) < 0] \to min_{\omega,\omega_0} \qquad (2)$$

где $M_i(\omega;\ \omega_0) = \langle x, \omega\rangle - \omega_0$, а y_i – отступ объекта x_i

Отступ – это расстояние до разделяющей поверхности, взятое со знаком. Иными словами, если отступ положительный – то ошибки нет, тогда выражение $\langle x, \omega\rangle - \omega_0$ того же знака, что и правильный ответ y_i.

Теперь, формализовав закономерности (1) и (2), необходимо применить математический прием, называемый аппроксимацией эмпирического риска для того, чтобы облегчить алгоритмическую сложность модели [5]. Эмпирическим риском называется число ошибок на обучающей выборке, а его аппроксимация возникает при замене пороговой функции потерь (3) какой-либо верхней оценкой [4].

$$Q(\omega, \omega_0) = \sum_{i=1}^{l}[M_i(\omega;\ \omega_0) < 0] \quad (3)$$

В зависимости от того, какую верхнюю оценку мы выбираем, можно получить тот или иной алгоритм обучения. В классификаторе рассматриваемая функция является составленной из двух отрезков прямых (лучей), представляющихся как значение $(1 - M)$ с положительной срезкой (4): если текущее значение положительно, то принимается само значение, если аргумент отрицателен, то принимаем значение функции «0».

$$Q(\omega, \omega_0) = \sum_{i=1}^{l} [M_i(\omega;\ \omega_0) < 0]$$
$$\leq \sum_{i=1}^{l} \left(1 - M_i(\omega;\ \omega_0)\right)_+ + \frac{1}{2C}\left|\left|\omega\right|\right|^2 \rightarrow min_{\omega,\omega_0} \quad (4)$$

Однако, формула **(6)** учитывает также и линейную разделимость выборки, так как позволяет контролировать параметр $\frac{1}{2C}\left|\left|\omega\right|\right|^2$, за счет чего достигается контроль мультиколлинеарности [5].

Отобразим представление пороговой функции на графике, сравнивая её со стандартной пороговой функцией линейного классификатора (3). Сравнение в геометрическом виде выполнено на графике и представлено на рисунке 1.

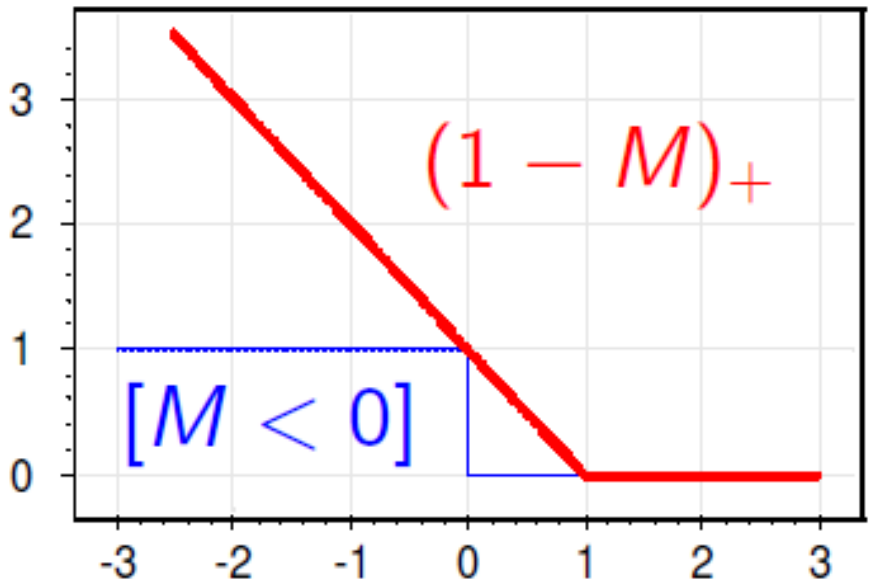

Рисунок 1 – Сопоставление пороговых функций классического линейного классификатора и метода опорных векторов

Параметр ω_0 в формуле (4) определяет положение разделяющей гиперплоскости, то есть разделяющую гиперплоскость все также можно перемещать параллельным переносом в любую точку пространства и на это математическое действие никаких ограничений накладываться не будет [3].

Благодаря аппроксимации (4) реализована возможность штрафовать объекты за приближение к границе классов, поэтому сохраняющийся зазор позволит сохранять устойчивость модели классификатора даже в условиях близкорасположенных объектов классов.

Неустойчивые решения будут «оштрафованы» при мультиколлинеарности (линейной разделимости классов) во избежание появления эффекта переобучения, поэтому в условиях возможного линейного

разделения групп, алгоритм будет работать с высокой точностью, не попадая в положения переобучения [5].

Рассмотрим теперь конкретную задачу линейно разделимой выборки. В формализованном математическом виде это означает, что выборка $X^l = (x_i; y_i)_{i=1}^l$, тогда:

$$\exists\, \omega, \omega_0 \colon M_i(\omega;\ \omega_0) = y_i(\langle \omega, x_i \rangle - \omega_0) > 0, \qquad i = 1, \dots, l$$

Из записи следует, что существует такое положение разделяющей гиперплоскости, что отступы положительны для всех объектов.

Коль скоро выполняется система неравенств (неравенств столько, сколько объектов обучающей выборки $i = 1, \dots, l$), то возникает возможность перенормировки.

Если система неравенств $y_i(\langle \omega, x_i \rangle - \omega_0) > 0$ выполнена, то коэффициенты ω и ω_0 можно одновременно умножить на одно и тоже число и от проведенной операции система уравнений не изменится.

Теперь мы вправе распорядиться возможностью перенормировки. Так как в уравнении неравенств $y_i(\langle \omega, x_i \rangle - \omega_0) > 0$ знак строгий (строго больше нуля), то левую часть можно взять в качестве коэффициента и принять за минимум функции.

Тогда главная особенность выполнения операции нормировки состоит в условии минимизации:

$$min_{i=1,\dots,l} M_i(\omega;\ \omega_0) = 1$$

Благодаря нормировке определяется оптимальная гиперплоскость разделения объектов выборки, относящихся к разным классам (рисунок 2).

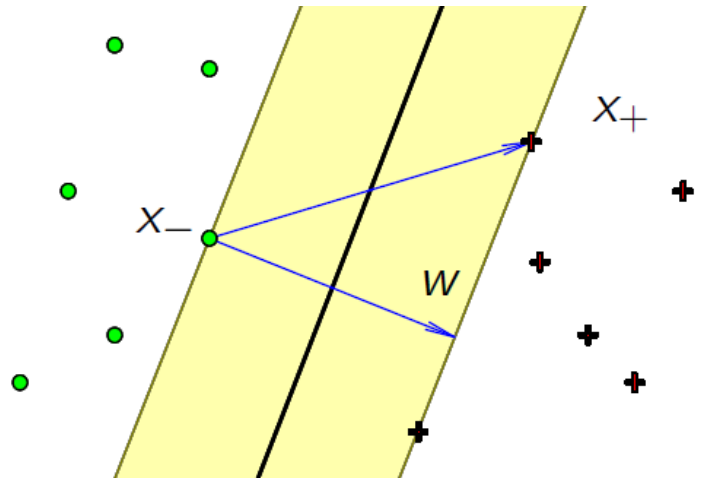

Рисунок 2 – Оптимальная гиперплоскость разделения классов

Следующим шагом определяем математические условия оптимальной гиперплоскости (рисунок 2).

Из рисунка 2 следует, что разделяющая полоса (выделенная желтым цветом) максимально широкая. Это означает, что классификация выполнена наиболее надежным способом – выборка разделена так, что объекты обоих

классов отстоят как можно дальше от разделяющей поверхности. Кроме этого, очевидно, что разделяющая полоса упирается обеими сторонами в объекты: с одной стороны ее подпирает объект класса «**-1**», а с другой стороны – объекты класса «**+1**». Расширить полосу дальше не представляется возможным.

Утверждается [6, 7], что, если полоса максимально широкая (оптимальная разделяющая гиперплоскость – лежащая посередине максимально широкой полосы, которая разделяет два класса), то условия оптимальной ширины полосы выражаются в формализованном виде условия (5).

$$\frac{\langle x_+ - x_-, \omega \rangle}{||\omega||} = \frac{2}{||\omega||} \to max \quad (5)$$

Таким образом, решение задачи классификации состоит в задаче минимизации квадратичного функционала при линейных ограничениях неравенств.

При ***линейно неразделимой*** выборке система неравенств $y_i(\langle \omega, x_i \rangle - \omega_0) > 0$ перестанет быть совместной и множество решений для такой задачи будет пустым. Соответственно, решить задачу данным алгоритмом при текущей модификации (в условиях линейно неразделимой выборке) не будет представляться возможным с точки зрения математического программирования [3 – 5, 7]. Отметим, что такая ситуация в рассматриваемой прикладной задаче невозможна, так как признаки являются номинативными (то есть их численное представление является отображением взаимосвязи текстовых объектов с их количественными представлениями).

Список литературы:

1. Свидетельство о государственной регистрации программы для ЭВМ № 2023680124 Российская Федерация. BrainPower : № 2023669010 : заявл. 16.09.2023 : опубл. 26.09.2023 / Р. В. Майтак. – EDN QXBJIM.
2. Математические и программные методы построения моделей глубокого обучения : Учебное пособие / А. В. Протодьяконов и др. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2023. – 176 с. – ISBN 978-5-9729-1484-5. – EDN PZLUAH.
3. Свидетельство о государственной регистрации программы для ЭВМ № 2023680335 Российская Федерация. Maitak Intelligence Natural Language Processing Module : № 2023669704 : заявл. 27.09.2023 : опубл. 28.09.2023 / Р. В. Майтак.
4. Методы восстановления непараметрической регрессии в условиях несбалансированных данных / А. Д. Салычева и др. – Вологда :

Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2024. – 192 с. – ISBN 978-5-9729-1856-0. – EDN AAJATW.

5. Свидетельство о государственной регистрации программы для ЭВМ № 2023684619 Российская Федерация. Efficient Network: № 2023684038: заявл. 14.11.2023: опубл. 16.11.2023 / П. А. Пылов.
6. Свидетельство о государственной регистрации программы для ЭВМ № 2023680070 Российская Федерация. Модернизированная модель DBSCAN для определения скрытых взаимосвязей : № 2023668841 : заявл. 13.09.2023 : опубл. 26.09.2023 / Р. В. Майтак. – EDN KQUUKF.
7. Асимптотический анализ поведения прикладных моделей машинного обучения : Учебное пособие / А. В. Протодьяконов и др. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2023. – 144 с. – ISBN 978-5-9729-1455-5. – EDN APHQME.