Санкт–Петербургский государственный университет
**Кафедра технологии программирования**

**Бельков Роман Андреевич**

**Выпускная квалификационная работа**

# Калибровка рекомендательных систем. Поиск гетерогенного эффекта и нетипичных пользователей для задач рекомендаций контента

Направление 01.03.02
«Прикладная математика и информатика»
Основная образовательная программа СВ.5005.2015 «Прикладная математика, фундаментальная информатика и программирование»

Научный руководитель:
кандидат техн. наук,
доцент
Блеканов И. С.

Рецензент:
старший преподаватель
Давыденко А. А.

Санкт-Петербург
2020 г.

# Содержание

# Введение

В настоящее время, количество информации растет очень быстрыми темпами и чтобы предоставлять наиболее релевантную и полезную для пользователей информацию, в веб сервисах начали использовать рекомендательные системы. Рекомендательные системы обеспечивают персонализированный пользовательский опыт во многих различных областях применения, включая интернет-магазины, социальные сети и потоковое воспроизведение музыки/видео. Рекомендательная система – это алгоритм, который предсказывает наиболее интересные объекты для конкретного пользователя на основе некоторой информации о нем.

Если пользователь посмотрел, 80 артхаусных фильмов и 20 комедий, то вполне разумно ожидать, что персонализированный список рекомендуемых фильмов будет состоять примерно из 80% артхаусных и 20% комедий. Это важное свойство, известное как калибровка, недавно получило новое внимание в контексте справедливости машинного обучения. В рекомендуемом списке элементов калибровка гарантирует, что различные области интересов пользователя будут отражены в соответствующих пропорциях. Калибровка особенно важна в свете того факта, что рекомендательные системы, оптимизированные в сторону точности в обычном автономном режиме, могут легко привести к рекомендациям, где меньшие интересы пользователя вытесняются основными интересами пользователя. В этой статье мы покажем, что рекомендательные системы, обученные точности, могут легко генерировать списки рекомендуемых элементов, которые фокусируются на основных областях интересов пользователя, в то время как меньшие области интересов пользователя, как правило, недопредставлены или даже отсутствуют. Со временем такие несбалансированные рекомендации несут в себе риск постепенного сужения областей интересов пользователя – что аналогично эффекту пузыря фильтров. Эта проблема также применима в случае нескольких пользователей, совместно использующих одну учетную запись, когда интересы менее активных пользователей в рамках одной учетной записи могут быть вытеснены в рекомендациях. Калибровка – это общая концепция машинного обучения, и в последнее время она переживает возрождение в контексте справедливо-

сти алгоритмов машинного обучения. Алгоритм классификации называется калиброванным, если прогнозируемые пропорции различных классов согласуются с фактическими пропорциями точек данных в имеющихся данных.

## Постановка задачи

Основная цель работы заключается в реализации метода калибровки рекомендательных систем, который бы не сильно ухудшал точность работы рекомендательной системы, но в то же время учитывал все интерес пользователя.

Для достижения цели были поставлены следующие задачи:

1. Обзор существующих методов калибровки рекомендательных систем.
2. Реализация алгоритма на языке Python.
3. Поиск или сбор данных и проверка реализованного метода на данных.

# Глава 1. Обзор литературы

## 1.1 Расстояние Кульбака-Лейблера

В ходе работы был проведен обзор литературы и найдено два метода потенциально подходящих для решения поставленной проблемы. Первый метод заключается в пересчете целевой метрики с помощью расстояния Кульбака-Лейблера (1). [1]

$$C_{KL}(p, q) = KL(p||\tilde{q}) = \sum_{g} p(g|u) \log \frac{p(g|u)}{\tilde{q}(q|u)}, \tag{1}$$

где $p$ это целевое распределение жанра $g$ для пользователя $u$, $q$ – полученное распределение жанров для пользователя. Во избежание случая $q(g|u) = 0$, будем использовать

$$\tilde{q}(q|u) = (1 - \alpha) \cdot q(g|u) + \alpha \cdot p(g|u)$$

с очень маленьким $\alpha > 0$, такое что $q \approx \tilde{q}$.

Сама же калибровка выполняется по формуле:

$$I^* = \arg \max_{I,|I|=N} (1 - \lambda) \cdot s(I) - \lambda \cdot C_{KL}(p, q(I)), \tag{2}$$

где $\lambda \in [0, 1]$, которая определяет компромисс между расстоянием Кульбака-Лейблера и значением метрики полученным рекомендательной системой. $s(I) = \sum_{i \in I} s(i)$, где $s(i)$ – степень уверенности, что фильм $i$ подойдет пользователю, предсказанная рекомендательной системой.

## 1.2 Метод Сент-Лагю

Второй алгоритм является адаптацией метода Сент-Лагю. [2] Метод Сент-Лагю, был изобретен французским математиком Андре Сент-Лагю для пропорционального распределения мандатов в правительстве. Суть метода заключается в поочередном присуждении мандатов партии с наибольшей квотой, которая на кажом шаге считается по формуле $\frac{V}{2s+1}$, где $V$ – количесвто голосов, полученных партией, s – количество мандатов, выделенных партии

на данном шаге.

Можно модифицировать данный метод под наш случай. Имея список наиболее релевантных фильмов, мы будем формировать новый список, выбирая фильмы по одному методом Сент-Лагю, только вместо партий у нас будут жанры, а голоса, полученные партией, заменятся на количество понравившихся пользователю фильмов конкретного жанра.

# Глава 2. Реализация

## 2.1 Набор данных

Мною был выбран датасет MovieLens 1M [3], включающий в себя 1 миллион оценок 4,000 фильмов от 6,000 пользователей.

## 2.2 Scikit SurPRISE

SurPRISE [4] – библиотека на языке Python, которая включает в себя реализацию построения и анализа рекомендательных систем. Из этого пакета, был использован метод SVD для построения рекомендательной системы.

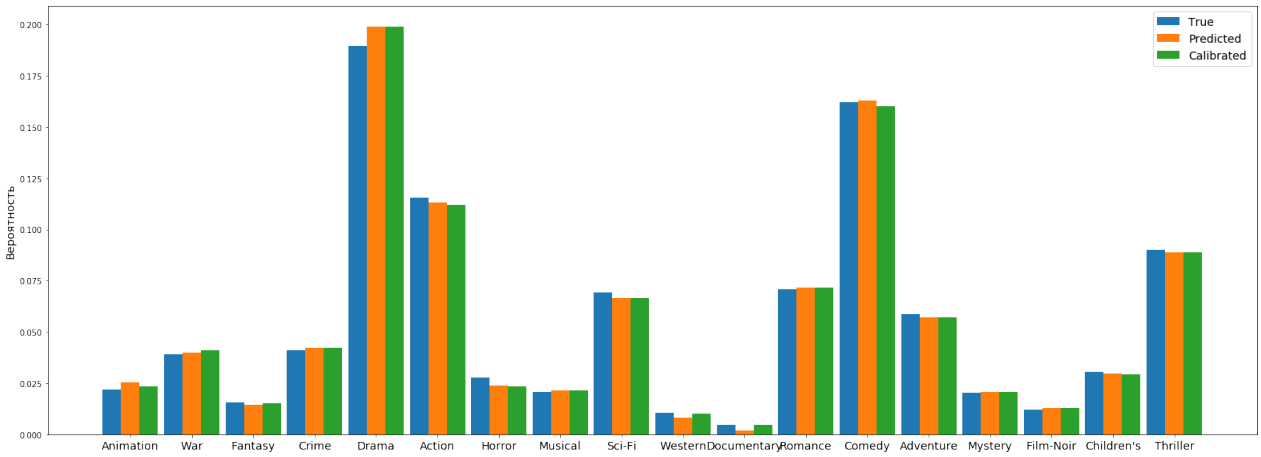


**Рис. 1:** Распределение жанров.

## 2.3 Результаты

Получив вектор предсказаний, я построил графики реального, предсказанного и откалиброванного распределения жанров фильмов для всех пользо-

вателей, они изображены на рисунке 1. Как видно на графике, распределение, полученное рекомендательной системой, плохо отражает такие жанры, как вестерн и документальное кино, а после калибровки распределение стало ближе к реальному. Калибровка была произведена с помощью расстояния Кульбака-Лейблера по формуле (2), с коэффициентом $\lambda = 0.2$.

## Список литературы


[1] Harald Steck «Calibrated Recommendations». RecSys '18: Proceedings of the 12th ACM Conference on Recommender Systems, 2018, pp. 154–162, doi.org/10.1145/3240323.3240372.

[2] Van Dang and W. Bruce Croft «Diversity by Proportionality: An Election-based Approach to Search Result Diversification». SIGIR '12: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, 2012, pp. 65-74, doi.org/10.1145/2348283.2348296.

[3] F. Maxwell Harper and Joseph A. Konstan «ACM Transactions on Interactive Intelligent Systems». ACM Trans. Interact. Intell. Syst. 5, 4, Article 19, 2015, doi.org/10.1145/2827872.

[4] Nicolas Hug «Surprise a Python library for recommender systems». 2017, surpriselib.com.