

МАСЛОВА
Арина Олеговна

магистратура, Петрозаводский государственный университет
(Петрозаводск, Россия),
ari.maslova03@yandex.ru

МЕТОДЫ МОНТЕ-КАРЛО ПО СХЕМЕ МАРКОВСКИХ ЦЕПЕЙ ДЛЯ ОБУЧЕНИЯ ФАКТОРИЗАЦИОННЫХ МАШИН

Научный руководитель:

Лукашенко Олег Викторович

Рецензент:

Воронов Роман Владимирович

Статья поступила: 06.11.2023;

Принята к публикации: 28.11.2023;

Размещена в сети: 01.12.2023.

Аннотация. В работе предложен подход для обучения факторизационных машин с логистической функцией активации на основе специального варианта схемы Гиббса с дополнительными переменными, имеющими распределение Пойа-Гамма. Были проведены численные эксперименты на синтетических и реальных данных, иллюстрирующие качество предложенного подхода. Результаты численных экспериментов дают основание полагать, что разработанный метод может быть использован для практических задач машинного обучения, в том числе для рекомендательных систем.

Ключевые слова: факторизационные машины, распределение Пойа-Гамма, схема Гиббса, байесовский подход

Для цитирования: Маслова А. О. Методы Монте-Карло по схеме марковских цепей для обучения факторизационных машин // StudArctic Forum. 2023. Т. 8, № 4. С. 103–109.

1. Введение

В последнее время модели факторизации привлекли большое внимание исследователей в области интеллектуальных информационных систем и машинного обучения. Такие модели продемонстрировали отличные возможности прогнозирования в нескольких важных программах, например, в рекомендательных системах. Эти системы разработаны для того, чтобы предоставлять рекомендации пользователю на основе разнообразных факторов. Они обрабатывают обширный объем сведений, фильтруя наиболее значимую информацию на основе пользовательских входных данных и других факторов, учитывая при этом предпочтения и интересы пользователя. Наиболее хорошо изученной моделью факторизации является матричная факторизация, которая позволяет нам предсказать связь между двумя категориальными переменными. Факторизационные машины являются в некотором смысле обобщением матричной факторизации. Основное достоинство данной модели состоит в том, что она обладает хорошей способностью к обучению на разреженных данных и позволяет строить рекомендательные системы на основе большого количества категориальных и вещественных признаков. Постоянные исследования в данной области продолжают расширять сферу применения факторизационных машин, делая их все более эффективными и способными адаптироваться к различным задачам, что, в свою очередь, способствует развитию и совершенствованию рекомендательных систем.

Рассмотрим линейную модель с учетом взаимодействия факторов:

$$y(\mathbf{x};\theta) = w_0 + \sum_{j=1}^p w_j x_j + \sum_{i=1}^p \sum_{j=i+1}^p w_{ij} x_i x_j. \quad (1)$$

В данном случае возникает проблема, так как число параметров представленной модели квадратично зависит от числа признаков: $(p-1)/2 + p + 1$. Если среди признаков есть категориальные переменные с большим числом категорий, то после бинарного кодирования число параметров значительно увеличится. В статье [Rendle, 2010] было предложено решение этой проблемы. Предположим, что вес взаимодействия определяется низкоразмерными векторами $\mathbf{v}_i, \mathbf{v}_j \in \mathbb{R}^k$ через их скалярное произведение:

$$y(\mathbf{x};\theta) = w_0 + \sum_{j=1}^p w_j x_j + \sum_{i=1}^p \sum_{j=i+1}^p (\mathbf{v}_i, \mathbf{v}_j) x_i x_j. \quad (2)$$

В этом случае число параметров линейно зависит от числа признаков: $pk + p + 1$.

Современные рекомендательные системы часто используют сложные модели машинного обучения. Эти модели обычно включают большое количество параметров и несколько параметров регуляризации, чтобы предотвратить переобучение. Для факторизационных машин было предложено несколько методов обучения. В зависимости от значений, которые могут принимать наблюдения, различают задачу регрессии и задачу классификации. В исходной статье 2010 года [Rendle, 2010] для факторизационных машин был предложен метод стохастического градиентного спуска (SGD), который можно использовать как для задачи регрессии, так и для задачи классификации. Кроме того, для задачи регрессии в [Rendle, 2011] авторы предложили использовать метод чередующихся наименьших квадратов (ALS). И в 2011 году были предложены байесовские факторизационные машины [Freudenthaler]. В рамках иерархической байесовской модели авторы предложили использовать Методы Монте-Карло по схеме марковских цепей (MCMC – Markov Chain Monte Carlo), а точнее схему Гиббса. При этом ранее методы Монте-Карло рассматривались только для задачи регрессии и задачи бинарной классификации с пробит-функцией связи. На основании проведенных экспериментов в [Freudenthaler], было обнаружено, что MCMC продемонстрировали более высокое качество прогнозирования по сравнению с другими методами. Превосходство заключается в том, что Методы Монте-Карло по схеме марковских цепей учитывают неопределенность параметров модели, что обеспечивает точные прогнозы даже для моделей с большим количеством категориальных и вещественных признаков. Также использование байесовского подхода естественным образом включает регуляризацию через априорное распределение параметров, что помогает избежать переобучения модели.

Анализируя литературу и документацию к различным библиотекам для факторизационных машин, мы выяснили, что для факторизационных машин с логистической функцией активации был предложен только метод стохастического градиентного спуска. Поэтому в работе была сделана попытка разработать схему Гиббса для логистической функции связи.

Структура статьи организована следующим образом: в разделе 2 приводится описание байесовских факторизационных машин. Подробное описание применения схемы Гиббса для байесовских факторизационных машин представлено в разделе 3. Раздел 4 посвящен результатам, полученным в ходе численных экспериментов, проведенных как на синтетических, так и на реальных данных, с целью продемонстрировать эффективность предложенного метода. В заключительном разделе 5 представлены выводы, основанные на полученных результатах, и рассматриваются дальнейшие планы и направления исследований в контексте проделанной работы.

2. Байесовские факторизационные машины

Байесовский подход в машинном обучении состоит в задании априорных распределений на оцениваемые параметры. При этом на практике часто используют так называемые байесовские иерархические модели, когда для параметров априорных распределений (которые будем называть гиперпараметрами) также задаются свои априорные распределения. В настоящей работе рассматривается байесовская модель, предложенная в работе [Freudenthaler]. Кратко опишем структуру данной модели, для чего выделим три множества параметров:

1. Исходные параметры $\theta = \{w_0, w_j, v_{j,s}, j = \overline{1, p}, s = \overline{1, k}\}$;
2. Параметры априорных распределений (гиперпараметры) $\theta_H = \{\mu_w, \lambda_w, \mu_{v_s}, \lambda_{v_s}, s = \overline{1, k}\}$;
3. Фиксированные параметры априорных распределений для гиперпараметров $\theta_0 = \{\mu_0, \mu_{w_0}, \lambda_{w_0}, \gamma_0, \alpha_\lambda, \beta_\lambda\}$.

Пусть имеется обучающая выборка $\{x_i, y_i\}$, $i = \overline{1, n}$, где $x_i \in \mathbb{R}^p, y_i \in \{0, 1\}$.

Статистическая модель для факторизационных машин имеет следующий вид:

$$y_i \sim \text{Ber}(p_i), \quad i = \overline{1, n}$$

$$p_i = F(y(x_i, \theta)), \quad i = \overline{1, n}$$

$$F(x) = 1 / (1 + e^{-x}).$$

Рассматриваемая байесовская иерархическая модель может быть задана в виде графовой модели, представленной на рис. 1.

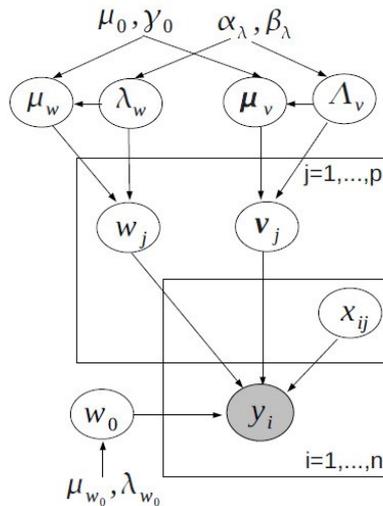


Рис. 1. Иерархическая байесовская модель [Freudenthaler]

Для всех параметров из множеств θ и θ_H задаются следующие априорные распределения:

$$w_0 \sim N(\mu_{w_0}, \lambda_{w_0}),$$

$$w_j \sim N(\mu_w, \lambda_w), \quad j = \overline{1, p},$$

$$v_{j,s} \sim N(\mu_{v_s}, \lambda_{v_s}), \quad j = \overline{1, p}, s = \overline{1, k},$$

$$\mu_w \sim N(\mu_0, \gamma_0 \lambda_w),$$

$$\mu_{v_s} \sim N(\mu_0, \gamma_0 \lambda_{v_s}), \quad s = \overline{1, k},$$

$$\lambda_w \sim \text{Ga}(\alpha_\lambda, \beta_\lambda),$$

$$\lambda_{v_s} \sim \text{Ga}(\alpha_\lambda, \beta_\lambda), \quad s = \overline{1, k}.$$

Здесь $N(\mu, \lambda)$ означает нормальное распределение с математическим ожиданием μ и дисперсией $1/\lambda$; $\text{Ga}(\alpha, \beta)$ – гамма-распределение с параметром формы α и параметром масштаба β .

3. Схема Гиббса для байесовских факторизационных машин

Для нахождения оценок в сложных байесовских моделях применяются методы приближенного байесовского вывода, в том числе на основе методов Монте-Карло по схеме марковских цепей.

Метод Н. Полсона, Дж. Скота и Дж. Уиндла был предложен в статье 2013 года [Polson] для логистической регрессии, но с некоторыми изменениями его можно адаптировать для модели факторизационных машин. В соответствии с подходом задается случайный вектор $\omega = (\omega_1, \dots, \omega_n)^T$ так, что ω_i имеет распределение Пуассона-Гамма:

$$\omega_i \sim PG(1, y(x_i, \theta)), \quad i = \overline{1, n}. \quad (3)$$

Для многомерной случайной величины $(\omega, \theta, \theta_H)$ применяется схема Гиббса (Gibbs sampling). Для этого требуется, чтобы была возможность генерировать реализации из условных распределений $\omega | \theta, \theta_H$, $\theta | \omega, \theta_H$, $\theta_H | \omega, \theta$. Исходя из предположений о графовой модели, получаем такое совместное распределение модели:

$$\begin{aligned} p(\theta, \theta_H, \omega, y | X, \theta_0) &\propto p(\omega, y | \theta, X) p(\theta | \theta_H, \theta_0) p(\theta_H | \theta_0) \\ &\propto \prod_{i=1}^n p(\omega_i, y_i | \theta, X) \prod_{\theta \in \Theta} p(\theta | \theta_H, \theta_0) \prod_{\theta_h \in \Theta_H} p(\theta_h | \theta_0), \end{aligned} \quad (4)$$

где y - вектор откликов, $X = \{x_i^T\}$ - матрица признаков.

Для применения схемы Гиббса из совместного распределения модели находятся полные условные распределения для всех $\theta \in \Theta$ и $\theta_h \in \Theta_H$:

$$p(\theta | \theta \setminus \{\theta\}, \theta_H), \quad (5)$$

$$p(\theta_h | \theta_H \setminus \{\theta_h\}, \theta). \quad (6)$$

Для гиперпараметров получаем следующие распределения:

$$p(\lambda_h | \theta_H \setminus \{\lambda_h\}, \theta, \theta_0) = Ga(\lambda_h | \bar{\alpha}_h, \beta_h), \quad (7)$$

$$p(\mu_h | \theta_H \setminus \{\mu_h\}, \theta, \theta_0) = N(\mu_h | \bar{\mu}_h, \bar{\sigma}_h^2), \quad (8)$$

где $h \in \{w, v_1, \dots, v_k\}$, $\bar{\alpha}_h, \beta_h, \bar{\mu}_h, \bar{\sigma}_h^2$ зависят от остальных значений параметров. Здесь $Ga(\cdot | \alpha, \beta)$ - плотность гамма-распределения с параметром формы α и параметром масштаба β , а $N(\cdot | \mu, \sigma^2)$ - плотность нормального распределения с математическим ожиданием μ и дисперсией σ^2 .

Полные условные распределения для каждого исходного параметра:

$$p(\theta | \theta \setminus \{\theta\}, \theta_H, \theta_0, \omega, y, X) = N(\theta | \bar{\mu}_\theta, \bar{\sigma}_\theta^2), \quad (9)$$

где $\bar{\mu}_\theta, \bar{\sigma}_\theta^2$ зависят от значений остальных параметров, обучающей выборки и вектора ω . Прямой расчет $\bar{\mu}_\theta, \bar{\sigma}_\theta^2$ требует $O(npk)$ операций (для каждого элемента обучающей выборки $O(pk)$ операций). Принимая во внимание свойство полилинейности модели

$$y(x, \theta) = g_\theta(x) + \theta h_\theta(x), \quad \forall \theta \in \Theta, \quad (10)$$

можно сократить число операций до $O(n)$, следуя подходу, предложенному ранее в статье [Rendle 2011], с незначительными изменениями.

4. Результаты

Численные эксперименты проводились как на синтетических, так и на реальных данных. В частности, в серии M экспериментов для синтетических данных рассчитывалась оценка исходных параметров модели помощью схемы Гиббса на основе распределения Пуассона-Гамма и находилось эмпирическое распределение для показателя AUC (площадь под ROC-кривой).

Данный показатель был рассчитан для метода стохастического градиентного спуска и для реализованной схемы Гиббса, при этом рассматривается ситуация, когда гиперпараметры θ_H фиксированы (Рис. 2).

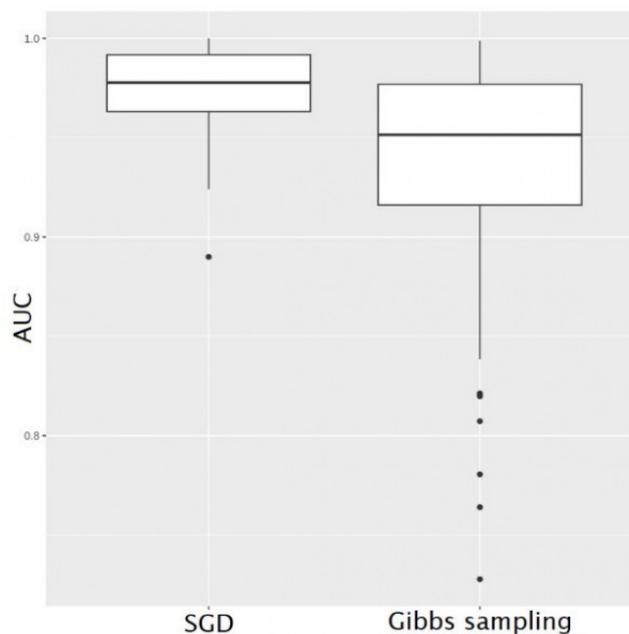


Рис. 2. SGD vs. Gibbs sampling

Исходя из диаграммы, видно, что значения для двух разных методов близки между собой и лежат в диапазоне от 0.9 до 1. Кроме того, эксперименты были проведены на реальных данных, в частности, показатели AUC были подсчитаны для стандартных наборов данных (Pima и т.д.) [Chopin].

Таблица

Показатели AUC для стандартных наборов данных

Dataset	AUC Gibbs sampling	AUC SGD
Pima (Indian diabetes)	0.776	0.617
Sonar	0.814	0.885
Banknote	0.92	0.85
Australian Credit Approval	0.943	0.691
Musk (Version 2)	0.97	0.715

Из таблицы видно, что только в случае набора данных Sonar показатель AUC для схемы Гиббса ниже, чем для AUC SGD, что, вероятнее всего, связано с небольшим количеством переменных в представленном наборе данных.

5. Заключение

В результате работы был реализован новый подход для обучения байесовских факторизационных машин с логистической функцией активации на основе специального варианта схемы Гиббса с дополнительными переменными, имеющими распределение Пуа-Гамма. Проведены вычислительные эксперименты на синтетических и на реальных данных, иллюстрирующие качество предложенного подхода. В дальнейшем предполагается провести сравнительный анализ различных методов обучения факторизационных машин, что поможет выявить преимущества и недостатки каждого метода и определить наиболее эффективные

методы обучения и применения факторизационных моделей в различных прикладных задачах. Кроме того, планируется провести дополнительные эксперименты, включая апробацию алгоритма на более крупных наборах данных, содержащих большое количество категориальных признаков, исследовать различные расширения и обобщения методов факторизации для задач многоклассовой классификации, учета пропущенных значений, скрытых факторов и т. п.

СПИСОК ЛИТЕРАТУРЫ

Chopin N. Leave Pima Indians Alone: Binary Regression as a Benchmark for Bayesian Computation / N. Chopin, J. Ridgway. *Statist. Sci.* 32 (1) 64–87, 2017. DOI: 10.1214/16-STS581

Freudenthaler C. Bayesian Factorization Machines / C. Freudenthaler, L. Schmidt-Thieme, S. Rendle // *Proceedings of the NIPS Workshop on Sparse Representation and Low-rank Approximation* 995-1000, 2011.

Polson N.G. Bayesian Inference for Logistic Models Using Polya–Gamma Latent Variables / Nicholas G. Polson, James G. Scott & Jesse Windle // *Journal of the American Statistical Association*. 2013.

Rendle S. Factorization machines // *Proceedings of the 10th IEEE International Conference on Data Mining*. IEEE Computer Society, 2010.

Rendle S. Fast context-aware recommendations with factorization machines // S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme // *Proceedings of the 34th ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011.

Arina O. MASLOVA

master's degree, Petrozavodsk State University
(Petrozavodsk, Russia),
ari.maslova03@yandex.ru

MARKOV CHAIN MONTE CARLO METHODS FOR FACTORIZATION MACHINE LEARNING

Scientific adviser:

Oleg V. Lukashenko

Reviewer:

Roman V. Voronov

Paper submitted on: 11/06/2023;

Accepted on: 11/28/2023;

Published online on: 12/01/2023.

Abstract. This paper presents a new approach for training factorization machines with the logistic activation function based on a special version of the Gibbs sampling incorporating additional variables with the Polya-Gamma distribution. To evaluate the effectiveness of the proposed method, numerical experiments were performed on both synthetic and real data. The results of these experiments indicate that the developed method can be used for various practical machine learning tasks, including recommendation systems.

Keywords: factorization machines, Polya-Gamma distribution, Gibbs sampling, Bayesian approach

For citation: Maslova, A. O. Markov Chain Monte Carlo Methods for Factorization Machine Learning. *StudArctic Forum*. 2023, 8 (4): 103–109.

REFERENCES

Chopin N., Ridgway J. Leave Pima Indians alone: Binary regression as a benchmark for Bayesian computation. *Statist. Sci*, 2017, No. 32 (1), pp. 64–87. DOI: 10.1214/16-STS581.

Freudenthaler C., Schmidt-Thieme L., et al. Bayesian factorization machines. In *Proceedings of the NIPS Workshop on Sparse Representation and Low-rank Approximation*, 2011, pp. 995-1000.

Polson N.G., Scott J.G., et al. Bayesian inference for logistic models using Polya–Gamma latent variables. *Journal of the American Statistical Association*, 2013, Vol. 108, Issue 504, pp. 1339-1349.

Rendle S. Factorization machines. In *Proceedings of the 10th IEEE International Conference on Data Mining. Sidney, IEEE Computer Society*, 2010.

Rendle S., Freudenthaler C., et al. Fast context-aware recommendations with factorization machines. In *Proceedings of the 34th ACM SIGIR Conference on Research and Development in Information Retrieval*, 2011.