Расширение метода Expectation Propagation на случай логистического правдоподобия

Ветров Д. П., Кропотов Д. А., Пташко Н. О.
vetrovd@yandex.ru, dkropotov@yandex.ru, ptashko@inbox.ru
Москва, ВМиК МГУ, ВЦ РАН

Рассматривается стандартная задача классификации на два класса. Дана обучающая выборка $D = \{(x_i, t_i)\}_{i=1}^N$, $\mathbf{x}_i \in \mathbb{R}^d$, $t_i \in \{-1, 1\}$. Алгоритм классификации строится в виде $y(\mathbf{x}) = \text{sign}(\mathbf{w}^T\varphi(\mathbf{x}))$, где $\mathbf{w} \in \mathbb{R}^M$, $\varphi(\mathbf{x}) = [\varphi_1(\mathbf{x}), \ldots, \varphi_M(\mathbf{x})]^{\text{T}}$ — некоторые базисные функции. Значения $\mathbf{w}$ находятся путем максимизации регуляризованного правдоподобия:

$$p(\mathbf{t}|X, \mathbf{w})p(\mathbf{w}|\alpha) = p(\mathbf{w}|\alpha)\prod_{i=1}^N p(t_i|\mathbf{x}_i, \mathbf{w}) = p(\mathbf{w}|\alpha)\prod_{i=1}^N \Psi(t_i\mathbf{w}^T\varphi(\mathbf{x}_i); 0, 1),$$

где $\mathbf{t} = \{t_i\}_{i=1}^N$, $X = \{\mathbf{x}_i\}_{i=1}^N$, $\Psi(y; m, s^2) = \frac{1}{\sqrt{2\pi}s}\int\limits_{-\infty}^{y-m} \exp\left(-\frac{x^2}{2s^2}\right) dx$ — гауссовская функция распределения (пробит-функция), $p(\mathbf{w}|\alpha) \sim \mathcal{N}(0, A^{-1})$, где $A = \text{diag}(\alpha_1, \ldots, \alpha_M)$. Значения гиперпараметров α находятся с помощью максимизации обоснованности [1]:

$$p(\mathbf{t}|X, \alpha) = \int p(\mathbf{t}|X, \mathbf{w})p(\mathbf{w}|\alpha)d\mathbf{w} \to \max_\alpha.$$

Данный интеграл, как правило, не удается вычислить аналитически. Поэтому возникает проблема его адекватной аппроксимации. Одним из популярных способов решения этой задачи является алгоритм expectation propagation (EP) [2].

Алгоритм Expectation Propagation

Алгоритм EP использует тот факт, что правдоподобие является произведением простых множителей. Аппроксимируя каждый из них, получаем аппроксимацию всего апостериорного распределения.

$$p(\mathbf{w}|X, \mathbf{t}, \alpha) \propto p(\mathbf{w}|\alpha)\prod_{i=1}^N p(t_i|\mathbf{x}_i, \mathbf{w}) \equiv p(\mathbf{w}|\alpha)\prod_{i=1}^N g_i(\mathbf{w}) \approx$$
$$\approx p(\mathbf{w}|\alpha)\prod_{i=1}^N \widetilde{g}_i(\mathbf{w}) = q(\mathbf{w}).$$

Алгоритм аппроксимирует каждый множитель таким образом, чтобы результирующее апостериорное распределение было близко к аппроксимированному в смысле меры Кульбака-Лейблера (называемой также

Алгоритм 1. Expectation Propagation.

Вход: $g_i(x), i = 1, \ldots, N$;
Выход: $\widetilde{g}_i(x), i = 1, \ldots, N$;

1: инициализация: $\widetilde{g}_i = 1$; $v_i = \infty$; $m_i = 0$; $s_i = 1$; $q(\mathbf{w}) = p(\mathbf{w}|\alpha)$;
2: **цикл** *// пока все* (m_i, v_i, s_i) *не сойдутся*
3: **для** $i = 1, \ldots, N$
4: (a) Удаляем $\widetilde{g}_i$ из $q(\mathbf{w})$.
Получим $q^{\setminus i}(\mathbf{w}) \propto q(\mathbf{w})/\widetilde{g}_i \sim \mathcal{N}(\mathbf{m}_{\mathbf{w}}^{\setminus i}, \mathbf{V}_{\mathbf{w}}^{\setminus i})$;
$\mathbf{V}_{\mathbf{w}}^{\setminus i} = \mathbf{V}_{\mathbf{w}} + \frac{V_{\mathbf{w}}\varphi_i(V_{\mathbf{w}}\varphi_i)^{\mathrm{T}}}{v_i - \varphi_i^{\mathrm{T}} V_{\mathbf{w}}\varphi_i}$; $\mathbf{m}_{\mathbf{w}}^{\setminus i} = \mathbf{m}_{\mathbf{w}} + (\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i)v_i^{-1}(\varphi_i^{\mathrm{T}}\mathbf{m}_{\mathbf{w}} - m_i)$;
5: (b) Полагаем $\widehat{p}(\mathbf{w}) \propto g_i(\mathbf{w})q^{\setminus i}(\mathbf{w})$.
Находим $q(\mathbf{w})$, минимизирующее $KL(\widehat{p}(\mathbf{w})\|q(\mathbf{w}))$;
$\mathbf{m}_{\mathbf{w}} = \mathbf{m}_{\mathbf{w}}^{\setminus i} + \mathbf{V}_{\mathbf{w}}^{\setminus i}\rho_i\varphi_i$; $\mathbf{V}_{\mathbf{w}} = \mathbf{V}_{\mathbf{w}}^{\setminus i} + (\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i)\frac{\rho_i(\varphi_i^{\mathrm{T}}\mathbf{m}_{\mathbf{w}} + \rho_i)}{\varphi_i^{\mathrm{T}}\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i + 1}(\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i)^{\mathrm{T}}$;
$Z_i = \int_{\mathbf{w}} g_i(\mathbf{w})q^{\setminus i}(\mathbf{w})d\mathbf{w} = \Psi(z_i)$, где
$z_i = \frac{(\mathbf{m}_{\mathbf{w}}^{\setminus i})^{\mathrm{T}}\varphi_i}{\sqrt{\varphi_i^{\mathrm{T}}\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i + 1}}$; $\rho_i = \frac{1}{\sqrt{\varphi_i^{\mathrm{T}}\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i + 1}}\frac{\mathcal{N}(z_i;0,1)}{\Psi(z_i)}$;
6: (c) Используя $\widetilde{g}_i = Z_i\frac{q(\mathbf{w})}{q^{\setminus i}(\mathbf{w})}$, получаем:
$v_i = \varphi_i^{\mathrm{T}}\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i\left(\frac{1}{\rho_i(\varphi_i{}^{\mathrm{T}}\mathbf{m}_{\mathbf{w}} + \rho_i)} - 1\right) + \frac{1}{\rho_i(\varphi_i{}^{\mathrm{T}}\mathbf{m}_{\mathbf{w}} + \rho_i)}$;
$m_i = \varphi_i^{\mathrm{T}}\mathbf{m}_{\mathbf{w}}^{\setminus i} + (v_i + \varphi_i^{\mathrm{T}}\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i)\rho_i$;
$s_i = \Psi(z_i)\sqrt{1 + v_i^{-1}\varphi_i^{\mathrm{T}}\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i}\exp\left(\frac{1}{2}\frac{\varphi_i^{\mathrm{T}}\mathbf{V}_{\mathbf{w}}^{\setminus i}\varphi_i + 1}{\varphi_i^{\mathrm{T}}\mathbf{m}_{\mathbf{w}}^{\setminus i} + \rho_i}\rho_i\right)$;
7: Подсчет нормализующей константы и обоснованности:
$B = (\mathbf{m}_{\mathbf{w}})^{\mathrm{T}}(\mathbf{V}_{\mathbf{w}})^{-1}(\mathbf{m}_{\mathbf{w}}) - \sum_i \frac{m_i^2}{v_i}$;
$p(\mathbf{t}|X, \alpha) \approx \int\prod_{i=1}^N \widetilde{g}_i(\mathbf{w})d\mathbf{w} = \frac{|\mathbf{V}_{\mathbf{w}}|^{1/2}}{(\prod_j \alpha_j)^{1/2}}\exp(B/2)\prod_i s_i$;

KL-дивергенцией):

$$KL(p\|q) = \int q(x)\log\frac{p(x)}{q(x)}dx.$$

Каждый аппроксимирующий член $\widetilde{g}_i(\mathbf{w})$ выбирается равным $\widetilde{g}_i(\mathbf{w}) = s_i\exp\left(-\frac{1}{2v_i^2}(t_i\mathbf{w}^T\varphi(\mathbf{x}_i) - m_i)^2\right)$, где (m_i, v_i, s_i) — параметры.

Можно показать, что если функция $q(\mathbf{w})$ — гауссиана, то условие близости эквивалентно условию равенства первых и вторых моментов функций $p(\mathbf{w}|X, \mathbf{t}, \alpha)$ и $q(\mathbf{w})$. Алгоритм 1 представляет собой метод EP (для краткости $t_i\varphi_i$ обозначено как φ_i, а $\Psi(z; 0, 1)$ через $\Psi(z)$).

Модификация EP

Заметим, что приведенный алгоритм неприменим для популярной модели логистической регрессии, в которой правдоподобие задается выражением

$$p_\sigma(\mathbf{t}|X,\mathbf{w}) = \prod_{i=1}^{N} \sigma(t_i \mathbf{w}^T \varphi(\mathbf{x}_i)) = \prod_{i=1}^{N} \frac{1}{1+\exp(-t_i \mathbf{w}^T \varphi(\mathbf{x}_i)))}.$$

Для этого при текущем векторе α находим максимум $\mathbf{w}_{MP}$ функции $p(\mathbf{w}|X,\mathbf{t},\alpha)$. Далее для каждого $i = 1,\ldots,N$ приближаем функцию $\sigma(y)$ в точке $y^i_{MP} = t_i \mathbf{w}^T_{MP} \varphi(\mathbf{x}_i)$ пробит-функцией $\Psi(y; m_i, {s_i}^2)$, получая приближение $p_\Psi(\mathbf{w}|\mathbf{t},\alpha) = \prod_{i=1}^{N} \Psi(y; m_i, {s_i}^2)$. Параметры пробит-функции m_i и ${s_i}^2$ находятся из требований совпадения значения нулевой и первой производных логистической и пробит-функции в точке y^i_{MP}

$${s_i}^2 = \frac{1}{\sqrt{2\pi} S(1-S)} \exp\left(-\frac{1}{2}\Psi^{-1}(S;0,1)\right), \quad m_i = y^i_{MP} - \Psi^{-1}(S;0,1)s_i,$$

где $S = \sigma(y^i_{MP})$, а $\Psi^{-1}(y;0,1)$ — функция, обратная к пробит-функции[1]. Далее применяем алгоритм EP для поиска приближения функции $p_\Psi(\mathbf{w}|\mathbf{t},\alpha) = \prod_{i=1}^{N} \Psi(y; m_i, {s_i}^2)$.

Следует отметить, что предлагаемое в работе приближение использует тот факт, что логистическая и пробит-функция очень близки по значениям (и значит можно приблизить одну другой), в то время как их логарифмы (которые используются при обучении логистической и пробит-регрессии соответственно) сильно отличаются, поэтому сами методы классификации приводят к существенно различным решающим правилам. В частности, пробит-регрессия значительно менее робастна.

Работа выполнена при поддержке РФФИ, проекты №№ 07-01-00211, 06-01-08045, 05-07-90333.

Литература

[1] *Tipping M.* Sparse Bayesian Learning and the Relevance Vector Machine // Journal of Machine Learning Research. — 2001. — Vol. 1, № 5. — P. 211–244.

[2] *Qi Y. A., Minka T. P., Picard R. W., Ghahramani Z.* Predictive Automatic Relevance Determination by Expectation Propagation // 21-st International Conference on Machine Learning, Banff, Canada, 2004.

[1] Реализованная, например, в среде MATLAB в виде процедуры norminv.