Курсовая работа за 3 курс ВМК МГУ имени М.В. Ломоносова на тему: "Анализ мультимодальных распределений в пространстве эмбеддингов, использование параметрических распределений von Mises–Fisher."

Федоров Артем Максимович

Содержание

1 Введение

Прикладные задачи, решаемые исследователями и инженерами, обладают уникальной природой, диктующей желаемый результат, семантику решения и особенности входных данных. Однако эта природа не всегда облегчает процесс решения. Наоборот, она может стать одной из сложнейших проблем для специалистов. Неинтерпретируемость компьютером, неформализуемость и высокая сложность данных, которые на первый взгляд легко понимаемы людьми, приводят к неприменимости классических методов решения. В связи с этим первостепенным становится создание интерпретируемых представлений для таких объектов, которые далее будут так же упоминаться как "эмбеддинги". Будем считать, что получаемые эмбеддинги являются объектами n-мерного вещественного пространства $\mathbb{R}^n$ с заданной евклидовой метрикой и соответствующим скалярным произведением. Рассмотрим задачу поиска этого отображения в общем случае.

Эмбеддинги должны адекватно моделировать исходные объекты, ввиду чего не каждое отображение будет удовлетворять условиям соответствия. Дополнительно предположим, что исходные объекты обладают свойством интерпретируемости человеком: человек может различать объекты по выделяющимся чертам,

сравнивать их между собой и находить сходства. Следовательно, логично требовать от искомого оператора $A : K \to \mathbb{R}^n$, переводящего описанную фиксированную категорию объектов K, объединяемых единой природой, в искомые эмбеддинги из $\mathbb{R}^n$, соответствовать трем условиям:

1. Оператор моделирует объекты в пространстве эмбеддингов посредством описания их черт.
2. Обратный оператор перевода объектов определен быть не обязан, однако описание пытается запечатлеть характерные черты объекта, позволяя различать эмбеддинги так же, как это делает человек с первоначальными объектами, обеспечивая возможность их сравнения и нахождения сходств.
3. Оператор восстанавливает семантические группы объектов в новом пространстве.

Последнее условие качественно отличается от предыдущих двух, хотя и частично из них следует. Его выполнение даёт исследователям возможность не только анализировать единичные представления объектов и сравнивать их между собой, но и судить о зависимостях между объектами по распределениям эмбеддингов в моделирующем пространстве. В этом случае само пространство можно интерпретировать как пространство смыслов, где объектами являются "идеи"из некоторых распределений, поддающихся анализу классическими методами.

И хотя на практике нельзя говорить о точном существовании такого "универсального"оператора и пространства смыслов в целом, методы машинного обучения предлагают собственные подходы к построению его приближения, от которого и будет зависеть дальнейший анализ структуры исследуемых объектов, переходящий в анализ получаемых распределений. Здесь возникает первый переломный момент: большинство исследований, решающих поставленную задачу, стремятся максимально упростить результирующее пространство, отказываясь от мультимодальных распределений в пользу унимодальных и отбрасывая большую часть полезной информации, заложенной в первоначальном представлении.

Идея данного исследования заключается в том, что локальное сходство мультимодальных распределений может дополнять наши представления о них, служа сигналом наличия условной зависимости между объектами при определённых обстоятельствах, не утверждая ничего о глобальном положении дел. Таким образом, утверждается:

- Классы (области смыслов) факторизуются на простые идеи, совокупность которых и дает весь его образ.
- Подтемы даже разных классов способны быть схожи между собой, что в общем случае не говорит о схожести самих классов.
- Распределения классов представимы в виде мультимодальных распределений, что в свою очередь представимы в виде смесей более простых.
- Анализ мод распределений и попарных расстояний между ними способен дать полезную информацию не только о качестве представлений оператора перевода, но и об изначальной задаче и системе классов.

Для задач анализа и классификации текста примерами такого подхода могут выступать слова "сильный"и "злой "доблестный"и "добрый что в текстах определенной тематики имеют сильную корреляцию, когда сами они не являются схожими по смыслу. В данном исследовании рассматривается способ параметрического приближения таких мультимодальных распределений для последующего анализа, исследуются и разрешаются проблемы, возникающие при построении таких распределений, а так же исследуется подход имплементации аппарата в проект.

2 Аналитическое представление распределений

Первостепенным шагом является восстановление распределений классов в результирующем пространстве. Будем считать, что первоначальная выборка известных объектов категории K порождена некоторыми классами T, для которых определены соответствующие функции распределения над объектами $\hat{\mathbb{P}}_K$. Таким образом, наша выборка является реализацией некоторых случайных величин.

Потребуем от оператора $A : K \to \mathbb{R}^n$, переводящего изначальные объекты, поделенные на классы, в интерпретируемое компьютером пространство, отвечать условию измеримости относительно первоначального измеримого вероятностного пространства и всем условиям, описанным во вступлении: схожие по смыслу объекты он должен поставить ближе, чем фрагменты с противоположным или контрастирующим содержанием (гипотеза компактности в пространстве смыслов). Тогда действие этого оператора на случайную величину будет случайной величиной, а распределения в новом пространстве смыслов останутся информативными. Задача восстановления распределения над классами обобщается на задачу восстановления плотности распределения эмбеддингов фрагментов заданного класса в пространстве смыслов.

Здесь возникают две первоначальных проблемы:

1. Оператор A (из эмпирического опыта) представляет собой очень сложное преобразование, вывод влияния которого на аналитическое распределение над объектами K либо невозможен, либо крайне затруднителен. Если же данные распределения $\hat{\mathbb{P}}_K$ не известны, либо сам оператор A является черным ящиком, то сам подход аналитического выведения распределений не подходит.

2. Размерность вещественного пространства $\mathbb{R}^n$ может быть крайне большим, что является ожидаемым результатом попытки наиболее точно вобрать свойства моделируемых объектов. В таком случае методы восстановления распределений такие как Монте-Карло, что требуют экспоненциально возрастающую по объему выборку от размерности пространства так же не подходят.

Таким образом наиболее подходящим методом восстановления мультимодальных распределений, знания о которых мы получаем из реализации выборки в $\mathbb{R}^n$, является приближение смесью парамметрически заданных унимодальных распределений.

2.1 Распределения на сфере

Теперь построим пространство эмбеддингов и потребуем от получаемых представлений быть нормированными на единичную гиперсферу. Такое ограничение обусловлено получаемым упрощением интерпретации модуля эмбеддинга, особенно при последующих преобразованиях в пространстве смыслов, а так же из вычислительной эффективности подсчета косинусной схожести и расстояния для векторов с нормой равной 1.

$$A : K \to \mathbb{S}_{n-1}, \text{где } \mathbb{S}_{n-1} = \{x \in \mathbb{R}^n \mid \|x\|_{l^2} = 1\}$$

Будем считать, что для распределений в новом пространстве определена плотность, а значит восстановление распределений происходит не во всем пространстве $\mathbb{R}^n$, а лишь на поверхности сферы единичного радиуса, что выражается в $p_K(x : \|x\| \neq 1) = 0$ почти всюду в $\mathbb{R}^n$, что позволяет нам перейти от декартовых координат к полярным:

$$\begin{cases} \mathbb{S}_{n-1} : \sqrt{x_1^2 + x_2^2 + ... + x_n^2} = 1 \\ x_1 = r\cos\theta_1 \\ x_2 = r\sin\theta_1\cos\theta_2 \\ ... \\ x_k = r\sin\theta_1\sin\theta_2...\sin\theta_{k-1}\cos\theta_k \\ ... \\ x_n = r\sin\theta_1\sin\theta_2...\sin\theta_{n-1}\sin\theta_n \end{cases}$$

Однако такой переход радикальным образом сказывается на представлениях распределений. Каждый объект на $\mathbb{S}_n$ описывается $n-1$ координатой с точностью до периода $2\pi k, k \in \mathbb{Z}$. При этом мера поверхности сферы имеет конечное значение, а значит и носитель распределений так же конечен. Это делает невозможным использование таких распределений как Гауссовское напрямую с данными:

$$\begin{cases} p(x|w) = p(\theta|w) \\ p(\theta|w) = \sum\limits_{k_i=-\infty}^{\infty} p(\theta + 2\pi\hat{k}_i|w) \\ p_w(\boldsymbol{\theta}) = \sum\limits_{k_i=-\infty}^{\infty} p_w(\boldsymbol{\theta} + 2\pi k_1\mathbf{e}_1 + \cdots + 2\pi k_F\mathbf{e}_F) \end{cases}$$

2.2 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 $\mathbb{R}^2$:

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

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:

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

- 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.

3 Разделимость распределений

Восстановление распределений классов T в вещественном пространстве эмбеддингов позволяет определить для любого вектора на $\mathbb{S}_n$ оценки плотностей вероятностей его принадлежности конкретной компоненте каждого распределения. В свою очередь это делает возможным решение ряда задач на принятие/отвержение гипотез, получения оценок качества оператора-эмбеддера или же качества составления системы классов при должных предположениях об адекватности классификатора и разметчиков, самого оператора соответственно.

В рамках данных задач важно понимать, насколько хорошо отделимы распределения друг от друга и насколько они попарно идентифицируемы. Более того, для конкретной задачи классификации объектов, идентифицируемость распределений в пространстве эмбеддингов способна дать представление о том, как эффективно будет работать конечное решение. Парную идентифицируемость будем определять индексом различимости (или, как его называют в некоторых источниках, Баесовским индексом различимости), который имеет корни в классической задаче о простых гипотезах в теории **Signal Detection** (где он известен как **discriminability index**).

Определим такой индекс через метрику $i(f_1, f_2)$ от функций плотностей распределений на сфере, действующих в одном вероятностном пространстве. Очевидно, что не каждая метрика способна выступать в качестве такой оценки. Для адекватности получаемых результатов, метрика должна учитывать разницу на всей сфере и отражать ее (например метрика Леви не подходит под данный параметр), быть аналитически вычислимой и устойчивой к размерности пространства и параметрам распределений. Далее предлагается 3 наиболее часто используемых метода построения такой оценки: d', Kullback-Leibler divergence и χ^2 метрики.

3.1 Метрика d'

Статистика $d'(\cdot,\cdot)$ является наиболее классической оценкой разделимости распределений в теории **Signal Detection**, где была выведена из для задачи о разделимости двух нормальных распределений соответствующих гипотезе H_0 и H_1. Для двух гипотез и выбранного порога, статистика $d' = z(P(\text{True positive})) - z(P(\text{False positive}))$, $d' \in (-\infty, +\infty)$, где функция $z(\cdot)$ есть обратная к нормальной функции распределения, и разность вероятностей правильно определить класс H_1 при попарном сравнении с классом H_0 для заданного критерия, из-за чего данная метрика напрямую не применима к введенным ранее распределениям. Попытка же адаптировать d' для распределения **von Mises-Fisher** так же является сложной задачей, ведь подразумевает аналитический вывод кумулятивной функции распределения на сфере.

Изменим d', отказавшись от его неограниченности, убрав обратные кумулятивные функции распределений. При этом идея, заложенная в нем останется: мы все так же будем оценивать вероятности ошибки при принятии гипотезы H_1. Представим новую статистику в виде интеграла по области принятия гипотезы H_1 и построим метрику для любых абсолютно гладких распределений:

$$\begin{cases} \overline{d}'(S_1, S_2) = \displaystyle\int_{\Omega_2} \{f(\omega|S_2) - f(\omega|S_1)\}\, d\omega \\ i_{d'} = \dfrac{\overline{d}'(S_1, S_2) + \overline{d}'(S_2, S_1)}{2} \end{cases}$$

Очевидно, что для $i_{d'}$ зависимость от выбора критерия задачи о гипотезах отпадает, и итоговая формула принимает вид:

$$i_{d'} = \frac{\overline{d}'_1 + \overline{d}'_2}{2} = \frac{1}{2}\int_{\mathbb{S}_{p-1}} |f_1(x|k_1, \mu_1) - f_2(x|k_2, \mu_2)|dx$$

Данное выражение носит под собой значение полной вариационной метрики распределений, однако не имеет аналитического вывода для выбранного распределения Мисеса-Фишера. Это осложняется тем, что численные методы так же не способны принести результата, так как для больших размерностей нормировочные константы распределений ведут себя крайне неустойчиво. Тем не менее данную метрику можно приблизить частотной оценкой **Accuracy** для задачи с одинаковым априорным распределением над объектами. Но такой подход потребовал бы слишком большого объема выборки в условиях высокой размерности пространства.

3.2 Метрика Kullback-Leibler

Более широко используемая оценка распределений, что все так же является несимметричной, а потому требует построения по подобию предыдущего пункта для метрики $i_{KL} = \frac{KL(f_1\|f_2)+KL(f_2\|f_1)}{2}$. Ее главное отличие от предыдущего метода подсчета индекса разделимости: оно оценивает, насколько хорошо одно распределение приближает дифференциальный энтропийный параметр второго, из-за чего значения метрики перестают быть ограниченными.
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 ($\eta = n/2 - 1$):

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

3.3 Метрика χ^2

Точно так же определим метрику как среднее от дивергенций в обоих направлениях $i_{\chi^2} = \frac{\chi^2(f_1\|f_2)+\chi^2(f_2\|f_1)}{2}$. Получаемая метрика так же, как и i_{KL} не является ограниченной, но при этом обладает более экстремальным поведением на "близких"и "дальних"распределениях (стремление к нулю или к бесконечности выражено много сильнее, чем для метрики предыдущего пункта).

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

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: $\overline{\mu} = \frac{(2k_1\mu_1 - k_2\mu_2)}{\|2k_1\mu_1 - k_2\mu_2\|_{\ell^2}}$; $\overline{k} = \|2k_1\mu_1 - k_2\mu_2\|_{\ell^2} \Longrightarrow$

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

3.4 Общие положения

Переход от мультимодальных распределений к унимодальных **von Mises-Fisher** сводит задачу анализа классов к рассмотрению трех групп параметров:

1. Число унимодальных распределений каждого класса, достаточное для его адекватного моделирования.
2. Центры унимодальных распределений.
3. Параметры кучности распределений.

В контексте изначальных предположений, число унимодальных моделирующих распределений отражает сложность или разнообразие составных идей внутри класса, каждое из которых стремится описать свою. То есть, компоненты в смеси аппроксимируют распределение данных подидей. Таким образом, анализ компонент смесей, проводимый при помощи индекс разделимости i, представления которого выведены ранее, позволяет декомпозировать классы и работать уже с их частями, способен показать, насколько чётко различимы составные части класса, позволяет идентифицировать степень перекрытия между классами.

4 Вычисление параметров восстанавливающих смесей

Работа с описанными распределениями требует точного определения параметров, которые наиболее адекватно описывают структуру данных. В случае смесей распределений **von Mises-Fisher** процесс нахождения параметров глобально включает два шага:

1. Нахождение компонент смесей.
2. Нахождение параметров распределений.

Классическим способом решения такой задачи является применение EM алгоритма для смесей без фиксированного числа компонент, тем не менее неприменимым в рамках поставленной задачи ввиду высокой размерности пространства эмбеддингов и потенциальной порядковой разности между требуемым объемом выборки и располагаемой в действительности. Естественным образом возникает потребность в определении более устойчивого метода нахождения параметров.

4.1 Нахождение компонент смесей

В условиях малой обучающей выборки требуется использовать альтернативные методы, устойчивые к высокой размерности, самостоятельно определяющие число компонент и детектирующие выбросы. При этом от алгоритма требуется адекватно находить центры сгущений точек, что соответствуют модам. Одним из таких методов является алгоритм кластеризации **MeanShift**, ядром которого является косинусная метрика для векторов на сфере $\mathbb{S}_{n-1}$.

Алгоритм **MeanShift** представляет собой метод кластеризации, не требующий предварительного задания их числа, обнаруживая его автоматически, основываясь на плотности данных. Основной идеей реализации **MeanShift** является итеративное перемещение каждой точки данных в направлении увеличения

плотности до тех пор, пока точки не сойдутся в локальные максимумы плотности, которые затем и формируют кластеры. Соответственно **MeanShift**:

- Автоматически адаптируется к числу кластеров в высокоразмерном пространстве эмбеддингов.
- Устойчиво себя ведет в обнаружении мод распределений, так как работает только с метрикой.
- Способен определять шумовые объекты, что могут быть следствием ошибки разметки.

Тем самым метод кластеризации определяет для каждой компоненты порожденные ею объекты, что в дальнейшем будут использованы для восстановления параметров.

4.2 Нахождение параметров распределений

В условиях известной выборки точек, порожденных конкретным распределением **von Mises-Fisher**, параметры распределения представимы как максимизаторы оценки правдоподобия:

$$\mathscr{L}(\overline{X}, \mu, k) = \prod_{x \in \overline{X}} c_n(k) \exp\{\langle\, k\mu, x\rangle\} \Longrightarrow \log \mathscr{L}(\overline{X}, \mu, k) = k \sum_{x \in \overline{X}} \mu^\intercal x + \log c_n(k) \to \max_{\mu^\intercal \mu = 1}$$

Из чего следует решение для параметров μ и k:

$$\mu = \frac{\sum_{x \in \overline{X}} x}{\| \sum_{x \in \overline{X}} x \|} \quad k = M_{n/2-1}^{-1}\left(\frac{\| \sum_{x \in \overline{X}} x \|}{n} = \overline{D}\right) = \left(\frac{I_{n/2}(\cdot)}{I_{n/2-1}(\cdot)}\right)^{-1}(\overline{D})$$

Нахождение устойчиво вычислимой обратной функции для M в общем случае не представляется возможным, а потому используются приблизительные оценки. Данная статья использует простейший способ, не включающий в себя использование функций Бесселя, что тем не менее с ростом размерности пространства и выборки асимптотически стремится к действительному решению:

$$\mu = \frac{\sum_{x \in \overline{X}} x}{n\overline{D}} \quad k = \frac{\overline{D}(n - \overline{D}^2)}{1 - \overline{D}^2}$$

5 Устойчивость вычислений

Полученные параметры μ и k позволяют без явного построения получить по выведенным формулам индексы разделимости, и как итог, проанализировать распределения в пространстве эмбеддингов эффективным и формальным образом. Однако каждый подход подсчета i так или иначе требует вычисления функций Бесселя I высоких порядков, а так же произведений и делений экстремально малых или экстремально больших по модулю чисел. Данная глава посвящена численно-устойчивым методам получения промежуточных и конечных значений i для двух дивергенций.

5.1 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:

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

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 $A \cdot Tail$

5.2 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:

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

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

5.3 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:

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

6 Эксперименты

Разработанный инструмент статистического анализа позволяет формально моделировать вероятностную природу объектов в удобном для работы пространстве эмбеддингов, без необходимости учитывать их исходную форму, что достигается путем рассмотрения модели отображения объектов как черного ящика. Таким образом, становится возможным решать как строгие *прямые* задачи классификации наблюдаемых объектов на основе определения порождающего распределения и принятия или отвержения гипотез, так и *обратные* задачи оценки адекватности наблюдаемой выборки. Последний аспект особенно важен, так как включает в себя два ключевых вопроса, связанных с размеченными данными:

1. При условии корректного отображения объектов в пространство эмбеддингов, насколько хорошо и адекватно справились разметчики со своей работой.

2. При условии корректной разметки, насколько успешно эмбеддер создал адекватные представления объектов.

Каждая из полученных постановок задач основывается на определенных идеальных условиях, которые трудно полностью соблюсти на практике. Тем не менее, посредством интерпретации индексов разделимости распределений, рассматриваемых в исследовании, эти задачи могут быть решены на эмпирическом уровне. От инструмента ожидается устойчивость к большим размерностям пространства эмбеддингов и малым размерам обучающих выборок, соответствие значений индексов адекватным ожиданиям, наглядность в отображении свойств данных, а также способность решать задачи классификации объектов не хуже более простых прямолинейных методов.

6.1 Датасет

В исследовании использовался датасет над корпусом русских текстов, размеченных лабораторией Семантического анализа при университете МГУ. Далее будем отождествлять понятие **метка** и соответсвующий ей **класс**. Рассматривалась разметка ценностей в текстах по классификатору, состоящему из меток относящих ценность к соответствующим этническим или социальным группам, выражающих материальность/духовность ценности в различном проявлении, черты или характер проявления данной ценности. Разметка соответствует задаче multilabel-classification, где каждый фрагмент помечен набором меток, определяющих, какая ценность в конкретном фрагменте содержится, какое отношение к ней сложилось у автора, какое воздействие приведение данной ценности должно было произвести на читающего. Такой функционал отчасти выражается в
существовании 4 специальных меток тональности фрагмента: *отрицательная*, *нейтральная*, *положительная*, *конфликт тональности*.

Данная задача декомпозирована на задачу multilabel-classification, где каждый объект, помеченный набором меток, представлялся как множество одинаковых объектов, каждый из которых относится к одному соответствующему классу по изначальной группе.

6.2 Постановка задачи

Пусть определены множества документов $\mathscr{D} = \{d_1, ..., d_m\}$, множество классов $T = \{t_1, ..., t_\tau\}$, всего фрагментов F, разницей в работе разметчиков можно пренебречь, считая их равными, тем самым не рассматривая отдельно взятого разметчика как случайное воздействие. Тогда будем считать токенами в тексте (неделимыми по смыслу единицами) слова. Разметчик, смотря на слово в контексте документа, порождает дискретное распределение на множестве классов T

$$p(t_k|token_i, d_j), t_i \in T, token_i \in d_j, d_j \in \mathscr{D}$$

Откуда случайным образом выбирает метку и ставит ее в соответствие слову $token_i$. И оптимизационная задача становится задачей максимизации оценки правдоподобия:

$$L_{prob} = \prod_i p(t_i|token_i, d_{j_i}) \longrightarrow \max_{t_i \in T}$$

Пусть существует оператор, переводящий фрагменты текста осмысленно в пространство, интерпретируемое компьютером. Действуя в изначальных предположениях об этом операторе, его способности схожие по смыслу фрагменты ставить ближе, чем фрагменты с противоположным или контрастирующим содержанием (гипотеза компактности в пространстве смыслов озвученная во вступлении), будем приближать данный оператор работой **LLM**. В экспериментах использовался **ruBERT-base-cased-sentence**, обученный на задаче предсказания фрагментов текста.

6.3 Анализ распределений классов

Рассматривалась выборка из 84 классов, разметка по которым присутствует в датасете.

1. Для каждого класса по отдельности были найдены все области сгущения представителей с помощью алгоритма кластеризации **Mean-Shift**, каждая из которых сопоставлялась отдельному распределению.
2. На основе точек сгущения находились параметры моделирующего распределения **von Mises-Fisher**.
3. Для каждой пары распределений находились индексы i_{KL} и i_{χ^2}

В результате проведения данной операции получено 141 компонента, из которых для каждого отдельного класса соответствует не более 2. Тем самым каждый класс соответствует унимодальному или бимодальному распределению, а потому в предположениях об идейной содержательности такого представления, каждый класс несет в себе одну общую, либо две достаточно разных идеи. Теперь возможен анализ, какие классы между собой разделяют наиболее близкие и наиболее дальние идеи. Для этого будем брать наименьшее расстояние между составными распределениями двух классов. Результаты представлены ниже:

Таблица 1: Наиближайшие классы по своим *"подтемам"*

Таблица 2: Метрика i_{χ^2}

Класс 1	Класс 2	i
Достоинство	Культура и искусство	7.45e-9
Еда	Важность общ. мнения	3.17e-8
Достоинство	Честность	3.43e-8
Достоинство	Историческая память	5.87e-8
Гордость	Потворство желаниям	7.478e-8
Любовь	Культура и искусство	1.21e-7
Этничность	Удовольствие жизнью	6.27e-7
Воспитание	Познание	1.19e-6
Образование	Благочестие	1.72e-6
Патриотизм	Природа	2.44e-6

Таблица 3: Метрика i_{KL}

Класс 1	Класс 2	i
Мат. ценности	Нейтральный тон	36.86
Мор. ценности	Чувство принадлежности	78.77
Смелость	Положительный тон	115.85
Патриотизм	Уважение традиций	122.34
Полит. ценности	Права и свободы	126.34
Заботливость	Положительный тон	130.55
Нац. безопасность	Выборность власти	133.64
Жизнь	Чувство принадлежности	134.82
Безопасность	Конфликт тональности	135.90
Патриотизм	Этничность	176.16

Таблица 4: Самые удаленные классы по своим *"подтемам"*

Таблица 5: Метрика i_{χ^2}

Класс 1	Класс 2	i
Язык	Справедливость	6.24e16
Счастье	Твердая воля	1.40e15
Справедливость	Творчество	7.36e14
Заботливость	Нейтральный тон	2.88e13
Гуманизм	Смелость	1.69e13
Культура	Здоровье	5.14e11
Творчество	Смелость	2.46e11
Язык	Счастье	1.95e11
Творчество	Заботливость	5.44e10
Счастье	Жилище	2.50e10

Таблица 6: Метрика i_{KL}

Класс 1	Класс 2	i
Историческая память	Впечатления от жизни	7465.42
Смелость	Гуманизм	6651.31
Этничность	Мат. ценности	6433.21
Природа	Образование	6379.33
Талант	Чувство юмора	6235.82
Красота	Впечатления от жизни	6190.90
Власть	Жизнь	6186.52
Творчество	Верность	6174.12
Язык	Честность	6165.45
Еда	Твердая воля	5818.10

Полученные результаты для обеих метрик по нахождению 10 наиболее близких и отдаленных друг от друга классов хорошо согласуются с эмпирическим опытом и ожиданиями разметчиков/людей. Тем не менее в зависимости от выбора метрики состав таблицы меняется. Так, метрика на основе KL-дивергенции смогла

определить близкими между собой не только смысловые классы, но и классы, определяющие тональности фрагментов. Вместе с этим и разброс значений метрик рознится: i_{χ^2} принимает значения от 10^{-9} до 10^{16}, когда для i_{KL} разброс представлен разницей в два порядка. Важно отметить, что в каждой из таблиц присутствует доля повторяющихся несколько раз тегов. Такое поведение объясняется высокой схожестью фрагментов, представляющих соответственные классы, в контексте данного корпуса текстов.

6.4 Задача multilabel классификации

Рассмотрим работу полученного алгоритма на основе простейшей задачи multilabel-classification на подвыборке меток из датасета $\tilde{T}$. Выбор меток осуществлялся по принципу 10 наиболее широко представленных классов в датасете. Для каждого объекта будем ставить в соответствие вектор из нулей и единиц - маркируется ли данный объекта соответственной меткой. Для этого определим процедуру классификации:

$$token \sim \left\{ i \in \tilde{T} : j = \arg\max_j \left(p_{i_j}(token) \right) ; \ \mathbb{1}[p_{i_j} > h_{i_j}] \right\}$$

Где p_{i_j} - вероятность того, что изначальный объект порожден j компонентой распределения i класса; h_{i_j} - порог для j компоненты распределения i класса. Тогда определим обучающую выборку как 80% от всей полученной для $\tilde{T}$ и тестовую выборку как 20% оставшейся, и будем решать задачу оптимизации, как $\max_i F_1 score(i)$ - максимизируя $F_1 score$ определения каждого класса для объектов. Конечный результат подводится как макро усреднение по всем классам и равняется **0.643**, что говорит о состоятельности полученного алгоритма.

7 Выводы

Полученный алгоритм перехода от мультимодальных распределений к смесям простых унимодальных способен не просто упростить анализ реализации выборки, но и позволяет раскладывать классы на составные части, качественно решать на их основе задачи проверки работы разметчиков, работы эмбеддера, простые задачи классификации, на основе введенных индексов разделимости распределений.

8 Ссылки

1. Michael J. Hautus, Neil A. Macmillan & C. Douglas Creelman (2022) Detection theory. Third Edition
2. Chapman & Hall (2017) Modern Directional Statistics
3. Xie, Junyuan, Ross Girshick, и Ali Farhadi. "Deep Clustering and Mixture Model for Unsupervised Learning."arXiv preprint arXiv:1602.03591 (2016)