Новый метод обучения байесовской логистической регрессии с использованием лапласовского регуляризатора

Ветров Д. П., Кропотов Д. А., Курчин О. В.
VetrovD@yandex.ru, DKropotov@yandex.ru, 4education@mail.ru
Москва, ВМиК МГУ, ВЦ РАН, ВМиК МГУ

Рассмотрим стандартную задачу классификации на два класса по заданной обучающей выборке $\mathcal{D} = (X, T) = \{\mathbf{x}_i, t_i\}_{i=1}^m$, где $\mathbf{x} \in \mathbb{R}^d$, $t \in \{-1, 1\}$. Основным недостатком многих алгоритмов классификации является эффект переобучения. Одним из наиболее популярных подходов к его устранению является байесовская регуляризация, суть которой заключается в том, что задаются априорные распределения вероятности некоторых (возможно всех) весов классификатора. При этом настройка весов производится путем оптимизации суммы некоторого функционала качества, связанного с ошибкой на обучающей выборке, и регуляризатора, предотвращающего перенастройку на данные. Данную концепцию используют такие широко известные модели, как метод опорных векторов, метод релевантных векторов, логистическая регрессия.

Логистическая регрессия

Классическим подходом, особенно популярным при решении задач классификации в медицине, позволяющим вычислить апостериорную вероятность принадлежности объекта к одному из двух классов, является логистическая регрессия, основанная на линейной комбинации признаков объектов:

$$p(t|\mathbf{x}) = \frac{1}{1 + \exp(-t\hat{y}(\mathbf{x}))},$$

где $t \in \mathcal{Y} = \{-1; 1\}$, $\hat{y}(\mathbf{x}) = w_0 + w_1 x^1 + \cdots + w_d x^d$. Тогда отрицательный логарифм правдоподобия принимает вид:

$$L(\mathcal{Y}^m|\mathcal{X}^m, \mathbf{w}) = \sum_{i=1}^{n} \ln\bigl(1 + \exp(-t_i\hat{y}(\mathbf{x}_i))\bigr). \tag{1}$$

Поиск значений весов $\mathbf{w}$ осуществляется путём максимизации (1). Полученный классификатор обладает следующей особенностью: практически ни один из его весов w_i не равен в точности нулю. Добавление к критерию (1) лапласовского регуляризатора позволяет получать более разреженные решения [1]:

$$F = L(\mathcal{Y}^m|\mathcal{X}^m, \mathbf{w}) + \lambda R(\mathbf{w}), \tag{2}$$

где $R(\mathbf{w}) = \sum_{i=1}^d |w_i|$.

Устранение параметра распределения λ

Заметим, что параметр распределения λ в критерии (2) заранее неизвестен. Для его определения может быть проведено усреднение по данному параметру при использовании для него некоторого априорного распределения. Поскольку данный параметр является параметром масштаба, то закономерным является использование несобственного распределения Джеффри, $p(\lambda) \propto 1/\lambda$ — аналог равномерного распределения в логарифмической шкале. Условная плотность вероятности $p(\mathbf{w}|\lambda)$ для лапласовского распределения вычисляется как

$$p(\mathbf{w}|\lambda) = \left(\frac{\lambda}{2}\right)^N \exp\bigl(-\lambda R(\mathbf{w})\bigr) = \prod_{i=1}^{N} \frac{\lambda}{2} \exp\bigl(-\lambda|w_i|\bigr),$$

где N — количество ненулевых весов классификатора.

Для определения априорной плотности распределения весов классификатора и устранения параметра λ производится усреднение $p(\mathbf{w}) = \int p(\mathbf{w}|\lambda)p(\lambda)d\lambda$. В результате приходим к следующему критерию оптимизации:

$$Q = L(\mathcal{Y}^m|\mathcal{X}^m, \mathbf{w}) + N \ln R(\mathbf{w}). \tag{3}$$

Полученная модель (BLogReg), в основе которой лежит логистическая регрессия, была разработана в 2006 году английскими учеными Коули и Тэлбот [3]. Ими же была предложена процедура настройки данной модели. Основным достоинством данной модели является разреженность получаемого решения. Однако, разрывность целевой функции (3) привела к тому, что данная процедура могла использовать только метод оптимизации первого порядка, что существенно сказывалось на длительности процедуры обучения. В работе [2] был предложен похожий подход усреднения путем интегрирования по параметру масштаба с использованием распределения Джеффри для гауссовского распределения.

Разработанный метод настройки

Целевая функция (3) не является всюду гладкой, и, более того, является разрывной, поэтому мы не можем использовать методы оптимизации второго порядка. Для решения данной проблемы предлагается заменить критерий (3) на его непрерывный аналог:

$$\hat{Q} = L(\mathcal{T}|\mathcal{X}, \mathbf{w}) + \hat{N} \log R(\mathbf{w}); \tag{4}$$

$$\hat{N} = n - \sum_{i=1}^{n} \exp\left(-\frac{w_i^2}{2\sigma^2}\right). \tag{5}$$

Здесь $\sigma > 0$ — некоторый положительный коэффициент нечеткости. Рассмотрим гипероктант $\mathcal{H}_{ML}$, в котором находится точка максимума

правдоподобия (1). Тогда можно показать, что достаточно проводить оптимизацию (4) в области $\mathcal{M} = \{\bar{w} \in \mathcal{H}_{ML}, |w_i| \geqslant \varepsilon > 0, \forall i = 1, \ldots, n\}$. При этом критерий (4) является гладким в области $\mathcal{M}$ и для оптимизации может быть использован метод Ньютона второго порядка с ограничениями, что приводит к значительному увеличению быстродействия процедуры обучения.

Теорема 1. *Пусть ε — погрешность оптимизации, а σ — коэффициент нечеткости в выражении* (5)*. Тогда применение нечеткой версии критерия* (4) *возможно тогда и только тогда, когда*

$$\frac{1}{\sigma^2} = \bar{o}(\varepsilon^{-2} \ln^{-1} \varepsilon).$$

Эксперименты

Был проведен ряд экспериментов на задачах из UCI репозитория [4] по сравнению разработанного метода с методами опорных и релевантных векторов, а также байесовской логистической регрессии. Полученные результаты свидетельствуют о том, что разработанный метод позволяет получить сравнимое с остальными методами качество распознавания и разреженность, но при этом обучается быстрее BLogReg. На выборках размерности порядка 10 скорость обучения выросла в 2–5 раз, на выборках размерности порядка 100 — в 8–20 раз.

Литература

[1] *Williams, P. M.* Bayesian regularization and pruning using a Laplace prior. // Neural Computation. — 1995. — Vol. 7. — P. 117–143.

[2] *Figueiredo M.* Adaptive sparseness for supervised learning. // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2007. — Vol. 25. — P. 1150–1159.

[3] *Cawley G. C., Talbot N. L.* Gene selection in cancer classiffication using sparse logistic regression with bayesian regularization. // Bioinformatics. — 2007. — Vol. 22. — P. 2348–2355.

[4] *Asuncion A., Newman D. J.* UCI Machine Learning Repository — 2007. — `www.ics.uci.edu/~mlearn/MLRepository.html`.