Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н. Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н. Э. Баумана)

ФАКУЛЬТЕТ Фундаментальные науки

КАФЕДРА Математическое моделирование

Научно – исследовательская работа

НА ТЕМУ:

«Регрессия распределения» (distribution regression problem). Обзор литературы

Студент ФН12-11м (Группа) _____ (Подпись, дата) В. А. Лосев (И. О. Фамилия)

Научный руководитель _____ (Подпись, дата) В. А. Панкратов (И. О. Фамилия)

2023 г.

Оглавление

1. Введение

Регрессия распределения — это случай регрессии, когда факторами модели являются распределения. Многие важные задачи машинного обучения и статистики вписываются в эту структуру. На практике наблюдаемы только выборки из распределений, и оценки должны основываться на сходствах, вычисленных между наборами точек. Регрессия распределения имеет широкий спектр применения, например, поиск аэрозолей с использованием многоспектральных спутниковых изображений или классификация документов, где каждый документ представлен в виде набора слов. В некоторых рассматриваемых статьях с помощью регрессии распределения решается проблема экологического вывода (**ecological inference problem**) — проблема изучения групп пользователей на индивидуальном уровне по совокупным данным. В данной научно–исследовательской работе будет представлен обзор некоторых статей, относящихся к анализу регрессии распределения и затрагивающие ее область применения, также будет сказано про методы, используемые при решении данной проблемы.

2. Статья

Who supported Obama in 2012? Ecological inference through distribution regression

В данной статье рассматривается решение проблемы “экологического вывода” (ecological inference) , заключающееся в изучении групп населения на индивидуальном уровне по совокупным данным. В отличие от других методов экологического вывода , представленный метод использует индивидуальные данные путем встраивания распределения по этим данным в вектор в гильбертовом пространстве. Представленный метод применим для работы с большим объемом данных, благодаря использованию метода аппроксимации FastFood при решении задачи гребневой регрессии.

Встраивание распределений в ядро (kernel embeddings of distribution) – это довольно мощный класс в RKHS, который переводит распределение вероятности в вектор признаков в многомерном или бесконечномерном пространстве. Пусть $\varphi : \mathbb{R} \to H$, если независимые одинаково распределенные величины $x \sim X$, то встраивание среднего (kernel mean embedding) определяется как

$$\mu_x = \mathbb{E}[\varphi(x)]$$

В этой работе используется простая оценка среднего значения для μ_x:

$$\widehat{\mu}_x = \frac{1}{N} \sum_j \varphi(x^j)$$

В этом разделе непосредственно рассматривается регрессия распределения, с помощью которой можно сопоставить распределения вероятностей с метками. Поскольку вероятностное распределение наблюдается через выборку, то в качестве входных параметров имеем:

$$\left(\{x_1^j\}_{j=1}^{N_1}, y_1\right), \left(\{x_2^j\}_{j=1}^{N_2}, y_2\right), \ldots, \left(\{x_1^j\}_{j=1}^{N_1}, y_n\right),$$

где каждая группа $x_i^j \in \mathbb{R}^d$ (в данном случае индивидуальные демографические данные каждого округа) имеет единственное целочисленное значение метки y_i (процент голосов за определенного кандидата), а N_i – количество наблюдений в каждой группе. Предполагается, что выборки данных для каждой группы произвольно взяты из какого-то неизвестного мета-распределения [1]. Реализация регресионной модели требует двух основных шагов. На первом этапе строится вектор признаков в новом пространстве, и далее находится среднее вложение (kernel mean embedding):

$$\widehat{\mu}_1 = \frac{1}{N_1}\sum_j^{N_1}\varphi(x_1^j), \ldots, \widehat{\mu}_n = \frac{1}{N_n}\sum_j^{N_n}\varphi(x_n^j).$$

Далее, используя гребневую регрессию, находим функцию в виде

$$y = f(\widehat{\mu}) + \varepsilon,$$

где основная задача сводится к минимизации функционала и нахождении параметра λ:

$$\widehat{f} = argmin\sum_i [y_i - f(\widehat{\mu}_i)]^2 + \lambda||f||^2$$

Так как мы работаем в Гильбертовом пространстве над μ_i, выберем линейные ядра, такие, что $k(\widehat{\mu}_i, \widehat{\mu}_j) = \langle\widehat{\mu}_i, \widehat{\mu}_j\rangle$. Следуя стандартному выводу ядерной гребневой регрессии, мы также можем искать значение функции f для новой тестовой выборки x_* в замкнутой форме:

$$f(\mu_*) = k^*(\mathbb{K} + \lambda I)^{-1}[y_1, \ldots, y_n]^T,$$

где $k^* = [\langle\widehat{\mu}_1, \mu_*\rangle, \ldots, \langle\widehat{\mu_n}, \mu_*\rangle$ и $\mathbb{K}_{ab} = \langle\widehat{\mu}_a, \widehat{\mu}_b\rangle$. Стоит отметить, что при решении задачи регрессии распределения вместо самих данных используется только вложение среднего.

Решение данной задачи с использованием **kernel trick** ($K_{ab} = \langle\widehat{\mu}_a, \widehat{\mu}_b\rangle = \frac{1}{N_a N_b}\sum_j\sum_l k(x_a^j, x_b^l)$) является трудозатратным в вычислительном плане. Так как для нахождении матрицы Грама $\mathbb{K}$ понадобится $O(n^2N^2)$ операций. Так как в конечном счете нужно работать со средними вложениями μ_i, а не с отдельными наблюдениями x_i^j, явное представление признаков, особенно если оно многомерное, значительно снизит вычислительные затраты. В данной статье используется метод аппроксимации φ под названием FastFood, которое находит приближение $\widehat{\varphi(x)}$ в $\mathbb{R}^d$ для каждого $\mathbf{x}$. При этом φ может быть любым ядром радиальной базисной функции. Для примера берется Гауссово ядро, тогда φ будет иметь следующий вид:

$$\varphi(x) = p^{-1/2}\exp i[Vx],$$

где i – мнимая единица, а V – соответствующем образом масштабируемая случайная матрица, $V = [V_1^T, V_2^T, \ldots, V_{[p/d]}^T]^T$, где каждая матрица представлена в виде

$$V_j = \frac{1}{\sigma\sqrt{d}}SHG\Pi HB,$$

где S, B, C – диагональные случайные матрицы;

Π – случайная перестановка;

H – матрица Уолша-Адамара.

Таким образом, все преобразование $\varphi(x)$ может быть вычислено за O(p log d). Это на порядок быстрее, чем метод Random kitchen Sinks и Random Fourie Features и гораздо быстрее вычислений с использованием Kernel trick. Проиллюстрируем данный метод на примере.

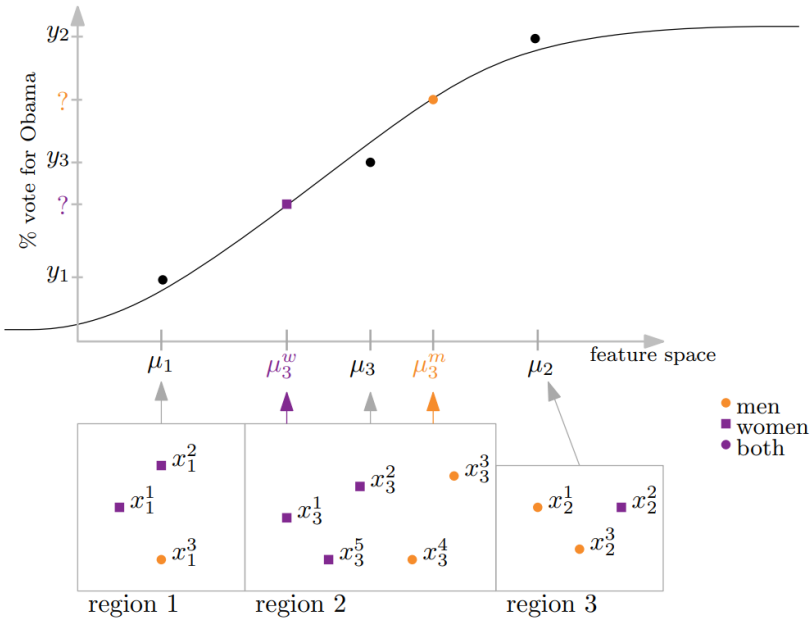

Рис. 1. Регрессия в задаче экологического вывода

Метки y_1, y_2, y_3 доступны на уровне регионов, они равны доли голосов, полученных Обамой в регионах 1, 2 и 3. Также доступны индивидуальные данные в регионах 1, 2 и 3. Спроецируем данные из каждой группы в пространство признаков, используя $\varphi(x)$, и берем среднее значение по группе, чтобы найти векторы μ_1, μ_2, μ_3, например $\mu_1 = \frac{1}{3}(\varphi(x_1^1) + \varphi(x_2^1) + \varphi(x_3^1))$. Теперь наша задача сводится к обучению с учителем, где мы хотим найти значение $f : \mu \to y$. Как только мы узнаем f, мы делаем прогнозы по подгруппам для мужчин и женщин в регионе 3, вычисляя средние значения вложений для мужчин. $\mu_3^m = \frac{1}{2}(\varphi(x_3^3)+\varphi(x_4^3))$ и женщин $\mu_3^w = \frac{1}{3}(\varphi(x_1^3)+\varphi(x_2^3)+\varphi(x_5^3))$, а затем вычисляем $f(\mu_3^m)$ и $f(\mu_3^w)$.

Дальнейшие рассуждения сводятся непосредственно к решению ecologiacal inference problem и не представляют глобальный интерес к задаче регрессии распределения.

3. Статья

Distribution regression and its application to ecological inference

Ранее было дано описание регресии распределения. Для большей наглядности проиллюстрируем основную задачу, которую необходимо решить.

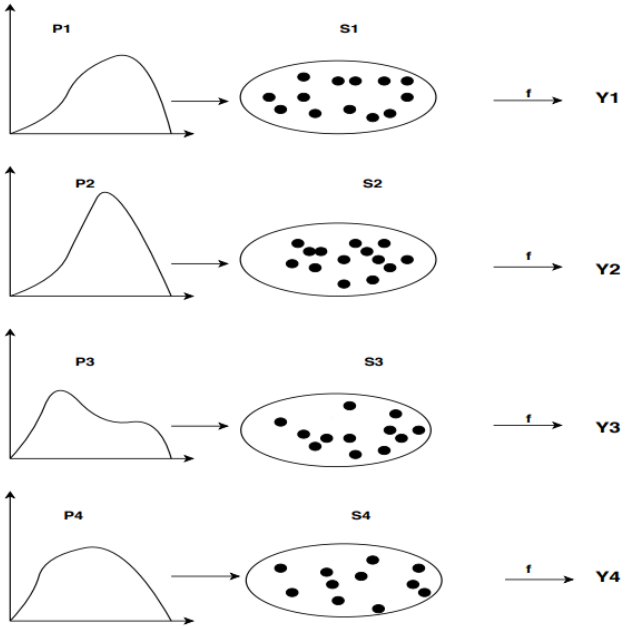

Рис. 2. Регрессия распределения

Регрессия распределение только недавно получила свое развитие и имеет широкую область применимости:

- Классификация изображений, заданных в виде набора точек;
- измерение массы скоплений галактик с использованием дисперсии их скоростей;
- классификация документов, где каждый документ представлен в виде набора слов;
- классификация трехмерных геометрических данных, где каждый объект представлен в виде облака точек;
- прогнозирование поведения при голосовании для демографических подгрупп географических регионов.

Задача, указанная в последнем пункте относится к проблеме экологического вывода: предположим, что есть n регионов, и для каждого региона у нас есть таблица

с общим количеством голосов по категориям и демографическим подгруппам, и проблема состоит в том, чтобы вывести долю голосов в зависимости от демографических показателей. Для этого предполагается, что регионы имеют что-то общее, и мы можем извлечь общую модель из всех них.

Входные данные для регресии распределения:

$$[(\{\mathbf{x_1}\}, y_1), \ldots, (\{\mathbf{x_n}\}, y_n)],$$

где $\mathbf{x_i}$ – выборочные данные из области i, а y – соответствующие им метки. Пусть:

$$\boldsymbol{X} = \begin{bmatrix} \varphi(\{x_1\}) \\ \varphi(\{x_2\}) \\ \vdots \\ \varphi(\{x_n\}) \end{bmatrix} \tag{1}$$

и $\boldsymbol{y} = [y_1, y_2, \ldots, y_n]^T$, где $\mathbf{X}$ – вектор средних вложений ядра для каждой области. Задача гребневой регрессии заключается в поиске $\hat{\beta}$, такой, что

$$\hat{\beta} = argmin(\boldsymbol{y} - \boldsymbol{X}\beta)^T(\boldsymbol{y} - \boldsymbol{X}\beta) + \lambda||\beta||^2,$$

где λ – регуляризацинное слагаемое, которое находится с использованием кросс-валидации. Решение в виде

$$\hat{\beta} = ((\boldsymbol{X}^T\boldsymbol{X} + \lambda I)^{-1}\boldsymbol{X}^T\boldsymbol{y}).$$

Тогда предсказание значения метки y^* для новой выборки $\{x_*\}$:

$$\hat{\boldsymbol{y}}^* = \varphi(\{x_*\})\hat{\beta}.$$

Также решение можно записать в следующем виде:

$$\hat{\beta} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T + \lambda I_n)^{-1}\boldsymbol{y},$$

где $\boldsymbol{X}\boldsymbol{X}^T = K$ – матрица Грама.
Если мы определим

$$k^* = [\langle\widehat{\mu_1}, \mu_*\rangle, \ldots, \langle\widehat{\mu_n}, \mu_*\rangle = [K_{*1}, \ldots, K_{*n}],$$

тогда предсказание для новой выборки $\{x_*\}$:

$$y_* = \boldsymbol{k}^*(K + \lambda I)^{-1}\boldsymbol{y} = \boldsymbol{k}^*\alpha = [K_{*1}, \ldots, K_{*n}][\alpha_1, \ldots, \alpha_n]^T$$

Для вычисления матрицы K можно использовать Гауссово ядро, тогда ее компоненты вычисляются в следующем виде:

$$K_{ab} = \frac{1}{N_1 N_2} \sum_j \sum_l \exp -\frac{||x_a^j - x_b^l||}{2\sigma^2}$$

Но ввиду вычислительной сложности такого подхода часто используют явное представление признаков $z : \mathbb{R}^d \to \mathbb{R}^D$, такое, что $z(x)^T z(y) \approx k(x, y)$. Существует метод аппроксимации положительно определенного ядра k с использованием преобразования признаков (feature map) более низкой размерности путем проецирования данных на случайно выбранную линию, а затем применения функции масштабирования. Данный результат основан на теореме Бохнера, которая гарантирует, что если ядро инвариантно относительно сдвига, положительно определено, то его преобразование Фурье $p(\omega)$ является соответсвующим правильным распределением вероятности.

$$k(x, y) \approx \frac{1}{D} \sum_j^D z_{wj}(x) z_{wj}(y) = z(x)^T z(y),$$

где $z(x) = [z_{w1}(x), \ldots, z_{wD}(x)]$. Таким образом, опишем общий алгоритм [2]

- Выбираем положительно определенное инвариантное ядро k, в данном случае Гауссово:
$$k(\nabla) = \exp \frac{-||\nabla||^2}{2\sigma^2}$$
- находим преобразование Фурье для k. Для гауссова ядра:
$$p(w) = \left(\frac{\sigma}{\sqrt{2\pi}}\right)^D \exp -\frac{\sigma^2 ||w||^2}{2}$$
- генерируем D независимых, одинаково распределенных величин $w_1, \ldots, w_n$ из p;
- генерируем D независимых, одинаково распределенных величин $b_1, \ldots, b_n$ из равномерного распределния на отрезке $[0; 2\pi]$;
- вычисляем $z(x) = [\sqrt{2}\cos(w_1^T x + b_1), \ldots, \sqrt{2}\cos(w_D^T x + D_1)$.

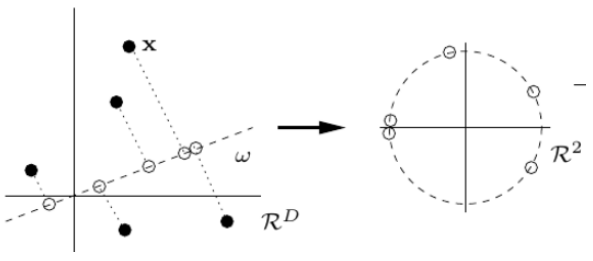

Рис. 3. Преобразование признаков random Fourie features

Таким образом, с помощью данного представления признаков, можно вычислить mean embedding μ_p и далее элементы матрицы Грама $\mathbb{K}$:

$$K_{ab} = \frac{1}{N_a N_b} \sum_j \sum_l k(x_a^j, x_b^l) \approx \frac{1}{N_a N_b} \sum_j \sum_l k(x_a^j, x_b^l)$$
$$= \left[\sum_j \frac{1}{N_a} z(x_a^j) \right]^T \sum_l \frac{1}{N_b} z(x_b^l) \quad (2)$$

Также одно из преимуществ такого явного представления пространства признаков является возможность сделать предсказание для новой выборке по следующей формуле:

$$\hat{y}^* = \left[\sum_j \frac{1}{N_*} z(x_*^j) \right]^T \hat{\beta}.$$

Либо

$$\hat{y}^* = \mathbf{k}^*(\mathbb{K} + \lambda I_n)^{-1} y$$

3.1. Встраивание распределений в ядро (kernel mean)

Ранее было упомянуто такое понятие, как **kernel mean**, в данном разделе раскроем это понятие более подробно. В машинном обучении встраивание распределений в ядро (среднее ядра) включает класс непараметрических методов, в которых распределение вероятностей представляется как элемент воспроизводящего гильбертова пространства ядра (RKHS). Отображения признаков отдельных точек данных, выполненное в классических методах ядра, встраивание распределений в бесконечномерные пространства признаков может сохранить все статистические особенности произвольных распределений, позволяя при этом сравнивать распределения, используя операции Гильбертова пространства. Теория, лежащая в основе встраивания распределений в ядро, была в основном разработана Алексом Смолой, Ле Сонгом, Артуром Греттоном и Бернхардом Шелкопфом.

$$\mu_{\hat{X}_i} = \int_\Omega k(\cdot, u) \, \mathrm{d}\hat{X}_i(u) = \frac{1}{N_i} \sum_{n=1}^{N_i} k(\cdot, X_{i,n})$$

Методы, основанные на встраивании распределений в ядро обладают следующими преимуществами [3]:

- Моделирование можно проводить без явных предположений о форме распределений и взаимосвязях между переменными;
- промежуточная оценка плотности не требуется;

- если используется характеристическое ядро, то встраивание однозначно сохраняет всю информацию о распределении, а также с использованием **kernel trick** вычисления в возможно бесконечномерных RKHS могут быть реализованы на практике в виде простых операций.

Также с использованием среднего вложения ядра можно вычислить расстояние между распределениями $P(X)$ и $Q(Y)$, которое определяется как расстояние между их вложениями в RKHS

$$MMD(P,Q) = \|\mu_X - \mu_Y\|_{\mathcal{H}}.$$

На практике, данная оценка хороша тем, позволяет найти показатель близости распределения по выборочным данным, при этом нет необходимости знать плотность распределения, как при вычислении расстояния Кульбака – Лейблера.

Также на практике могут использоваться различные ядра [4]:

$$K(\mu_P, \mu_Q) = e^{-\left\|\mu_P - \mu_Q\right\|^2_{H(k)}/(2\sigma^2)} - \text{Гауссово},$$

$$K(\mu_P, \mu_Q) = e^{-\left\|\mu_P - \mu_Q\right\|_{H(k)}/(2\sigma^2)} - \text{Экспоненциальное},$$

$$K(\mu_P, \mu_Q) = \left(1 + \|\mu_P - \mu_Q\|^2_{H(k)}/\sigma^2\right)^{-1} - \text{Коши}$$

и т.д.

4. Kernel Mean Embedding of Distributions: A Review and Beyond

Ранее были определены понятия среднего вложения ядра и даны соответствующие формулы. Встраивание распределений в ядро имеет следующий смысл, представленный на рисунке: каждое распределение отображается в воспроизводящее гильбертово пространство ядра с помощью вычисления среднего вложения.

Основные идеи использования среднего ядра могут быть представлены следующим образом [5]:

- используемые данные: $D = x_1, x_2, ..., x_n$;
- определяем преобразование признаков φ;
- применяем к выборке $D_\varphi = \{\varphi(x_1), \ldots, \varphi(x_n)\}$;
- решаем более легкую задачу в пространстве $\mathbb{H}$, с использованием D_φ.

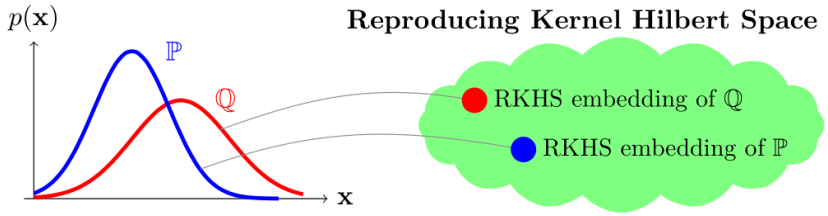

Рис. 4. Отображение распределения в Гильбертово пространство

Для примера, рассмотрим полиномиальное преобразование признаков $\varphi(\mathbf{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$, где $\mathbf{x} \in \mathbb{R}^2$. Тогда

$$\langle \varphi(x), \varphi(x') \rangle = x_1^2 x_1'^2 + x_2^2 x_2'^2 + 2x_1x_2x_1'x_2' = \langle x, x' \rangle^2.$$

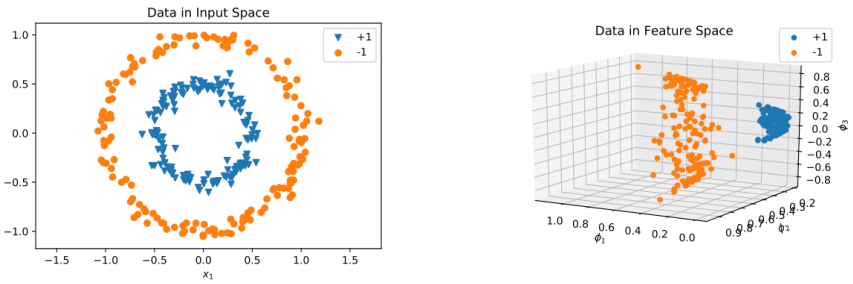

Рис. 5. Преобразование признаков в пространство повышенной размерности

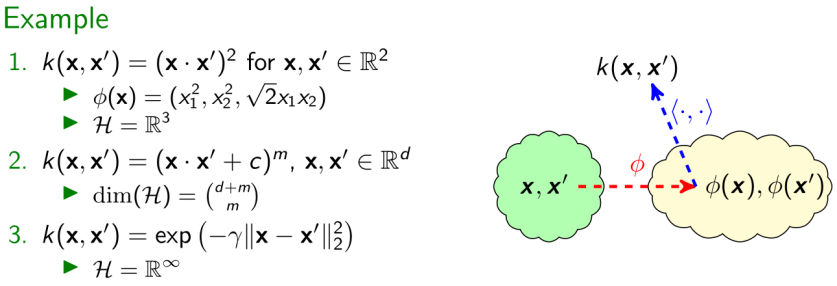

Рис. 6.

Зная μ_p, можно получить информацию о распределении:

- $k(x, x') = \langle x, x' \rangle$ – первый момент P;
- $k(x, x') = (\langle x, x' \rangle + 1)^n$ – n-ый момент P;
- $k(x, x')$ универсальное/характеристическое – вся информация о P.

Таким образом, зная значение μ_p для соответствующего распредления, с использованием характеристических ядер, к которым, например, относится Гауссово ядро, можно оценить моменты, а также восстановить плотность распределения [6]. Обычно, в задачах машинного обучения используется Гауссово ядро и ядро Лапласса.

$$k(x, x') = \exp\left(\frac{-||x - x'||_2^2}{2\sigma^2}\right), \quad k(x, x') = \exp\left(\frac{-||x - x'||_1}{\sigma}\right)$$

Эти ядра принадлежат к классу функций ядра, называемых радиальными базисными функциями (RBF). Также в данной статье упомянуты две важные теоремы.
Теорема Мерсера: Предположим, что K — непрерывное симметричное положительно определенное ядро. Тогда существует ортонормированный базис φ_i из $L^2[a,b]$, состоящий из собственных функций таким образом, что соответствующая последовательность собственных значений λ_i неотрицательна. Собственные функции, соответствующие ненулевым собственным значениям, непрерывны на $[a,b]$, а K может быть представлено в виде:

$$K(u,v) = \sum_{j=1}^{\infty} \lambda_j \, \varphi_j(u) \, \varphi_j(v)$$

Важным следствием теоремы Мерсера является то, что из представленого выше соотношения можно вывести явную форму отображения признаков $\varphi(x)$, т.е. отображение признаков может иметь вид

$$\varphi(x) = [\sqrt{\lambda_1}\varphi_1(x), \sqrt{\lambda_2}\varphi_2(x), \ldots]$$

Также, особого внимания заслуживает теорема Бохнера: Комплекснозначное ограниченное непрерывное ядро $k(x,x') = \varphi(x-x')$ на $\mathbb{R}^d$ положительно определено тогда и только тогда, когда существует конечная неотрицательная борелевская мера Λ на $\mathbb{R}^d$ такая, что

$$\varphi(x-x') = \int_{R^d} \exp(i\omega^T(x-x'))d\Lambda(\omega)$$

Важным применением теоремы Бохнера является возможность аппроксимации функции ядра (метод random fourie features), что полезно для уменьшения вычислительной сложности, поскольку отображение функций φ многих функций ядра, таких как гауссово ядро, является бесконечномерным. Далее в данной статье рассматриваются различные свойства вложения среднего, а также методы, в которых необходимо использование ядер, отличные от темы работы. Некоторые способы аппроксимации были представлены в предыдущих разделах.

5. Заключение

В рамках работы был произведен обзор литературы, относящейся к такому понятию, как **регрессия распределения**. В качестве литературы были взяты статьи об использовании регрессии распределений в задаче экологического вывода, а также книга, посвященная использованию такого понятия, как среднее ядра в новом пространстве признаков (kernel mean embedding). Было дано описание основных идей

и реализации регрессии распределения, а также ее область применения. Также были рассмотрены различные подходы к реализации регрессии распределения, с использованием различных методов, таких, как «трюк с ядром» и различные подходы аппроксимации значения ядра в случае работы с большим объемом данных или в пространстве высокой/бесконечномерной размерности.

Список использованных источников

1. S.R. Flaxman, Y. Wang, A. Smola. Who Supported Obama in 2012? Ecological Inference through Distribution Regression, 2016. 10 p.
2. L. L. Gonzalez Distribution Regression and its application to Ecological Inference. Ciencias de la Computación y Matemáticas Industriales, 2019. 64 p.
3. K Muandet, K Fukumizu. Kernel Mean Embedding of Distributions: A Review and Beyond. 2020. 147 p.
4. Kernel embedding of distributions URL: https://en.wikipedia.org/wiki/Kernel_embedding_of_di (Дата обращения: 30.10.2023).
5. Kernel method URL: https://en.wikipedia.org/wiki/Kernel_method (Дата обращения: 5.11.2023).
6. M Kanagawa. Recovering Distributions from Gaussian RKHS Embeddings 2020. 9 p.