Использование согласованности разметок для статистической оценки качества рубрикатора

Федоров Артем Максимович

October 2024

Аннотация

Данная статья предлагает новый подход к анализу качества рубрикаторов, независящий от домена рассматриваемых объектов. Выдвигается предположение о мультимодальной природе распределения объектов рубрик в латентных пространствах современных моделей-векторизаторов на основе механизма внимания и предлагается пайплайн решения таких задач через восстановление смесей распределений в структурированном пространстве. Выводятся оценки разделимости распределений смесей von-Mises-Fisher семейства и проводятся статистические выкладки их устойчивости.

1 Вступление

В различных задачах рубриктор может носить различные имена: тематизатор для задач подразделения текстовых массивов по критерию их содержания [9, 14], ценностный или культурный код для описания систем ценностей в языке, классификатор для задач дискриминатного анализа в машинном обучении. В общем случае задача создания рубрикатора подразумевает под собой построение системы классов для некторой категории объектов, что могла бы приблизить ее природную структуру, при помощи которого объект мог бы быть описан совокупностью рубрик, к которым он принадлежит.

Тем не менее, основной проблемой создания рубрикатора является его потимальность: из прикладных целей ожидается, что рубрики будут между собой непересекающимися, легко различимыми и полными по своей наполненности. Первые работы на тему создания оптимального рубрикатора приводились в начале 2000 годов при изучении сложных объектов, требующих соответствуенно сложную стректуру классов для их описания, таких как язык и категоризация текстовых массивов [10, 24, 1]. Такие работы основывались на мнении экспертов соответсвующей области [10, 34], либо на применении простейших алгоритмов машинного обучения (LDA [28, 29] или же SVD [24]).

Тем не менее с развитием машинного обучения и все большим его применением в новых областях, создание рубрикатора стало применяться в качестве составной части конечного задания или решения, где полагаться на конструктивный подход организации рубрикатора стало невозможно либо из семантики задачи (автоматическое создание рубрикатора), либо по причине потребности в валидации его качества в бизнес целях. В данном контексте оценка качества должна быть нацелена на тесную работу с экспертами области для доведения руббрикатора до оптимальной в контексте задачи формы.

Несмотря на то, что данная область машинного обучения появилась на заре образования nlp, большинство методов получения мер качесвте рубрикаторов являются слишком невыразительными: если они и способны показать конечную метрику, то под возможность дать взглянуть на саму проблему при ее наличии и проанализировать ее – они не заточены. Однако возможность получить качесвтенный анализ получаемого рубрикатора является важной составляющей создания заданий для крауд сорс платформ (как составить интуитивно понятное задания для разметчиков) или в задачах автоматического построения рубрикаторов, для анализа уместности используемых подходов.

Данная работа является результатом решения потребности лаборатории Семантического анализа текстов Московского Государственного Университета им. М. В. Ломоносова в новой универсальной метрике качества рубрикатора, применяемом в проектах по разметке корпусов текстов, требующих сложные и большие по количеству системы рубрик (для разметки ценностей, эмоций, поляризации мнений в текстах и т. д.), позволяющей анализировать набор рубрик, найти в нем повторяющие друг друга элементы, выявить элементы, которых недостает.

2 Связанная работа

Задача оценки качества рубрикатора возникает в первую очередь в областях машинного обучения, где создание структуры классов является составной частью решения. К таким задачам в первую очередь относятся пайплайны крайдсорсинга и применения активного обучения [32, 21], где она выступает в качестве одной из ступеней валидации, в выявлении скрытой структуры данных через тематическое моделирование, [35, 37], где она сводится к оценке результата unsupervised или semi-supervised методов.

Методы оценки качества рубрикатора конструктивно строятся на векторизации данных. Классическим способом например получения аналитического

представления текстовых данных считается TF-IDF, широко применяемый на заре рубрикации данных [37, 36], применение методов PCA [17] и ICA [16]. Появляение же в последнее время сильных моделей векторизации данных различных доменов (текст, фото, звук, видео) ввиду развития нейросетевого подхода позволило приблизить решение задач компьютером к уровню разметчиков людей, позволяя рассматривать задачу оценки качества рубрикатора с изначальных пространств объектов (текст, видео, прочая медиа) на изучение многообразия объектов в латентном пространстве модели (будь то BERT [15] или clip [25] поднобный энкодер).

2.1 Особенности областей применения

В автоматической рубрикации данных [37, 33] оценка качества результата носит валидационный характер. В таких задачах постановка либо включает в себя предположительный набор рубрик с малым числом представителей, либо подразумевает выявление структуры классов с нуля с последующей передачей результатов разметчикам.

В задачах активного обучения и краудсорсинга основополагающим требованием является эффективность работы разметчиков на платформе [32], что диктуется прикладной областью (стоимость разметки для бизнеса и исследователей). Создание хорошо понимаемого и эффективно решаемого разметчиком задания является первой ступенью для удовлетворения данного требования. В таком случае, оценка рубрикатора классов, используемого разметчиками при обработке примеров, отражает согласованность разметчиков, насколько каждый из них пособен понимать устройство рубрикатора и определять соответсвие объектов предложенными рубриками. Задачи связанные с активной разметкой все чаще начали допускать разметчиков к самостаятельному определению множества рубрик, что наблюдается в исследовательских задачах при анализе работы LLM [13, 11, 7, 4], где оценка качества применяется на размеченных данных, или же ценностной наполненности языков.

2.2 Обзор известных методов

Такие задачи классически решаются методами восстановления класетрной структуры объектов, таких как k-Means [27, 18, 22], иерархическая кластеризация (чаще всего с расстоянием Уорда, максимальным или минимальным между подмножествами объектов) [19, 30], либо же алгоритмы восстановления плотностей в латентном пространстве, такие как MeanShift [6, 5] и DBSCAN [8]. Для первого подхода постановки задачи определения рубрикатора с semi-supervised данными, где для подмножества рубрик известны характерные представители, качество рубрикатора оценивается простейшими внешними метриками классификации PrecisionBCubed, RecallBCubed и так далее. В то же время для unsupervised подхода используются внутренние меры качества: кофенетический коэффициент корреляции (CPCC), индексы следа, Калинского или Гарабача [2, 20]

2.3 Противопоставление нашего подхода

Каждый из затронутых ранее подходов рассматривает векторизованные объекты как сгенерированную выборку из некоторой смеси распределений, каждая компонента которой соответсвует одной рубрике. Такой подход позволяет упростить модель данных в латентном пространстве и открывает путь к применению классических метрик качества для класетризации. Тем не менее утакого подхода есть ряд недостатков:

1. Предположение, что рубрики соответсвуют одномодальным распределениям, сильно ограничивает выразительность используемой модели.

2. Информация о том, что объекты были соотнесены разным рубрикам в процессе разметки используется при оценке качества рубрикатора не напрямую, но лишь является одним из параметров метода кластеризации, что приводит к невыразительности суждения о качетсве рубрикатора в целом.

Изучая устройство размеченных выборок для задач определения культурных ценностей, эмоций и полярностей мнений, мы имеем основания полагать, что распределения рубрик являются мультимодальными. И проведение кластеризации с последующим применением метрик качества не способно в действительности описать качество рубрикатора.

Наш подход предлагает новый взгляд. Мы считаем, что качественно обученный векторизатор способен хорошо описывать структуру объектов, позволяя оценивать уверенность разделимости рубрик на объекте мерой разделимости двух соответсвующих распределений в пространстве эмбеддингов. Тогда по размеченной выборке объектов мы способны восстановить кластерную структуру каждого класса при помощи EM алоритма [3], получить оценки максимального правдоподобия на параметры распределений и далее высчитать оценки разделимостей смесей, что можно интерпретировать в статистическом анализе.

Данный подход является логическим продолжением ROC анализа [23] в классической литературе по SignalDetection [12], и предполагает вывод нового индекса разделимости, ведущего себя как d' [26, 31], но учитывающего моменты до второго порядка распределений.

3 Анализ задачи

Мы ограничиваем наш спектр задач рассмотрением ситуации, где эксперты уже смогли составить рубрикатор, по которому разметчики смогли составить размеченную выборку.

Список литературы

[1] IA Bolshakov and A Gelbukh. Classification of collocations in databases by meaning of combined words. *AUTOMATIC DOCUMENTATION AND MATHEMATICAL LINGUISTICS TRANSLATIONS OF SELECTED ARTICLES FROM NAUCHNO-TEKHNICHESKAIA INFORMATSIIA*, 34(3):64–74, 2000.

[2] Francois Boutin and Mountaz Hascoët. Cluster validity indices for graph partitioning. In *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.*, pages 376–381. IEEE, 2004.

[3] Gilles Celeux and Gérard Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational statistics & Data analysis*, 14(3):315–332, 1992.

[4] Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17817–17825, 2024.

[5] Zi Li Chen. Research and application of clustering algorithm for text big data. *Computational Intelligence and Neuroscience*, 2022(1):7042778, 2022.

[6] Dorin Comaniciu and Peter Meer. Mean shift analysis and applications. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1197–1203. IEEE, 1999.

[7] Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. *arXiv preprint arXiv:2104.08696*, 2021.

[8] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, pages 226–231, 1996.

[9] Akhmedov Farkhod, Akmalbek Abdusalomov, Fazliddin Makhmudov, and Young Im Cho. Lda-based topic modeling sentiment analysis using topic/document/sentence (tds) model. *Applied Sciences*, 11(23):11091, 2021.

[10] Inna E Gendlina. The russian rubricator: A unified system of classificatory indexing languages. *KO KNOWLEDGE ORGANIZATION*, 19(3):126–130, 1992.

[11] Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. *arXiv preprint arXiv:2012.14913*, 2020.

[12] Michael J Hautus, Neil A Macmillan, and C Douglas Creelman. *Detection theory: A user's guide*. Routledge, 2021.

[13] Jinwen He, Yujia Gong, Zijin Lin, Yue Zhao, Kai Chen, et al. Llm factoscope: Uncovering llms' factual discernment through measuring inner states. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 10218–10230, 2024.

[14] Swapnil Hingmire, Sandeep Chougule, Girish K Palshikar, and Sutanu Chakraborti. Document classification by topic labeling. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 877–880, 2013.

[15] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota, 2019.

[16] Wei Lu and Jagath C Rajapakse. Approach and applications of constrained ica. *IEEE transactions on neural networks*, 16(1):203–212, 2005.

[17] Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

[18] J MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability/University of California Press*, 1967.

[19] Christopher Manning and Hinrich Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

[20] Ujjwal Maulik and Sanghamitra Bandyopadhyay. Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on pattern analysis and machine intelligence*, 24(12):1650–1654, 2002.

[21] David A McConnell, David N Steer, and Kathie D Owens. Assessment and active learning strategies for introductory geology courses. *Journal of Geoscience Education*, 51(2):205–216, 2003.

[22] MES Mendes and Lionel Sacks. Dynamic knowledge representation for e-learning applications. In *Enhancing the Power of the Internet*, pages 259–282. Springer, 2004.

[23] Charles E Metz. Basic principles of roc analysis. In *Seminars in nuclear medicine*, volume 8, pages 283–298. Elsevier, 1978.

[24] Arturo Montejo Ráez et al. Automatic text categorization of documents in the high energy physics domain. 2005.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[26] Barbara Sakitt. Indices of discriminability. *Nature*, 241(5385):133–134, 1973.

[27] Nayani Sateesh, Kuljeet Kaur, M Lakshminarayana, Vipul Vekariya, Harshal Patil, and Ramya Maranan. Development of a gui for automated classification of scientific journal articles using clustering. In *2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6. IEEE, 2024.

[28] Xing Wei and W Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, 2006.

[29] Dejian Yu and Bo Xiang. Discovering topics and trends in the field of artificial intelligence: Using lda topic modeling. *Expert systems with applications*, 225:120114, 2023.

[30] Oren Eli Zamir. *Clustering web documents: a phrase-based method for grouping search engine results*. University of Washington, 1999.

[31] Weiwen Zou and Pong C Yuen. Discriminability and reliability indexes: two new measures to enhance multi-image face recognition. *Pattern Recognition*, 43(10):3483–3493, 2010.

[32] Руслан Айдарович Гилязев and Денис Юрьевич Турдаков. Активное обучение и краудсорсинг: обзор методов оптимизации разметки данных. *Труды Института системного программирования РАН*, 30(2):215–250, 2018.

[33] М Дли, О Булыгина, П Козлов, and В Борисов. *Rubrication of text documents based on fuzzy difference relations*. Litres, 2022.

[34] Татьяна Сергеевна Ильина and Татьяна Сергеевна Ившина. Выявление духовных ценностей в текстах семейных преданий. *Балтийский гуманитарный журнал*, 7(3 (24)):41–45, 2018.

[35] ЯБ Калачев and АН Сибирмовская. Создание кластерного пространства текстовых документов в базе данных. *Новые информационные технологии в автоматизированных системах*, (13):130–133, 2010.

[36] ОВ Пескова. Методы автоматической классификации текстовых электронных документов. *Научно-техническая информация. Серия 2: Информационные процессы и системы*, (3):13–20, 2006.

[37] Ольга Вадимовна Пескова. Автоматическое формирование рубрикатора полнотекстовых документов. In *Тр. Десятой Всерос. науч. конф.«Электронные библиотеки: перспективные методы и технологии, электронные колекции»(RCDL'2008).—Дубна*, pages 139–148, 2008.

Appendices

A Выбор семейства распределений

Из предыдущих размышлений мы пришли к выводу, что многомодальное распределение сложной формы можно приближать смесью простых унимодальных распределений. Согласно условиям на пространство смыслов, каждая мода представляет собой одну идею в классе, выступая в роли прототипа в определенной точке. Соответственно, чем ближе объект к этому прототипу, тем выше вероятность того, что он был сгенерирован из данной части смеси. Вокруг таких мод распределений будет наблюдаться сгущение точек, которое быстро убывает по мере удаления от центра. В качестве моделирующего распределения очень удобно использовать Гауссовское. Однако область определения такового является все пространство $\mathbb{R}^{n-1}$ для записи в полярных координатах, а потому видоизменим его при помощи ранее записанных преобразований. В качестве примера, рассмотрим запись для одномерного случая окружности в пространстве $\mathbb{R}^2$:

$$\begin{cases} WN(\theta;\mu,\sigma) \sim \dfrac{1}{\sigma\sqrt{2\pi}} \displaystyle\sum_{k=-\infty}^{\infty} \exp\left[\frac{-(\theta-\mu+2\pi k)^2}{2\sigma^2}\right] \\ f_{WN}(\theta;\mu,\sigma) = \dfrac{1}{2\pi}\vartheta\left(\dfrac{\theta-\mu}{2\pi}, \dfrac{i\sigma^2}{2\pi}\right) \\ \vartheta(\theta,\tau) = \displaystyle\sum_{n=-\infty}^{\infty} (w^2)^n q^{n^2}, \text{ где } w \equiv e^{i\pi\theta} \\ \vartheta(\theta,\tau) - \text{Тэта функция Якоби} \end{cases}$$

Расчет подобных формул представляется неэффективным в контексте высоко нагруженных задач даже для простейшего случая. При этом положение лишь усугубляется при переходе к $n > 2$ мерному случаю. Потому поставим для подходящего распределения, каким оно должно быть для использования:

1. Распределение должно быть унимодальным с центром в моде.
2. При удалении от центра плотность должна стремительно убывать к нулю.
3. Распределение подобно Гауссовскому должно обладать параметром кучности.

Таковым распределением предлагается рассматривать известное распределение из циркулярной статистики, приближающее Гауссовское на сфере — Распределение **von Mises-Fisher**.

Определим поверхность сферы как пространство элементарных исходов. Важнейшей особенностью является конечность меры поверхности. Тогда элементарным приращением $dw = dx_1 \times ... \times dx_{d-1} = dx$, что через якобиан переходит в углы по формуле: $dw = \left(\prod_{i=2}^{p-2} \sin^{p-i}\theta_{i-1}\right) d\theta$. А само распределение будет выводиться, как зависящее от параметра кучности k и центра распределения μ:

$$\begin{cases} p(x|\mu,k) = c_n(k)e^{k\langle\mu,x\rangle} \\ c_n(k) = \dfrac{k^{n/2-1}}{(2\pi)^{n/2} I_{n/2-1}(k)} \\ I_n(k) = \left(\dfrac{k}{2}\right)^n \displaystyle\sum_{t\geq 0} \frac{1}{\Gamma(n+t+1)t!}\left(\frac{k}{2}\right)^{2t} \end{cases}$$

- k является параметром кучности распределения, и чем он ближе к нулю, тем более равномерно распределение.
- μ является центром распределения.
- $c_n(k)$ есть нормировочная константа для распределения.
- С ростом размерности n резко увеличиваются проблемы с высчитываем параметров из-за float-point вычислений

Для такого распределения первый момент $\mathbb{E}x = M_{n/2-1}(k)\mu$, где $M_{n/2-1}(k) = \frac{I_{n/2}(k)}{I_{n/2-1}(k)}$ есть функция от параметра k и размерности пространства. В свою очередь данная функция сложна в вычислении для больших размерностей n, что решается далее в статье.

B Вывод формул для индексов разделимости

B.1 Метрика Kullback-Leibler

Рассмотрим, чему в общем случае равна дивергенция по поверхности сферы:

$$KL(f_1\|f_2)=\int_{\mathbb{S}_{p-1}} f_1(x|k_1,\mu_1)\log\left(\frac{f_1(x|k_1,\mu_1)}{f_2(x|k_2,\mu_2)}\right)dx=\int_{\mathbb{S}_{p-1}}\left(\langle k_1\mu_1,x\rangle-\langle k_2\mu_2,x\rangle+\log\left(\frac{c_n(k_1)}{c_n(k_2)}\right)\right)f_1(x|k_1,\mu_1)dx$$

Тогда интегрирование обеих частей дает систему:

$$\begin{cases}\displaystyle\int_{\mathbb{S}_{p-1}}\langle k_1\mu_1-k_2\mu_2,x\rangle\, f_1(x|k_1,\mu_1)dx=\langle k_1\mu_1-k_2\mu_2,\mathbb{E}x_1\rangle\\ \displaystyle\int_{\mathbb{S}_{p-1}}\log\left(\frac{c_n(k_1)}{c_n(k_2)}\right)f_1(x|k_1,\mu_1)dx=\log\left(\frac{c_n(k_1)}{c_n(k_2)}\right)\end{cases}$$

Тогда получим выражение дивергенции ($\eta=n/2-1$):

$$KL(f_1\|f_2)=\eta\log\left(\frac{k_1}{k_2}\right)-\log\left(\frac{I_\eta(k_1)}{I_\eta(k_2)}\right)+r_\eta(k_1)\,\langle k_1\mu_1-k_2\mu_2,\mu_1\rangle$$

B.2 Метрика χ^2

$$\chi^2(f_1\|f_2)=\int_{\mathbb{S}_{p-1}}\frac{\left(f_1(x|k_1,\mu_1)-f_2(x|k_2,\mu_2)\right)^2}{f_2(x|k_2,\mu_2)}dx$$

Раскрывая скобки, упростим интеграл в условиях независимости распределений. Обозначим $\eta=2/n-1$:

$$\chi^2(f_1\|f_2)=\int_{\mathbb{S}_{p-1}}\frac{\left(f_1(x|k_1,\mu_1)\right)^2}{f_2(x|k_2,\mu_2)}dx-1\Rightarrow\frac{\left(f_1(x|k_1,\mu_1)\right)^2}{f_2(x|k_2,\mu_2)}=\exp\left\{\langle 2k_1\mu_1-k_2\mu_2,x\rangle\right\}\frac{c_\eta^2(k_1)}{c_\eta(k_2)}$$

Заметим, что получаемое выражение повторяет запись для распределения **von Mises-Fisher** со специальными параметрами: $\overline{\mu}=\frac{(2k_1\mu_1-k_2\mu_2)}{\|2k_1\mu_1-k_2\mu_2\|_{\ell^2}}$; $\overline{k}=\|2k_1\mu_1-k_2\mu_2\|_{\ell^2}\Longrightarrow$

$$\chi^2(f_1\|f_2)=\frac{c_n^2(k_1)}{c_n(k_2)c_n(\overline{k})}-1=\frac{k_1^{2\eta}I_\eta(k_2)I_\eta(\overline{k})}{k_2^\eta\overline{k}^\eta I_\eta^2(k_1)}-1$$

C Вычисление функции Бесселя

Модернизированная первая функция бесселя представима в виде степенного ряда:

$$I_n(k)=\left(\frac{k}{2}\right)^n\sum_{t\geq 0}\frac{1}{\Gamma(n+t+1)t!}\left(\frac{k}{2}\right)^{2t}$$

Точное получение значения такой функции затруднительно с точки зрения достижения вычислительной устойчивости, так как в степенном ряду для каждого члена приходится высчитывать значение дроби (на практике) со стремящимися к бесконечности знаменателем и числительным. При этом подсчет точного значения гамма-функции может быть вычислительно неэффективным при больших n. В данном случае мы стремимся упростить выражение, разбить его на части, чтобы избежать неустойчивых выражений при вычислении индексов i. Для этого, используя свойство Гамма-функции $\Gamma(x+1)=x\cdot\Gamma(x)$ и ее аппроксимацию Стирлинга, упростим выражение:

$$\begin{cases}\displaystyle I_n(k)=\frac{k^n}{2^n\Gamma(n)}\sum_{t\geq 0}\frac{1}{t(t+1)(t+2)...(t+n)t!}\left(\frac{k}{2}\right)^{2t}=\frac{k^n}{2^n\Gamma(n)}\cdot Tail(n,k)\\ \displaystyle Tail(n,k)=\sum_{t\geq 0}\frac{1}{t(t+1)(t+2)...(t+n)t!}\left(\frac{k}{2}\right)^{2t}\\ \displaystyle\Gamma(n)=\left(\frac{n}{e}\right)^n\sqrt{\frac{2\pi}{n}}\left(1+\frac{1}{12n}+\frac{1}{288n^2}+\overline{o}\left(\frac{1}{n^2}\right)\right)\end{cases}$$

Тогда высчитывание функции $I_n(k)$ производится итерационным способом. Член степенного ряда получается из предыдущего умножением на $\frac{k^2/4}{(t+1)(n+t+1)}$. Тогда получаем алгоритм подсчета, представленный ниже:

Algorithm 1 Алгоитм вычисления первой модифицированной функции Бесселя $I_n(ka)$

Input n $\gg 1$, $k \geq 0$
T $\leftarrow 1.0$
$A' \leftarrow \left(\frac{ke}{2n}\right)^n \cdot \sqrt{\frac{n}{2\pi}}$
$A \leftarrow A' / \left(1 + \frac{1}{12n} + \frac{1}{288n^2}\right)$
$Tail \leftarrow 0; \quad t \leftarrow 1$
while $\Delta T > \varepsilon(A)$ **do**
  $T \leftarrow T \cdot \frac{k^2/4}{(t+1)(n+t+1)}$
  $Tail \leftarrow Tail + T$
  $t \leftarrow t + 1$
end while
return $A \cdot Tail$

D Стабильное вычисление индексов разделимости

D.1 Вычисление KL дивергенции

Вернемся к записи $KL(f_1\|f_2)$ и применим представление для $I_n(k)$, взяв $\eta = n/2 - 1$

$$KL(f_1\|f_2) = \eta \log\left(\frac{k_1}{k_2}\right) - \log\left(\frac{I_\eta(k_1)}{I_\eta(k_2)}\right) + r_\eta(k_1)\left\langle k_1\mu_1 - k_2\mu_2, \mu_1\right\rangle$$

Рассмотрим второй член выражения:

$$\log\left(\frac{I_\eta(k_1)}{I_\eta(k_2)}\right) = \log\left(\frac{\left(\frac{k_1 e}{2\eta}\right)^\eta \frac{\sqrt{\eta/2\pi}}{\left(1+\frac{1}{12\eta}+\frac{1}{288\eta^2}\right)} Tail(\eta, k_1)}{\left(\frac{k_2 e}{2\eta}\right)^\eta \frac{\sqrt{\eta/2\pi}}{\left(1+\frac{1}{12\eta}+\frac{1}{288\eta^2}\right)} Tail(\eta, k_2)}\right) = \eta \log\left(\frac{k1}{k2}\right) + \log\left(\frac{Tail(\eta, k_1)}{Tail(\eta, k_2)}\right)$$

Рассмотрим представление $r_\eta(k_1)$ и выведем ее ассимптотическое поведение при $\eta \to \infty$:

$$r_\eta(k_1) = \frac{I_{\eta+1}(k_1)}{I_\eta(k_1)} = \left(\frac{k_1 e}{2(\eta+1)}\right)^{\eta+1} \cdot \left(\frac{2\eta}{k_1 e}\right)^\eta \cdot \sqrt{\frac{\eta+1}{\eta}} \cdot \frac{\left(1 + \frac{1}{12\eta} + \frac{1}{288\eta^2}\right)}{\left(1 + \frac{1}{12(\eta+1)} + \frac{1}{288(\eta+1)^2}\right)} \cdot \frac{Tail(\eta+1, k_1)}{Tail(\eta, k_1)} =$$

$$= \frac{k_1 e}{2(\eta+1)} \cdot \left(1 - \frac{1}{\eta}\right)^\eta \cdot \sqrt{1 + \frac{1}{\eta}} \cdot \frac{\left(1 + \frac{1}{12\eta} + \frac{1}{288\eta^2}\right)}{\left(1 + \frac{1}{12(\eta+1)} + \frac{1}{288(\eta+1)^2}\right)} \cdot \frac{Tail(\eta+1, k_1)}{Tail(\eta, k_1)} \xrightarrow[\eta\to\infty]{} \frac{k_1}{2(\eta+1)} \frac{Tail(\eta+1, k_1)}{Tail(\eta, k_1)}$$

Тогда в предположении $n \gg 1$ полная формула получения $KL(f_1\|f_2)$ будет иметь вид:

$$KL(f_1\|f_2) \cong -log\left(\frac{Tail(\eta, k_1)}{Tail(\eta, k_2)}\right) + \frac{k_1}{2(\eta+1)} \frac{Tail(\eta+1, k_1)}{Tail(\eta, k_1)} \left\langle k_1\mu_1 - k_2\mu_2, \mu_1\right\rangle$$

$$\Longrightarrow i_{KL} = \frac{k_1}{2(\eta+1)} \frac{Tail(\eta+1, k_1)}{Tail(\eta, k_1)} \left\langle k_1\mu_1 - k_2\mu_2, \mu_1\right\rangle + \frac{k_2}{2(\eta+1)} \frac{Tail(\eta+1, k_2)}{Tail(\eta, k_2)} \left\langle k_2\mu_2 - k_1\mu_1, \mu_2\right\rangle$$

D.2 Вычисление χ^2 дивергенции

Так же рассмотрим запись χ^2 дивергенции, и упростим, используя формулы для $I_n(k)$

$$\chi^2(f_1\|f_2) = \frac{c_n^2(k_1)}{c_n(k_2)c_n(\overline{k})} - 1 = \frac{k_1^{2\eta} I_\eta(k_2) I_\eta(\overline{k})}{k_2^\eta \overline{k}^\eta I_\eta^2(k_1)} - 1$$

Тогда подставим формулу $I_n(k)$:

$$\chi^2(f_1\|f_2) = \frac{k_1^{2\eta}}{k_2^\eta \overline{k}^\eta} \cdot \frac{\left(\frac{k_2 e}{2\eta}\right)^\eta \left(\frac{\overline{k} e}{2\eta}\right)^\eta Tail(k_2, \eta) Tail(\overline{k}, \eta)}{\left(1 + \frac{1}{12\eta} + \frac{1}{288\eta^2}\right)^2} \cdot \frac{\left(1 + \frac{1}{12\eta} + \frac{1}{288\eta^2}\right)^2}{\left(\frac{k_1 e}{2\eta}\right)^{2\eta} Tail^2(k_1, \eta)} = \frac{Tail(k_2, \eta) Tail(\overline{k}, \eta)}{Tail^2(k_1, \eta)}$$

И соответственный индекс разделимости распределений i_{χ^2} примет вид:

$$\Longrightarrow i_{\chi^2} = \frac{Tail(k_2, \eta)Tail(\overline{k}, \eta)}{Tail^2(k_1, \eta)} + \frac{Tail(k_1, \eta)Tail(\overline{k}, \eta)}{Tail^2(k_2, \eta)}$$