Автоматическое дополнение плейлистов в рекомендательной системе пользователей*

Кислинский В. Г., Фролов Е. , Воронцов К. В.

kislinskiy.vg@phystech.edu; evgeny.frolov@skolkovotech.ru; vokov@forecsys.ru

Московский физико-технический институт

Работа посвящена исследованию метода совместной матричной факторизации в задаче top-N рекомендаций для автоматического продолжения плейлистов. Предлагается модель матричной факторизации, учитывающий дополнительную информацию о плейлистах и треках. Данный метод будет иметь, не только преимущество алгоритмов коллаборативной фильтрации, которые способны выявить скрытые свойства пользователей и обьектов, но также сможеть учитывать контекстную информацию, что поможет решить проблему холодного старта для обьектов. В данном методе будет введена дополнительная регуляризация, основанная на предположение, что если обьекты близки в пространстве признаков, то они близки в латентном факторном пространстве. Для анализа качества представленного алгоритма проводятся эксперименты на выборке из миллиона плейлистов MPD.

Ключевые слова: *задача top-N рекомендаций, совместная матричная факторизация, алгоритм LCE, латентное факторное пространство, коллаборативная фильтрация.*

1 Введение

Большинство методов коллаборативной фильтрации имеют ряд недостатков, основным из которых является проблема холодного старта. Другой подход к задаче рекомендаций, основанный на дополнительный информации, не имеет этой проблемы. Поэтому комбинирования этих методов [?]. Поэтому выбор и поиск оптимальной структуры нейронной сети также является вычислительно сложной процедурой, которая сильно влияет на итоговое качество модели. Использование переусложненных моделей с избыточным количеством неинформативных параметров также является препятствием для использования глубоких сетей на мобильных устройствах в режиме реального времени.

Существуют разные подходы к построению оптимальной сети. В работах [?, ?] предлагается использовать модель градиентного спуска для оптимизации сети. В ряде работ [?, ?] используются байесовские методы [?] оптимизации параметров нейронных сетей.

Другим методом поиска оптимальной структуры является прореживание переусложненной модели [?, ?, ?]. В работе [?] предлагатся удалять наименее релевантные параметры на основе значений первой и второй производных функции ошибки.

Данная работа посвящена прореживанию структуры сети. Предлагается удалять наименее *релевантные* параметры модели [?]. Метод предлагает построение исходной избыточной сложности нейросети с большим количеством избыточных параметров. Для определения релевантности параметров предлагается оптимизацировать параметры и гиперпараметры в единой процедуре. Для удаления параметров предлагается использовать метод Белсли.

Эксперимент метода проводится на выборке MNIST и синтетических данных. Результат сравнивается с моделью полученной при помощи алгоритма AdaNet [?].

Работа выполнена при финансовой поддержке РФФИ, проект № 00-00-00000. Научный руководитель: Воронцов К. В. Консультант: Фролов Е.

2 Постановка задачи

Задана выборка

$$\mathfrak{D} = \{\mathbf{x}_i, y_i\}, \; i = 1...N, \tag{2.1}$$

где $\mathbf{x}_i \in \mathbb{R}^n$, $y_i \in \{1, \dots, Y\}$, где Y — число классов.

Рассмотрим модель $f(\mathbf{x}, \mathbf{w}) : \mathbb{R}^n \times \mathbb{R}^m \to \{1, \dots, Y\}$, где $\mathbf{w} \in \mathbb{R}^m$ — пространство параметров модели.

$$f(\mathbf{x}, \mathbf{w}) = \text{softmax}\big(f_1(f_2(...(f_l(\mathbf{x}, \mathbf{w})\big), \tag{2.2}$$

где $f_i(\mathbf{x}, \mathbf{w}) = \tanh(\mathbf{wx})$, l — число слоев нейронной сети, $i \in \{1 \dots l\}$.

Параметр w_i модели f называется активным, если $w_i \neq 0$. Множество индексов активных параметров обозначим $\mathcal{A}$.

Задано множество параметров удовлетворяющих множеству активных параметров:

$$\mathbb{W}_{\mathcal{A}} = \{\mathbf{w} \in \mathbb{R}^m \mid w_i \neq 0, \; i \in \mathcal{A}\}, \tag{2.3}$$

Для модели f с множеством активных параметров $\mathcal{A}$ и соответствующего ей вектора параметров $\mathbf{w} \in \mathbb{W}_{\mathcal{A}}$ определим логарифмическую функцию правдоподобия выборки $\mathcal{L}_{\mathfrak{D}}(\mathfrak{D}, \mathcal{A}, \mathbf{w})$:

$$\mathcal{L}_{\mathfrak{D}}(\mathfrak{D}, \mathcal{A}, \mathbf{w}) = \log p(\mathbf{y}|\mathbf{x}, \mathcal{A}, \mathbf{w}) = \log p(\mathfrak{D}|\mathcal{A}, \mathbf{w}), \tag{2.4}$$

где $p(\mathbf{y}|\mathbf{x},\mathbf{w}, \mathcal{A})$ — апостериорная вероятность вектора $\mathbf{y}$ при заданных $\mathbf{x},\mathbf{w}, \mathcal{A}$.

Оптимальные $\mathbf{w}, \mathcal{A}$ находятся из минимизации $-\mathcal{L}_{\mathcal{A}}(\mathfrak{D}, \mathcal{A}, \mathbf{w})$ — логарифма правдоподобия модели:

$$\mathcal{L}_{\mathcal{A}}(\mathfrak{D}, \mathcal{A}, \mathbf{w}) = \log p(\mathbf{y}|\mathbf{x}, \mathcal{A}, \mathbf{w}) = \log \int_{w \in \mathbb{W}_{\mathcal{A}}} p(\mathbf{y}|\mathcal{A}, \mathbf{w}) p(\mathbf{w}|\mathcal{A}) d\mathbf{w}, \tag{2.5}$$

где $p(\mathbf{w}|\mathcal{A})$ — априорная вероятность вектора параметров в пространстве $\mathbb{W}_{\mathcal{A}}$.

Рассмотрим вариационный подход для решения этой задачи. Пусть задано распределение q, аппроксимирующее неизвестное апостериорное распределение $p(\mathbf{w}|\mathfrak{D}, \mathcal{A})$:

$$q(\mathbf{w}) \sim \mathcal{N}(\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1}), \tag{2.6}$$

где $\mathbf{m}, \mathbf{A}_{\text{ps}}^{-1}$ — вектор средних и матрица ковариации.

$$p(\mathbf{w}|\mathcal{A}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1}), \tag{2.7}$$

где $\boldsymbol{\mu}, \mathbf{A}_{\text{pr}}^{-1}$ — вектор средних и матрица ковариации.

Приблизим интеграл (2.5) вариационной оценкой [?]:

$$\mathcal{L}(\mathfrak{D}, \mathcal{A}, \mathbf{w}) = \mathcal{L}_{\mathbf{w}}(\mathfrak{D}, \mathcal{A}, \mathbf{w}) + \mathcal{L}_E(\mathfrak{D}, \mathcal{A}), \tag{2.8}$$

Первое слагаемое формулы (2.8) это сложность модели, которое определяется расстоянием Кульбака-Лейблера:

$$\mathcal{L}_{\mathbf{w}}(\mathfrak{D}, \mathcal{A}, \mathbf{w}) = D_{KL}\big(q(\mathbf{w})||p(\mathbf{w}|\mathcal{A})\big), \tag{2.9}$$

Второе слагаемое формулы (2.8) является матожиданием правдоподобия выборки $\mathcal{L}_{\mathfrak{D}}(\mathfrak{D}, \mathcal{A}, \mathbf{w})$, которое интерпретируется как функция ошибки:

$$\mathcal{L}_E\mathfrak{D}, \mathcal{A}) = \mathsf{E}_{\mathbf{w} \sim q} \mathcal{L}_{\mathfrak{D}}(\mathbf{y}, \mathfrak{D}, \mathcal{A}, \mathbf{w}), \tag{2.10}$$

Требуется найти параметры, доставляющие минимум суммарному функционалу потерь $\mathcal{L}$:

$$\mathbf{w} = \underset{\mathbf{w}\in\mathbb{W}_{\mathcal{A}},\ \mathcal{A}\in 2^m}{\arg\min} -\mathcal{L}(\mathfrak{D}, \mathcal{A}, \mathbf{w}), \tag{2.11}$$

3 Базовый метод

3.1 Случайное удаление

Метод случайного удаления заключается в том, что мы случайным образом удаляем некоторые нейроны из сети.

Тоесть:

$$\xi \in U(\mathcal{A}), \tag{3.1.1}$$

где ξ удаляемый параметр.

3.2 Optimal Brain Damage

Метод [?], использует вторую производную целевой функции по параметрам для определения не релевантных параметров. Рассмотрим функцию потерь $\mathcal{L}$ из (2.4) — разложенную в ряд Тейлора в некоторой окрестности вектора параметров $\mathbf{w}$:

$$\delta\mathcal{L} = \sum_{i\in\mathcal{A}} g_i \delta u_i + \frac{1}{2}\sum_{i,j\in\mathcal{A}} h_{ij}\delta w_i \delta w_j + O(||\delta\mathbf{w}||^3), \tag{3.2.1}$$

где δw_i — компоненты вектора $\delta\mathbf{w}$, g_i — компоненты вектора градиента $\nabla\mathcal{L}$, а h_{ij} — компоненты гессиана $\mathbf{H}$:

$$g_i = \frac{\partial\mathcal{L}}{\partial w_i} \qquad h_{ij} = \frac{\partial^2\mathcal{L}}{\partial w_i \partial w_j}. \tag{3.2.2}$$

Задача является вычислительно сложной в силу размерности матрицы $\mathbf{H}$. Введем следующее предположение [?], о том что удаление нескольких параметров приводит к такому же изменению функции потерь $\mathcal{L}$, как и суммарное изменение при индивидуальном удалении:

$$\delta\mathcal{L} = \sum_{i\in\mathcal{A}}^{N} \delta\mathcal{L}_i, \tag{3.2.3}$$

где N — число удаляемых параметров, $\mathcal{L}_i$ — изменение функции потерь, при удалении одного параметра $\mathbf{w}_i$.

В силу данного предположения будем рассматривать только диагональные элементы матрицы $\mathbf{H}$. После введенного предположения, (3.2.1) принимает вид:

$$\delta\mathcal{L} = \frac{1}{2}\sum_{i\in\mathcal{A}} h_{ii}\delta u_i^2, \tag{3.2.4}$$

Релевантность параметров определяется следующим образом:

$$s_i = h_{ii}\frac{w_i^2}{2}. \tag{3.2.5}$$

Получаем следующую задачу оптимизации:

$$\xi = \underset{j\in\mathcal{A}}{\arg\min}\, s_j, \tag{3.2.6}$$

где ξ — наименее релевантный параметр.

3.3 Удаление неинформативных параметров с помощью вариационного вывода

Для удаления параметров в работе [?] предлагается удалить параметры, которые имеют наибольшую плотность апостериорной вероятности ρ в нуле.

Для гауссовского распределения в работе [?] была предложена следующая задача оптимизации:

$$\xi = \arg\min_{i \in \mathcal{A}} \left| \frac{\mu_i}{\sigma_i} \right|, \tag{3.3.1}$$

где ξ — наименее релевантный параметр.

4 Метод Белсли

Помимо вышеописанных методов, предлагается метод основанный на модификации метода Белсли.

Пусть $\mathbf{w}$ — вектор параметров доставляющий минимум функционалу потерь $\mathcal{L}$ на множестве $\mathbb{W}_{\mathcal{A}}$, а $\mathbf{A}_{\text{ps}}^{-1}$ соответствующая ему ковариационная матрица.

Выполним сингулярное разложение матрицы $\mathbf{A}_{\text{ps}}^{-1}$:

$$\mathbf{A}_{\text{ps}}^{-1} = \mathbf{L}\mathbf{L}^{\text{T}} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^{\text{T}}\mathbf{V}\boldsymbol{\Lambda}^{\text{T}}\mathbf{U}^{\text{T}} = \mathbf{U}\boldsymbol{\Lambda}^2\mathbf{U}^{\text{T}}$$

Долевой коэффициент q_{ij} определим как вклад j-го признака в дисперсию i-го элемента вектора параметра параметров $\mathbf{w}$.

$$q_{ij} = \frac{u_{ij}^2 \lambda_{jj}}{D(w_i)}, \tag{4.1}$$

где $D(w_i)$ — дисперсия параметра w_i.

Используя метод (3.3.1) находим наименее релевантный параметр из набора параметров $\mathcal{A}$ — обозначим его ξ. Затем находим максимальные долевые коэффициента, соответствующие данному параметру w_ξ:

$$\zeta = \arg\max_{j \in \mathcal{A}} q_{\xi j}. \tag{4.2}$$

Параметры ξ и ζ определим как наименее релевантные параметры нейросети.

5 Базовый вычислительный эксперимент

В базовом эксперименте сравнивается качество и скорость сходимости трех моделей — полной нейронной сети, сети с произвольно удаленными параметрами и сети полученной при помощи Optimal Brain Damage.

В качестве исходных данных была использована выборка из 178 результатов химического анализа вин [1], по которым нужно было распределить вино по классам.

В результате эксперимента были получены следующие результаты, показаные на рисунке 1. На графике 1a изображена зависимость для обучающей выборки. На графике 1b — зависимость среднего значения функции потерь для тестовой выборки.

Из графиков видно, что на тестовой выборке при небольшом количестве удаляемых параметров, метод OBD дает лучше результат, чет произвольное удаление параметров

[1]http://archive.ics.uci.edu/ml/datasets/Wine

results/train.pdf

(a)

results/test.pdf

(b)

Рис. 1. Зависимости значения функции потерь от процента удаленных параметров.