Relevance Tagging Machine

Молчанов Дмитрий Александрович[1]
Кондрашкин Дмитрий Андреевич[2]

1: *Студент, факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия*
2: *Аспирант, факультет компьютерных наук НИУ ВШЭ, Москва, Россия*
E-mail: dmolch111@gmail.com, kondra2lp@gmail.com

Рассмотрим задачу бинарной классификации объектов с бинарными признаками (тегами). $(X_i, T_i)_{i=1}^n$ — обучающая выборка, где объект X_i задается бинарным вектором $x = (x_1, x_2, ..., x_d)$, d — общее количество тегов, $T_i \in \{0, 1\}$ — метка класса. Здесь $x_j = 1$, если у объекта x есть тег с номером j, и $x_j = 0$, если этот тег не указан. Задача — восстановить зависимость значения метки t от тегов объекта x.

Считая теги независимыми, зададим вероятностную модель:

$$q_j = \mathsf{P}(t = 1|x_j), \quad \mathsf{P}(t = 1|x, q) = \prod_{j=1}^{d} q_j^{x_j} \cdot \left(\prod_{j=1}^{d} q_j^{x_j} + \prod_{j=1}^{d} (1 - q_j)^{x_j} \right)^{-1},$$

где $q = (q_1, \ldots, q_d)$, $q_j \in [0, 1]$ — параметры модели; значение q_j отвечает за влияние тега j на значение метки t.

В реальных данных бóльшая часть тегов может оказаться зашумленной или вообще не имеющей никакого отношения к метке, которую мы восстанавливаем. В данной работе мы рассматриваем задачу *автоматического отбора релевантных признаков* (ARD, automatic relevance determination [1]).

Для решения этой задачи можно воспользоваться *байесовским подходом*. Будем рассматривать параметры модели q как случайные величины и введем на них априорное распределение:

$$q_j \sim \mathrm{Beta}(\alpha_j + 1, \alpha_j + 1),$$

где $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d), \alpha_j \in [0, +\infty)$ — гиперпараметры модели. Отметим, что $\alpha_j = +\infty$ означает, что q_j может принимать только значение $q_j = \frac{1}{2}$, что соответствует исключению тега j из модели.

Согласно байесовскому подходу к выбору модели, оптимальное значение гиперпараметров должно максимизировать *обоснован-*

ность (evidence) модели:

$$p(T|X,\alpha) = \int \prod_{i=1}^{n} p(T_i|X_i,q)p(q|\alpha)dq = \int f(q,\alpha)dq.$$

Этот d-мерный интеграл нельзя посчитать ни аналитически, ни численно (в реальных задачах d может достигать десятков тысяч и более), поэтому мы предлагаем два подхода для его аппроксимации.

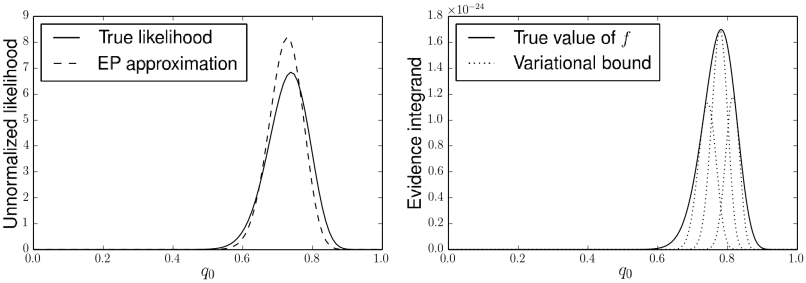

Рис. 1. Слева: Приближение функции правдоподобия, полученное с помощью EP. Справа: вариационные нижние оценки на функцию f при различных значениях вариационных параметров.

Первый подход основан на поиске ненормированного приближения для функции правдоподобия в виде произведения функций вида

$$g(u) = u^a(1-u)^b,\ a,b \in \mathbb{R}_+.$$

В таком случае приближение обоснованности можно посчитать аналитически и численно найти его максимум по α. Параметры такого приближения мы ищем методом *распространения ожидания* (Expectation Propagation, EP [2]).

Второй подход основан на поиске семейства нижних оценок на обоснованность [3]. В работе предложено семейство вариационных нижних оценок $L(q,\eta,\alpha)$ на функцию f такое, что d-мерный интеграл $\int L(q,\eta,\alpha)dq$ распадается на произведение d одномерных интегралов. Оптимальное значение α находится с помощью итерационного метода в духе EM–алгоритма — поочередно настраиваются вариационные параметры η и гиперпараметры α.

Эксперементально было показано, что оба подхода позволяют отсеивать зашумленные признаки. При этом второй подход, в отличие

от первого, позволяет также убирать и скоррелированные признаки. Со вторым подходом была проведена серия экспериментов на синтетических данных, показывающая, какая часть нерелевантных тегов отбрасывается в зависимости от общей зашумленности данных. Рассматривались выборки из 1000 объектов с 50 тегами, где у каждого объекта было не более 9 тегов. Видно, что предложенная модель успешно отсеивает не только зашумленные, но и скоррелированные признаки.

Число шумовых тегов	Тип шума	Убрано шумов
10	Случайные теги	95%
28		90.6%
46		89.7%
10	Скоррелированные теги	83.75%
28		72.32%
46		85.87%

В заключение авторы выражают признательность студенту 4-го курса кафедры ММП факультета ВМК МГУ А. С. Чистякову за помощь с применением подхода распространения ожидания.

Литература

1. Michael E. Tipping Sparse Bayesian Learning and the Relevance Vector Machine // In Journal of Machine Learning Research 1, 2001, P. 211-244
2. T. P. Minka Expectation Propagation for approximate Bayesian inference // In Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, August 2-5, 2001, P. 362–369.
3. T. S. Jaakkola and M. I. Jordan Bayesian logistic regression: a variational approach // Statistics and Computing, 10, 2000, P.25-37.