**Этап L0:** проработка и формализация задачи, определение таргета.

**Итог:** определен таргет, определена мл задача

# Решение. Общая постановка задачи

## Определение задачи

Мы имеем дело с задачей построения рекомендаций, что является **частным случаем задачи ранжирования**, где запросом является сам пользователь и его предпочтения. Факт заказа (или покупки) товара ещё не говорит о том, что товар понравился пользователю, поэтому фидбек неявный (implicit feedback). Однако количество покупок само по себе является важным отражением предпочтений пользователя, и это свойство учитывает ALS, который обсудим далее.

Решать такую задачу будем с помощью метода коллаборативной фильтрации, потому что взаимодействия пользователей несут самую важную информацию. В частности, будет необходимо факторизовать полученную матрицу взаимодействий пользователей с товарами. В качестве условного рейтинга товара взять количество покупок товара. Такая матрица разреженная: в ней много пустых ячеек. Для итогового ранжирования необходимо восстановить значения “предпочтений” в этой матрице. Такой метод называется матричной факторизацией.

## Матричная факторизация

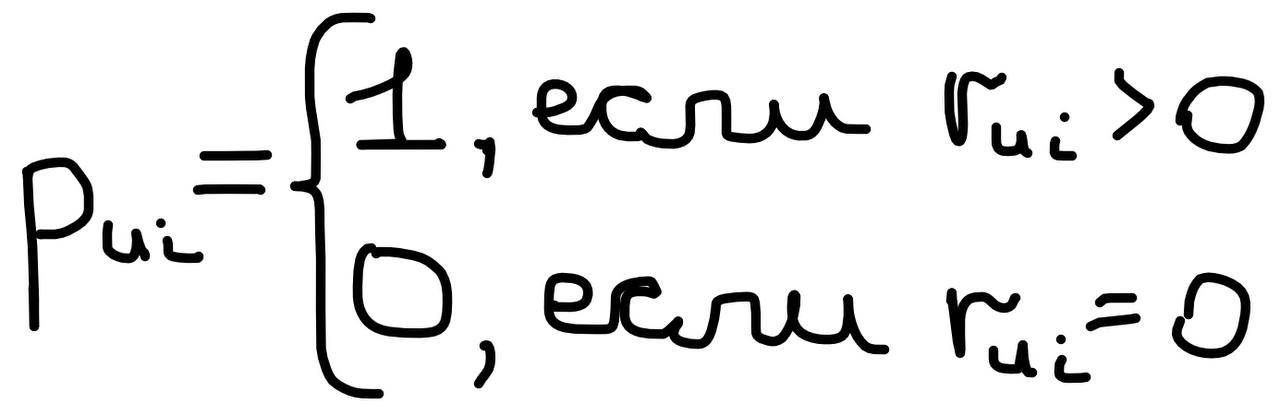
Разложим матрицу взаимодействий на 2 матрицы и перейдём в пространство меньшей размерности (подобно SVD, но сингулярная квадратная матрица для этой задачи нам не нужна): , где размерность матрицы равна , для – , для – . - это гиперпараметр, который будем подбирать по кросс-валидации, , .

Такое представления исходной матрицы взаимодействий позволяет выявить скрытые закономерности и поэтому данный подход относится к моделям со скрытыми переменными. Матрица условно отвечает за свойства пользователя: по строкам отложены числовые значения, которые отражают “предпочтения” пользователя. Столбцы матрицы - это такие же векторные числовые представления, но уже латентных качеств товара. Значения таких векторов не интерпретируемы: это “эмбединги” свойств пользователя и товара.

## Переход от частоты покупок к уверенности предпочтений

Все пропущенные значения матрицы заменим на 0. В этой связи необходима будут регуляризация, которую обсудим после того, как поговорим про обработку ненулевых элементов .

Далее , т.е. элементы матрицы бинаризуем, и от перейдём к по следующему принципу:



Во работах [[1]](#footnote-0) и [[2]](#footnote-1) бинаризация “предпочтений” давала более хороший результат в сравнении с моделью с реальным значениями (в целом во всех прочитанных статьях делается такой переход; в этих работах наглядно показана разница в показателях метрик). В случае перехода к бинарному представлению значений матрицы взаимодействий нам понадобятся дополнительные веса, отвечающие за степень уверенности в том, что товар пользователю по вкусу.

## Соберём функцию потерь

Перейдём к выводу функции потерь и отметим 2 её важных свойства.

Во-первых, мы заменили пропущенные значения нулями, и теперь имеем дисбаланс классов: ненулевых элементов, точнее говоря, элементов положительного класса (единиц) в матрице всего 14 070 857 из 6 937 772 930 (0.2%), поэтому необходимо дать бОльшие веса ненулевым (или для модели с бинарными значениями в матрице предпочтений, которая будет интересовать в первую очередь). Так, каждое слогаемое функции потерь взвесим на (или в некоторых русскоязычных источниках): , где подбирается кросс-валидацией, авторы метода рекомендуют брать . Авторы[[3]](#footnote-2) называют веса “уверенностью” в наблюдении : чем больше раз пользователь купил товар , тем с большей уверенностью можно говорить о том, что товар ему понравился. Мне кажется, что уместно сказать, что, задав таким образом веса, мы обращаем внимание, что ошибки, которые получаются на единицах, важнее ошибок на нулях.

Во-вторых, надо добавить регуляризацию для векторов матрицы со скрытыми свойствами пользователей и для товаров. Что на хабре, что в оригинальных статьях, все вносят обе нормы под один множитель , а не под разные, поэтому. Возможно, в дальнейшем следует попробовать разбить это составное слогаемое на 2 с разными множителями и , и тогда регуляризация будет иметь вид: .

Чтобы не отходить от оригинальных обозначений, обозначим за вектор скрытых предпочтений пользователя (т.е. строка матрицы в обозначениях выше), вектор скрытых свойств товара - за . Наконец, можем выписать функцию потерь для алгоритма:

Минимизация происходит сразу по двум переменным, поэтому вместо обычного градиентного спуска оптимизация будет выполняться с помощью ALS.

## Alternating least squares

Поочерёдно фиксируем то , то и дифференцируем по другой переменной. Это и есть ALS. После дифференцирования[[4]](#footnote-3) на каждом шаге итеративно обновляем значения в матрицах по формулам:

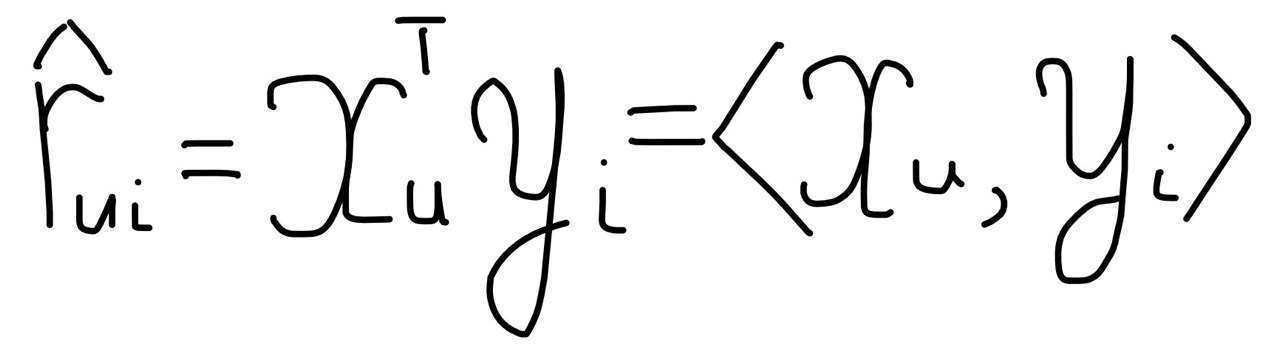
,

.

Оптимизация остановится, когда пройдёт установленное число итераций (в обеих работах за 10 шагов нашли оптимум).

## Таргет

Когда найдены для всех пользователей и для всех товаров (другими словами, заполнены матрицы и , с помощью которых можно восстановить исходную матрицу ), можем восстановить пропущенные значения:



Затем полученные значения сортируем по убыванию и в качестве рекомендуемых товаров берём первые K айтемов. Так, **таргетом в нашей задаче являются элементы матрицы взаимодействий** , часть из которых дана, а пропущенные спрогнозируем по формуле выше (с колпачком).

## Дополнительно про решение задачи

Алгоритм, рекомендующий самые популярные товары[[5]](#footnote-4), можно взять в качестве базового[[6]](#footnote-5), а пробовать его превосходить изложенной выше матричной факторизацией.

По кросс-валидации будут подбираться:

* - количество скрытых свойств товара и предпочтений юзера, которые будут использованы при разложении;
* - коэффициент во втором слагаемом весов , отвечающий за уверенность в предпочтениях пользователя;
* - коэффициент регуляризации. В целом, как отмечено выше, возможно следует попробовать регуляризовать с разными коэффициентами, и тогда подбирать будет надо и .

1. <http://yifanhu.net/PUB/cf.pdf> [↑](#footnote-ref-0)
2. <https://medium.com/radon-dev/als-implicit-collaborative-filtering-5ed653ba39fe> [↑](#footnote-ref-1)
3. Статья в сноске №1 [↑](#footnote-ref-2)
4. Выкладки можно посмотреть здесь: <https://habr.com/en/companies/prequel/articles/567648/>, причём в статье используются разные коэффициенты регуляризации для норм векторов. [↑](#footnote-ref-3)
5. Как, например, в статье, также приведённой выше: <http://yifanhu.net/PUB/cf.pdf> [↑](#footnote-ref-4)
6. <https://towardsdatascience.com/evaluation-metrics-for-recommender-systems-df56c6611093> [↑](#footnote-ref-5)