Децентрализованная доменная адаптация

.....

. . . . .

Рассматривается задача восстановления совместного распределения по маргинальным.

Ключевые слова:

Слайд 2.

Решается задача доменной адаптации. Суть этой адаптации заключается в обучении модели на данных из домена-источника так, чтобы она показывала сравнимое качество на целевом домене. Например, домен-источник может представлять собой синтетические данные, которые можно «дешево» сгенерировать, а целевой домен — фотографии пользователей. Тогда задача доменной адаптации заключается в тренировке модели на синтетических данных, которая будет хорошо работать с «реальными» объектами.

Данную задачу предлагается решить через построение функции преобразования одного домена в другой. Мотивация подхода заключается в том, что после того, как мы построим такую функцию, т.е. получим возможность переводить распределение одного домена в другой, и умея решать некую задачу на первом домене, мы получим сравнимое качество для второго домена на данной задаче.

Исследуется проблема построения и анализа вероятностного пространства параметров этого преобразования. Проблема усложнена тем, что домены могут принадлежать непересекающихся или слабо пересекающихся пространствах.

Метод исследования: Байесовский подход к нахождению параметров модели.

Задача - нахождение функции преобразования одного домена в другой. Параметры функции находятся для различных функций сходства.

Проблема - Различные функции сходства дают различные результаты. В работе предлагается ряд критериев для анализа качества.

Результат - Производится анализ функций преобразования для различных функций сходства распределений.

Слайд 3.

Определение 1. Каждый *домен* $\mathcal{D}$ описывается парой $\{G, \mathcal{J}\}$, где $\mathcal{J}$ - множество индексов, G - область пространства признакового описания, причем $G \subset \mathbb{R}^{\mathcal{J}}$.

Определение 2. *Функцией преобразования* домена $\{G_1, \mathcal{J}_1\}$ в домен $\{G_2, \mathcal{J}_2\}$ назовем функцию $f : \mathbb{R}^{\mathcal{J}_1} \to \mathbb{R}^{\mathcal{J}_2}$. На параметры этого преобразования может задаваться какая-то априорная информация.

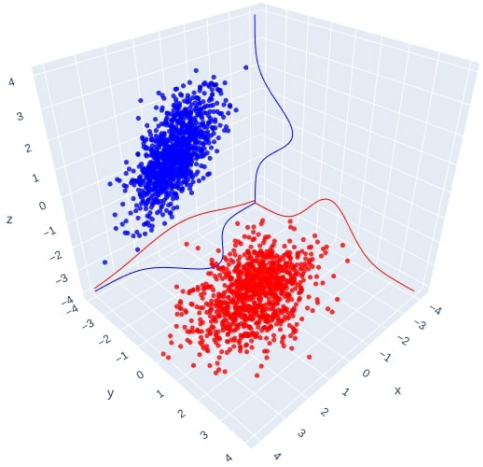

Рис. 1 Пример двух доменов в общем признаковом пространстве. Домен 1 (синий) принадлежит пространству z-x, домен 2 (красный) принадлежит пространству x-y. Для каждого домена задано распределение.

Определение 3. *Оптимальной функцией преобразования* домена $\mathcal{D}_1$ в домен $\mathcal{D}_2$ относительно функции сходства g назовем функцию $\hat{f}$, на которой достигается максимальное сходство распределений второго домена и образа функции f, т.е.:

$$\hat{f} = \arg\max_f g\Big(p(x, x \in \mathcal{D}_2), p(f(x), x \in \mathcal{D}_1)\Big)$$

Из [1]:

Определение 4. Назовем функцией сходства s_score пары распределений $g_1 : \mathbb{R}^n \to \mathbb{R}^+$ $+, g_2 : \mathbb{R}^n \to \mathbb{R}^+$, определенных на одном пространстве, функцию вида

$$s_0(g_1, g_2) = \frac{\int g_1(\mathbf{w}) g_2(\mathbf{w}) d\mathbf{w}}{\max\limits_{\mathbf{b} \in \mathbb{R}^n} \int g_1(\mathbf{v}) g_2(\mathbf{v} - \mathbf{b}) d\mathbf{v}}$$

В работе будут использоваться функции сходства построенные на дивергенции Кульбака-Лейблера, метрика Васерштейна.

Слайд 4. Постановка задачи.

В качестве функции преобразования будем брать обобщенно линейную функцию, вида

$$f_{\mathbf{v}_1,\mathbf{V}_2,\mathbf{b}_2,\mathbf{b}_1} = \mathrm{diag}(\mathbf{v}_1)\sigma(\mathbf{V}_2\mathbf{X} + \mathbf{b}_2) + \mathbf{b}_1,$$

где $\mathbf{v}_1, \mathbf{V}_2, \mathbf{b}_2, \mathbf{b}_1$ - параметры преобразования.

Будем искать оптимальные параметры функции относительно функции сходства g через задачу максимизации:

$$\mathbf{v}_1, \mathbf{V}_2, \mathbf{b}_2, \mathbf{b}_1 = \underset{\mathbf{v}_1,\mathbf{V}_2,\mathbf{b}_2,\mathbf{b}_1}{\arg\max}\, g\Big(p(x, x \in \mathcal{D}_2), p(f_{\mathbf{v}_1,\mathbf{V}_2,\mathbf{b}_2,\mathbf{b}_1}(x), x \in \mathcal{D}_1)\Big)$$

Слайд 4.1. Постановка задачи для дивергенции KL.

$$p(\theta_f|\mathbf{z}, \theta_d) \propto \Big(\prod_{i=1}^{n_f} D(f(\mathbf{z}^{(i)}; \theta_f); \theta_d)\Big) p(\theta_f|\alpha_f)$$

$$p(\theta_d|\mathbf{z}, \mathbf{X}, \theta_f) \propto \prod_{i=1}^{n_d} D(\mathbf{x}^{(i)}; \theta_d) \times \prod_{i=1}^{n_f} (1 - D(f(\mathbf{z}^{(i)}; \theta_f); \theta_d)) \times p(\theta_d|\alpha_d).$$

$p(\theta_f|\alpha_f)$ и $p(\theta_d|\alpha_d)$ являются априорными распределениями параметров функции преобразования и дискриминатора с гиперпараметрами α_f и α_d соответственно. n_d и n_f - размеры батчей для дискриминатора и функции преобразования соответственно.
$\mathbf{X} = \{\mathbf{x}^{(i)}\}_{i=1}^{n_d}$.

Тогда классическая мин-макс функция для GAN переписывается в виде:

$$V(f, G) \approx \mathbb{E}_{x\sim p(x)}[\log D(x)] + \mathbb{E}_{z\sim p(z)}[\log(1 - D(f(z)))] + \log p(\theta_f|\alpha_f) - \log p(\theta_d|\alpha_d)$$

Слайд 5. Постановка задачи для предлагаемого подхода.

Зададим классификатор как нейронную сеть и построим состязательную процедуру обучения. Задача функции преобразования сделать так, чтобы классификатор не мог угадать из какого домена взяты объекты, а задача классификатора научиться угадывать из какого домена взят объект. Чтобы не происходило коллапса моды предлагается добавить обратную функцию преобразования и сравнивать получаемые значения с изначальными по L_2 норме.

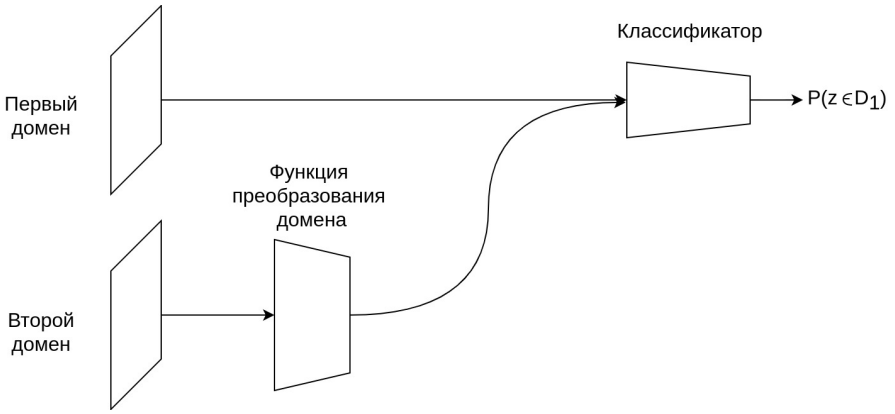

Рис. 2 Предлагаемая архитектура модели для решения задачи.

Для упрощения записи зададим C, f функцию классификатора и функцию преобразования, тогда итоговый вид задачи:

$$L(C, V) = \mathbb{E}_{x\sim\mathcal{D}_1}[\log C(x)] + \mathbb{E}_{x\sim\mathcal{D}_2}[\log(1 - C(f(x)))]$$

Слайд 6. Предлагаемый анализ ошибки.

Хотим:

1. Статистически проверить гипотезу о равенстве весов в задачах регрессии. Модели строятся для разных доменов и для разных функций преобразования.
2. Проверить качество функций преобразования построенных относительно различных функций сходства с помощью точности решения задачи классификации доменов.
3. Для архитектуры с классификатором проверить незначимость каждого признака через априорную матрица корреляции $\mathbf{A} = \mathrm{diag}(\boldsymbol{\alpha})$, которая обладает следующим свойством, если $\boldsymbol{\alpha}_j \to \infty$ это значит, что этот признак не важен. Цель функции преобразования сделать все признаки шумовыми для классификатора. Поэтому хочется проверить, что $\min\limits_i \boldsymbol{\alpha}_i^C \to \infty$.

1 Введение

В последнее время в машинном обучении становятся всё более актуальным задачи такие как zero-shot learning, когда нужно научиться аппроксимировать неизвестное распределение без возможности семплирования из него, и few-shot learning, когда есть небольшой набор семплов неизвестного распределения. В данной работе рассматривается теоретический подход к нахождению совместного распределения по маргинальным, с, возможно, малым размером выборки. Вводится гипотеза о строении совместного распределения всех объектов.

Данная задача решается в пространстве признаков объектов. Также данную задачу можно решать с помощью использования мультимодели, когда мы работаем в пространстве ответов локальных моделей (каждая построена на своем домене).

Определение 5. Под *доменом* понимается подмножество объектов выборки, которые обладают некоторыми одинаковыми признаками.

Когда у нас множество индексов двух доменов не пересекаются, то получается, что объекты этих доменов принадлежат ортогональной подпространствам, то есть проекция распределения одного домена на подпространство другого - функция Дирака и совместное распределение является просто произведением маргинальных. Такая постановка не позволяет найти совместное распределение, поэтому опишем свою гипотезу строения данных.

1.1 Гипотеза об объектах

Считаем, что существует общее пространство объектов (товары в магазине). Есть наблюдаемые параметры этих объектов (для карандашей - цвет грифеля и его мягкость, для книг - объем количество страниц и тип переплета), множество наблюдаемых параметров может как пересекаться (общий параметр для карандашей и книг - цена), так и не пересекаться (цвет грифеля, количество страниц).

1.2 Методы нахождения совместного распределения.

Пусть есть $\mathfrak{X}$ - множество объектов, X_1, X_2 - набор объектов из $\mathfrak{X}$, множества индексов $\mathcal{A}_1, \mathcal{A}_2 \subset \mathcal{J} = \{1, \dots, n\}$ - множество всех индексов.

- Тогда если $\mathcal{A}_1 \cap \mathcal{A}_2 = \varnothing$ или $X_1 \cap X_2 \neq \varnothing$, то расширяя множество индексов для какого-то домена присваиванием какого-то случайного значения из априорного распределения, будем получать разные результаты совместного распределения (у нас нет никакой общей информации для доменов, поэтому матрица корреляции для совместного распределения будет блочно-диагональной, где блоки - матрицы корреляции маргинальных распределений).
- Если $\mathcal{A}_1 \cap \mathcal{A}_2 \neq \varnothing$ и $X_1 \cap X_2 \neq \varnothing$, тогда аппроксимируя (описывая логистической регрессией) левую часть, можем получить правую:

$$\begin{cases} p(\mathcal{A}_1 \backslash \mathcal{A}_2 | \mathcal{A}_1 \cap \mathcal{A}_2) \\ p(\mathcal{A}_2 \backslash \mathcal{A}_1 | \mathcal{A}_1 \cap \mathcal{A}_2) \end{cases} \Rightarrow p(\mathcal{A}_1, \mathcal{A}_2 | \mathcal{A}_1 \cap \mathcal{A}_2)$$

В [1] вводится функция сходства s, определенная на паре распределений $g_1(\mathbf{w})$ и $g_2(\mathbf{w})$, которая удовлетворяет следующим свойствам:

1. определена в случае несовпадения носителей
2. $s(g_1, g_2) \leqslant s(g_1, g_1)$
3. $s \in [0, 1]$
4. $s(g_1, g_1) = 1$

5. близка к 1, если $g_2(\mathbf{w})$ - малоинформативное распределение
6. симметрична, т.е. $s(g_1, g_2) = s(g_2, g_1)$

Где под малоинформативностью понимается следующее:

Определение 6. Назовем распределение $g_2(\cdot) : \Omega \to \mathbb{R}^+$ *неинформативным* относительно распределения $g_1(\cdot) : \Omega \to \mathbb{R}^+$ с конечным носителем $\text{supp}(g_1) = A$, если $\exists B : A \subseteq B$, что

$$\forall \mathbf{w} \in B : g_2(\mathbf{w}) = \frac{1}{|B|},$$

то есть $g_2(\cdot)$ есть равномерное распределение на множестве B.

Обобщим теперь понятие неинформативности на случай двух распределений $g_1(\cdot), g_2(\cdot)$, которые определены на несовпадающих пространствах, то есть $g_1 : \Omega \times \Omega_1 \to \mathbb{R}^+, g2 : \Omega \times \times \Omega_2 \to \mathbb{R}^+$.

Определение 7. Назовем распределение $g_2(\cdot) : \Omega \to \mathbb{R}^+$ *неинформативным* относительно распределения $g_1(\cdot) : \Omega \to \mathbb{R}^+$ с конечным носителем $\text{supp}(g_1) = A$, если $\Omega_1 = \varnothing$, то есть g_1 определено на подпространстве области определения g_2 и $\exists \tau > 0, B : A \times [- -\tau, \tau]^{\dim(\Omega_2)} \subseteq B$, что

$$\forall \mathbf{w} \in B : g2(\mathbf{w}) = \frac{1}{|B|},$$

то есть $g_2(\cdot)$ есть равномерное распределение на множестве B.

Определение 8. Назовем функцией сходства s_score пары распределений $g_1 : \mathbb{R}^n \to \mathbb{R}^+ +, g_2 : \mathbb{R}^n \to \mathbb{R}^+$, определенных на одном пространстве, функцию вида

$$s_0(g_1, g_2) = \frac{\int g_1(\mathbf{w}) g_2(\mathbf{w}) d\mathbf{w}}{\max\limits_{\mathbf{b} \in \mathbb{R}^n} \int g_1(\mathbf{v}) g_2(\mathbf{v} - \mathbf{b}) d\mathbf{v}}$$

В [1] доказывается, что введенная функция сходства удовлетворяет вышеперечисленным свойствам, в отличие от дивергенциями Кульбака-Лейблера, f-дивергенциями, Брегмана и симметризованной дивергенцией Брегмана, а также расстояниями Дженсона-Шеннона, Хеллингера, Бхаттачарайа.

Лемма 1. Пусть заданы распределения $\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1), \mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2)$. Тогда

$$\text{s}_\text{s}\text{core}(\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1), \mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2)) = \exp\Big(-\frac{1}{2}(\mathbf{m}_1 - \mathbf{m}_2)^T (\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-1} (\mathbf{m}_1 - \mathbf{m}_2) \Big)$$

Доказательство. Обозначим g_1, g_2 как соответствующие функции плотности вероятности распределений, тогда

$$\text{s}_\text{s}\text{core}(\mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1), \mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2)) = s_0(g_1, g_2) = \frac{\int g_1(\mathbf{w}) g_2(\mathbf{w}) d\mathbf{w}}{\max\limits_{\mathbf{b} \in \mathbb{R}^n} \int g_1(\mathbf{v}) g_2(\mathbf{v} - \mathbf{b}) d\mathbf{v}}$$

Числитель и знаменатель дроби, используя симметричность нормального распределения, можно переписать через формулу свертки.

Пусть $\xi_1 \sim \mathcal{N}(\mathbf{m}_1, \mathbf{\Sigma}_1), \xi_2 \sim \mathcal{N}(\mathbf{m}_2, \mathbf{\Sigma}_2)$, тогда

$$\mathbb{P}_{\boldsymbol{\xi_1 - \xi_2}}(\mathbf{x}) = \int g_1(\mathbf{w}) g_2(\mathbf{w} + \mathbf{x}) d\mathbf{w}$$

Отсюда получаем, что

$$s_0(g_1, g_2) = \frac{\mathbb{P}_{\xi_1 - \xi_2}(\mathbf{0})}{\max\limits_{\mathbf{b}\in\mathbb{R}^n} \mathbb{P}_{\xi_1-\xi_2}(-\mathbf{b})} = \exp\Big(-\frac{1}{2}(\mathbf{m}_1 - \mathbf{m}_2)^T(\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2)^{-1}(\mathbf{m}_1 - \mathbf{m}_2)\Big)$$

■

1.3 Байесовская логистической регрессия

Пусть $\mathbf{X} \in \mathbb{R}^{m\times n}$ - признаковая матрица, а $\mathbf{y} \in \{1, -1\}^m$ - метки класса.

Тогда совместное распределение имеет вид:

$$p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{A}) = p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathbf{A}), \text{ где } p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{j=1}^{m} \sigma(y_j \mathbf{w}^T \mathbf{x}_j). \tag{1}$$

$p(\mathbf{w}|\mathbf{A}) = \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}), \mathbf{A} = \mathrm{diag}(\boldsymbol{\alpha})$ - априорное распределение на $\mathbf{w}$, гиперпараметр $\boldsymbol{\alpha}$ можем найти используя метод максимизации обоснованности.

Вычисление обоснованности

Тогда

$$\mathbf{A}^* = \arg\max_{\mathbf{A}} p(\mathbf{y}|\mathbf{X}, \mathbf{A}) = \arg\max_{\mathbf{A}} \int \underbrace{p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{A})}_{Q(\mathbf{w})} d\mathbf{w}$$

Так как данный интеграл аналитически не вычисляется, поэтому можно воспользоваться аппроксимацией Лапласа:

$$\log Q(\mathbf{w}) \approx \log Q(\mathbf{w}_{MAP}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{MAP})^T \overbrace{\nabla^2 \log Q(\mathbf{w}_{MAP})}^{-\mathbf{H}^{-1}} (\mathbf{w} - \mathbf{w}_{MAP}).$$

Выпишем $\mathbf{H}^{-1}$ явно:

$$\begin{aligned}
\mathbf{H}^{-1} = \nabla^2(-\frac{1}{2}\mathbf{w}^T\mathbf{A}^{-1}\mathbf{w} - \mathbf{1}^T \log(\mathbf{1} + \exp(\mathbf{y} \odot \mathbf{X}^T\mathbf{w}))) = \\
= \mathbf{A}^{-1} + \nabla^2 \mathbf{1}^T \log(\mathbf{1} + \exp(\mathbf{y} \odot \mathbf{X}^T\mathbf{w})) = \\
= \mathbf{A}^{-1} + \sum_{i=1}^{m} \nabla^2 \log(1 + \exp(y_i \mathbf{x}_i^T \mathbf{w})) = \\
= \mathbf{A}^{-1} + \sum_{i=1}^{m} \mathbf{x}_i\mathbf{x}_i^T \frac{\exp(y_i\mathbf{x}_i^T\mathbf{w})}{1 + \exp(y_i\mathbf{x}_i^T\mathbf{w})} - \mathbf{x}_i\mathbf{x}_i^T \frac{\exp(2y_i\mathbf{x}_i^T\mathbf{w})}{(1 + \exp(y_i\mathbf{x}_i^T\mathbf{w}))^2}
\end{aligned}$$

В итоге получаем:

$$\mathbf{A}^* = \arg\max_{\mathbf{A}} \Big(Q(\mathbf{w}_{MAP}) \int e^{-\frac{1}{2}(\mathbf{w}-\mathbf{w}_{MAP})^T\mathbf{H}^{-1}(\mathbf{w}-\mathbf{w}_{MAP})} d\mathbf{w} \Big) =$$

$$\arg\max_{\mathbf{A}} \Big(Q(\mathbf{w}_{MAP})(2\pi)^{n/2} \det(\mathbf{H})^{1/2} \Big) \tag{2}$$

Оптимальная $\mathbf{A}$ находится итеративно. Сначала производим итерацию по $\mathbf{w}_{MAP}$, максимизируя совместное распределение (1), затем происходит итерация по $\mathbf{A}$, которое находится из (2). Но так как аппроксимация Лапласа дает менее точный результат, чем VLB, поэтому в дальнейшем будет пользоваться вариационной нижней оценкой.

Для сигмоидной функции существует VLB:

$$\sigma(x) \geqslant \sigma(\xi)\exp\Big(-\frac{1}{4\xi}(2\sigma(\xi)-1)(x^2-\xi^2)+\frac{x-\xi}{2}\Big).$$

Применим нижнюю оценку к совместному распределению:

$$p(\mathbf{y},\mathbf{w}|\mathbf{X},\mathbf{A}) = \prod_{j=1}^{m}\sigma(y_j\mathbf{w}^T\mathbf{x}_j)\frac{\sqrt{\det\mathbf{A}}}{(2\pi)^{n/2}}e^{-\frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}} \geqslant \mathrm{VLB}(\mathbf{w},\boldsymbol{\xi},\mathbf{A}) =$$
$$\frac{\sqrt{\det\mathbf{A}}}{(2\pi)^{n/2}}e^{-\frac{1}{2}\mathbf{w}^T\mathbf{A}\mathbf{w}}\prod_{j=1}^{m}\sigma(\xi_j)\exp\Big(-\frac{2\sigma(\xi_j)-1}{4\xi_j}(\mathbf{w}^T\mathbf{x}_j\mathbf{x}_j^T\mathbf{w}-\xi_j^2)+\frac{y_j\mathbf{w}^T\mathbf{x}_j-\xi_j}{2}\Big) =$$
$$\frac{\sqrt{\det\mathbf{A}}}{(2\pi)^{n/2}}\prod_{j=1}^{m}\sigma(\xi_j)\exp\Big(\frac{2\sigma(\xi_j)-1}{4\xi_j}\xi_j^2-\frac{\xi_j}{2}\Big)\exp\Big(-\frac{1}{2}\mathbf{w}^T\mathbf{A}'\mathbf{w}+\mathbf{w}^T\mathbf{v}\Big) \quad (3)$$

где $\mathbf{A}' = \mathbf{A}+\sum_{j=1}^{m}\frac{2\sigma(\xi_j)-1}{2\xi_j}\mathbf{x}_j\mathbf{x}_j^T$, $\mathbf{v} = \frac{1}{2}\sum_{j=1}^{m}y_j\mathbf{x}_j$.

Тогда

$$p(\mathbf{y}|\mathbf{X},\mathbf{A}) \geqslant \mathrm{LB}(\mathbf{A},\boldsymbol{\xi}) = \int \mathrm{VLB}(\mathbf{w},\boldsymbol{\xi},\mathbf{A})d\mathbf{w} \to \max_{\mathbf{A},\boldsymbol{\xi}}.$$

$$\mathrm{LB}(\mathbf{A},\boldsymbol{\xi}) = \sqrt{\frac{\det\mathbf{A}}{\det\mathbf{A}'}}\prod_{j=1}^{m}\sigma(\xi_j)\exp\Big(\frac{2\sigma(\xi_j)-1}{4\xi_j}\xi_j^2-\frac{\xi_j}{2}\Big)\exp\Big(\frac{1}{2}\mathbf{v}^T\mathbf{A}'^{-1}\mathbf{v}\Big)$$

1.4 Домен-инвариантный автокодировщик

Задача домен-инвариантного автокодировщика заключается в том, чтобы найти такое латентное представление z для каждого x, чтобы $g_1 = p(z|z \in D_1)$ было как можно ближе к $g_2 = p(z|z \in D_2)$, это обеспечивается добавление состязательной процедуры между классификатором и кодировщиком, которые последовательно модифицируют эти распределения.

Определение 9. Назовем *домен-инвариантным* признаковым описанием объекта, такое описание, что его распределение не зависит от домена, при это содержит достаточно информации для описания объекта.

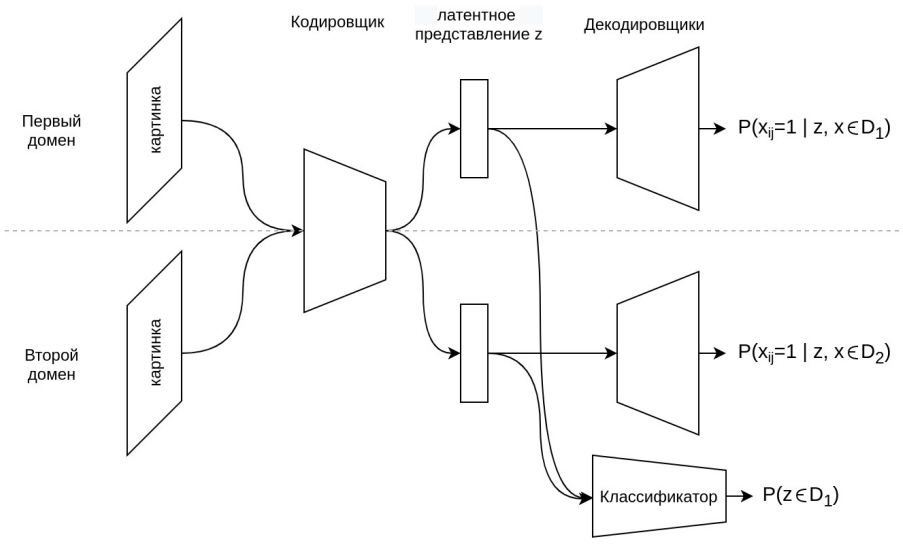

Рис. 3 Предлагаемая архитектура модели для решения задачи.

Классификатор, кодировщик и декодировщик являются моделями для логистической регрессии, с весами $\mathbf{w}^C \in \mathbb{R}^m$, $\mathbf{w}^E \in \mathbb{R}^{n\times m}$, $\mathbf{w}^D \in \mathbb{R}^{m\times n}$ соответственно. m - размерность латентного пространства, n - размерность пространства $\mathbf{x}$.

Задача классификатора научиться определять по латентному коду $\mathbf{z}_i$ из какого домена он взят, поэтому $y_i = 1$, если $\mathbf{z}_i$ соответствует первому домена и $y_i = -1$ иначе:

$$p(\mathbf{y}, \mathbf{w}^C|\mathbf{A}^C, \mathbf{Z}) = \prod_i p(\mathbf{z}_i)\mathcal{N}(\mathbf{w}^C|\mathbf{0}, \mathbf{A}^{C^{-1}})p(y_i|\mathbf{z}_i, \mathbf{w}^C)$$

где $p(y_i|\mathbf{z}_i, \mathbf{w}^C) = \frac{1}{1+\exp(y_i \mathbf{w}^{C^T}\mathbf{z}_i)}$.

По формуле Байеса:

$$p(\mathbf{w}^C|\mathbf{Z}, \mathbf{y}, \mathbf{A}^C) = \frac{p(\mathbf{y}, \mathbf{w}^C|\mathbf{Z}, \mathbf{A}^C)}{p(\mathbf{y}|\mathbf{Z}, \mathbf{A}}$$

Поэтому мы можем оценить $\mathbf{w}^C_{MAP}$:

$$\mathbf{w}^C_{MAP} = \arg\max_{\mathbf{w}\in\mathbb{R}^m} p(\mathbf{w}^C|\mathbf{Z}, \mathbf{y}, \mathbf{A}^C) = \arg\min_{\mathbf{w}\in\mathbb{R}^m} (-\log p(\mathbf{y}|\mathbf{Z}, \mathbf{w}^C) - \log p(\mathbf{w}|\mathbf{A}^C)) \tag{4}$$

где

$$p(\mathbf{y}|\mathbf{Z}, \mathbf{w}^C) = \prod_{j=1} \hat{p}_i^{y_i}(1-\hat{p}_i)^{1-y_i}; \quad -\log p(\mathbf{w}|\mathbf{A}) \propto \frac{1}{2}\mathbf{w}^T\mathbf{A}^{C^{-1}}\mathbf{w}; \quad \hat{\mathbf{p}}^C = \frac{1}{1+\exp(-\mathbf{Z}^T\mathbf{w}^C)} \tag{5}$$

Подставляя (5) в (4), получаем, что (4) можно переписать через кросс-энтропию:

$$H\big(p(z \in \mathcal{D}_1)|\hat{\mathbf{p}}^C\big) + \frac{1}{2}\mathbf{w}^T\mathbf{A}^{C^{-1}}\mathbf{w} \to \min_{\mathbf{w}\in\mathbb{R}^m}$$

Для декодировщика строится такая же задача:

$$\sum_{i,j} H\big(p(x_{ij})|\hat{\mathbf{p}}^D_{ij}\big) + \frac{1}{2}\mathbf{w}^T\mathbf{A}^{D^{-1}}\mathbf{w} \to \min_{\mathbf{w}\in\mathbb{R}^{m\times n}}$$

где $\hat{\mathbf{p}}^D = \frac{1}{1+\exp(-\mathbf{Z}^T\mathbf{w}^D)}$

В итоге задача нахождения домен-инвариантного признакового описания сводится к итеративной максимизации двух выражений (6), (7):

Оптимизация параметров классификатора и декодировщика:

$$\mathbf{w}^C_{MAP}, \mathbf{w}^{D_1}_{MAP}, \mathbf{w}^{D_2}_{MAP} = \arg \max_{\mathbf{w}^C \in \mathbb{R}^m, \mathbf{w}^{D_1} \in \mathbb{R}^{m\times n}, \mathbf{w}^{D_2} \in \mathbb{R}^{m\times n}} \sum_{\mathbf{x}\in\mathcal{D}_1} H\big(p(z \in \mathcal{D}_1)|\hat{\mathbf{p}}^C\big) + \sum_{i,j} H\big(p(\mathbf{x}_{ij})|\hat{\mathbf{p}}^{D_1}_{ij}\big) + \sum_{\mathbf{x}\in\mathcal{D}_2} H\big(p(z \in \mathcal{D}_2)|\hat{\mathbf{p}}^C\big) + \sum_{i,j} H\big(p(\mathbf{x}_{ij})|\hat{\mathbf{p}}^{D_2}_{ij}\big) + \frac{1}{2}\mathbf{w}^{C^T}\mathbf{A}^{C^{-1}}\mathbf{w}^C + \frac{1}{2}\mathbf{w}^{D_1^T}\mathbf{A}^{C^{-1}}\mathbf{w}^{D_1} + \frac{1}{2}\mathbf{w}^{D_2^T}\mathbf{A}^{C^{-1}}\mathbf{w}^{D_2} \quad (6)$$

Оптимизация параметров кодировщика:

$$\mathbf{w}^E_{MAP} = \arg \max_{\mathbf{w}^E \in \mathbb{R}^{n\times m}} \sum_{\mathbf{x}\in\mathcal{D}_1} -H\big(p(z \in \mathcal{D}_1)|\hat{\mathbf{p}}^C\big) + \sum_{i,j} H\big(p(\mathbf{x}_{ij})|\hat{\mathbf{p}}^{D_1}_{ij}\big) + \sum_{\mathbf{x}\in\mathcal{D}_2} -H\big(p(z \in \mathcal{D}_2)|\hat{\mathbf{p}}^C\big) + \sum_{i,j} H\big(p(\mathbf{x}_{ij})|\hat{\mathbf{p}}^{D_2}_{ij}\big) \quad (7)$$

где $\mathbf{z} = \sigma(\mathbf{X}^T\mathbf{w}^E)$

2 Методы проверки качества домен-инвариантного представления

2.1

Статистически проверить гипотезу о равенстве весов в задачах линейной регрессии определения размера и положения фигуры. Модели строятся для разных доменов (квадраты/круги/треугольники).

2.2

Статистически проверить гипотезу о равенстве весов в задачах линейной регрессии определения размера и положения фигуры. Модели строятся для разных доменов (квадраты/повернутые квадраты - ромбы).

2.3

Матрица корреляции обладает свойством отбора признаков, если $\boldsymbol{\alpha}_j \to \infty$ это значит, что этот признак не важен. Мотивация кодировщика найти такой $\mathbf{z}$, что все его компоненты будут шумовыми для классификатора, т.е. для переменной $\mathbf{y}$. Поэтому хочется проверить, что $\min\limits_i \boldsymbol{\alpha}^C_i \to \infty$.

При этом гиперпараметры декодировщика не должны обладать такими свойствами.

3 Задача аппроксимации параметров изображения

3.1 Описание данных

Дана бинарная картинка:

$$\mathbf{M} \in \{0,1\}^{m_1\times m_2}$$

где 1 отвечает черному пикселю картинки, 0 — белому.

3.2 Описание картинки с прямоугольником

На каждой картинке находится прямоугольник полученный следующим образом:

- Пусть $a, b \sim U[a_0, b_0]$ — две случайные величины, которые отвечают за размер прямоугольника. a_0, b_0 - известные границы возможных размеров прямоугольника.
- Пусть $x_0 \sim U[a, m_1 - a], y_0 \sim U[b, m_2 - b]$ — два числа, которые отвечают за положение прямоугольника.

Введем понятие *изображения* $\mathbf{C}$ - набор координат ненулевых пикселей x_i, y_i:

$$\mathbf{C} \in \mathrm{R}^{N \times 2}$$

где $N = m_1 * m_2$ - число пикселей картинки.

Тогда прямоугольник можно задать следующими уравнениями:

$$\left|\frac{x - x_0}{a} + \frac{y - y_0}{b}\right| + \left|\frac{x - x_0}{a} - \frac{y - y_0}{b}\right| = 1$$

либо:

$$\left\|\frac{x - x_0}{a}, \frac{y - y_0}{b}\right\|_\infty = \max\left(\frac{|x - x_0|}{a}, \frac{|y - y_0|}{b}\right) = 1$$

3.3 Описание картинки с эллипсом

На каждой картинке находится прямоугольник полученный следующим образом:

- Пусть $a, b \sim U[a_0, b_0]$ — две случайные величины, которые отвечают за размер полуосей эллипса. a_0, b_0 - известные границы возможных размеров.
- Пусть $x_0 \sim U[a, m_1 - a], y_0 \sim U[b, m_2 - b]$ — два числа, которые отвечают за положение эллипса.

Тогда эллипс можно задать следующим уравнением:

$$\left(\frac{x - x_0}{a}\right)^2 + \left(\frac{y - y_0}{b}\right)^2 = 1$$

3.4 Постановка задачи

Теорема 2. Пусть есть набор пар, доменов, $\{(\mathbf{X_1}, \mathbf{y_1}), \dots (\mathbf{X_k}, \mathbf{y_k})\}$ $\mathbf{X_i} \in \mathbb{R}^{n \times m_i}, \mathbf{y_i} \in \mathbb{R}^{m_i}$. И выполнен ряд условий:

- $\forall i \in [1, k]$ $\exists \mathbf{w_i} \in \mathbb{R}^n$: $\mathbf{y_i} = \mathbf{X_i w_i}$. Т.е. для каждого домена существует локальная аппроксимирующая модель.
- $\forall i \neq j, \mapsto \mathbf{X_i w_j} = \mathbf{0}$

Тогда, существует единая модель для всех доменов, веса которой зависят только от $\{\mathbf{w_1}, \dots, \mathbf{w_k}\}$.

Доказательство.

$$\mathbf{X_j}\Big(\sum_{i=1}^{n} \mathbf{w_i}\Big) = \underbrace{\mathbf{X_j w_1} + \dots \mathbf{X_j w_{j-1}}}_{=0} + \mathbf{X_j w_j} + \underbrace{\mathbf{X_{j+1} w_1} + \dots \mathbf{X_j w_n}}_{=0} = \mathbf{y_j}$$

■

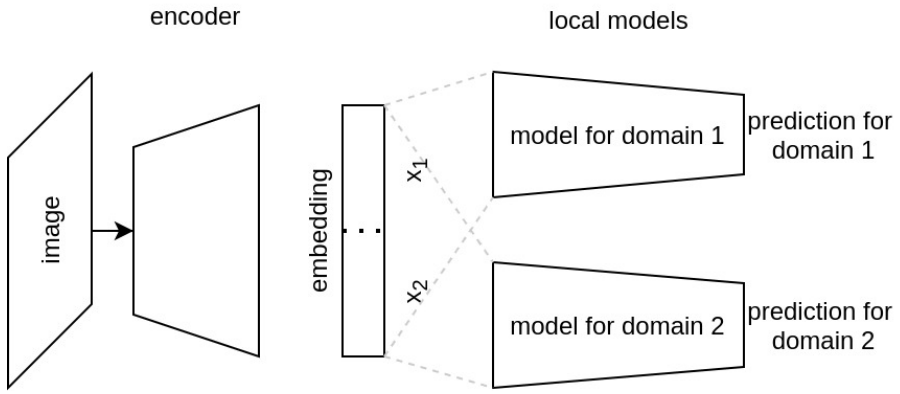

Рис. 4 Предлагаемая архитектура модели для решения задачи.

3.5 Анализ алгебраической структуры пространства домена изображений

Рассмотрим пространство изображений с прямоугольниками.

Определение 10. Назовем изображение с прямоугольником *запрещенным*, если фигура "разделяется"на 2 части границами изображения.

Определение 11. Назовем изображение с прямоугольником *разрешенным*, если оно не является запрещенным.

а) б)

Рис. 5 Примеры (а) запрещенного и (б) разрешенного изображений с прямоугольником.

На пространстве изображений с прямоугольником определены следующие операции, переводящие пространство в само себя:

- Сдвиг фигуры влево/вправо - циклическая перестановка на 1 базисных векторов.
- Увеличение ширины фигуры.

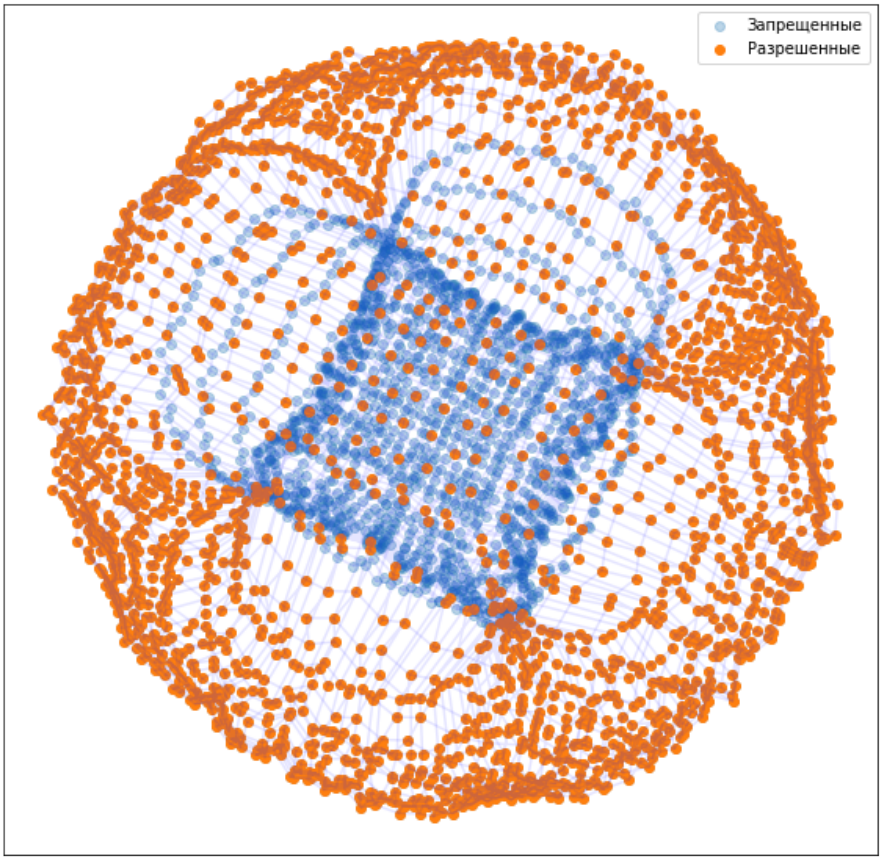

Рис. 6 Первые 2 компоненты PCA разложения пространства изображений прямоугольников для всех возможных положений с фиксированным размером фигуры.

Литература

[1] *А.А. Адуенко*. Выбор мультимоделей в задачах классификации, 2017 `http://www.frccsc.ru/sites/default/files/docs/ds/002-073-05/diss/11-aduenko/11-Aduenko_main.pdf`

[2] *Manuel Pérez-Carrasco and Guillermo Cabrera-Vives and Pavlos Protopapas and Nicolas Astorga and Marouan Belhaj*, Adversarial Variational Domain Adaptation, 2019, CoRR

[3] *Garrett Wilson and Diane J. Cook*, Adversarial Transfer Learning, 2018, CoRR

[4] *Jing Wang and Jiahong Chen and Jianzhe Lin and Leonid Sigal and Clarence W. de Silva*, Discriminative Feature Alignment: Improving Transferability of Unsupervised Domain Adaptation by Gaussian-guided Latent Alignment, 2020

[5] *Jing Jiang*, A Literature Survey on Domain Adaptation of Statistical Classifiers, 2008

[6] *А. В. Грабовой, В. В. Стрижов*, Анализ выбора априорного распределения для смеси экспертов

[7] *Guo, Jiang and Shah, Darsh J and Barzilay, Regina*, Multi-Source Domain Adaptation with Mixture of Experts, 2018, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, `http://aclweb.org/anthology/D18-1498`

[8] *Seniha Esen Yuksel; Joseph N. Wilson; Paul D. Gader*, Twenty Years of Mixture of Experts, 2012, IEEE Transactions on Neural Networks and Learning Systems

Определение 12. Назовем *p-ым расстоянием Васерштейна* $W_p(\mu, v)$ между двумя вероятностными мерами μ и v:

$$W_p(\mu, v) := \Big(\inf_{\gamma \in \Gamma(\mu,v)} \int d(x,y)^p d\gamma(x,y)\Big)^{1/p}$$

где $\Gamma(\mu, v)$ обозначает совокупность всех мер с маргинальными распределениями μ и v для первого и второго параметров соответственно.

Замечание 1. p-е расстоянием Васерштейна в случае распределений можно переписать в виде:

$$W_p = \Big(\inf_{\gamma \in \Gamma(\mu,v)} \mathbb{E}_{(x,y)\sim\gamma}\|x-y\|_p\Big)^{1/p}$$

Введем функцию сходства порожденную расстоянием Васерштейна:

$$s_{Wp}(g_1, g_2) = \exp(-W_p(g_1, g_2))$$

Определение 13 (по Адуенко). Назовем распределение $g_2(\cdot) : \Omega \times \Omega_2 \to \mathbb{R}^+$ *неинформативным* относительно распределения $g_1(\cdot) : \Omega \times \Omega_1 \to \mathbb{R}^+$ с конечным носителем $\mathrm{supp}(g_1) = A$, если $\Omega_1 = \varnothing$, то есть g_1 определено на подпространстве области определения g_2 и $\exists\ \tau > 0, B : A \times [-\tau, \tau]^{\dim(\Omega_2)} \subseteq B$, что

$$\forall\ \mathbf{v} \in B : g_2(\mathbf{w}) = \frac{1}{|B|},$$

то есть $g_2(\cdot)$ есть равномерное распределение на множестве B.

Замечание 2. Из определения неинформативного распределения следует, что распределение не является неинформативным относительно самого себя. Например, сужение нормального распределения на отрезок $[-1, 1]$, $\mathcal{N}(0,1)|_{[-1,1]}$ не будет равномерным распределением на отрезке $[-1, 1]$.

Определение 14 (по Адуенко). Назовем последовательность распределений $g_2^1(\cdot\ \cdot), \ldots, g_2^k(\cdot), \cdots \to \mathbb{R}^+$ *малоинформативной* на Ω, если выполнены следующие условия

$$\forall\ a > 0, g_2^k(\cdot)|_A \to U(A),\ \text{где}\ A = \{\mathbf{w} : \|\mathbf{w}\| \leqslant a\},$$
$$\exists\ 0 \leqslant B < \infty,\ \exists\ k_0 :\ \forall\ k \geqslant k_0 \sup_{\{\mathbf{w}:\|\mathbf{w}\|\geqslant B\}} g_2^k(\mathbf{w}) \leqslant \sup_{\{\mathbf{w}:\|\mathbf{w}\|\leqslant B\}} g_2^k(\mathbf{w}),$$

где $g_2^k(\cdot)|_A$ есть сужение распределения $g_2^k(\cdot)$ на множество A, то есть

$$g_2^k|_A(\mathbf{w}) = \begin{cases} 0,\ \textit{если}\ \|\mathbf{w}\| > a, \\ \frac{g_2^k(\mathbf{w})}{\int_A g_2^k(\mathbf{v})dv}, w \in A. \end{cases}$$

Сходимость понимается равномерная, то есть

$$g_2^k(\cdot)|_A \to U(A) \Leftrightarrow \sup_{\mathbf{w}\in A} |g_2^k(\mathbf{w})1/|A|| \to 0\ \textit{при}\ k \to \infty.$$

Определение 15 (по Адуенко). Функция сходства $s(g_1, g_2)$, определенная на паре распределений $g_1(\mathbf{w}) : \Omega \times \Omega_1 \to \mathbb{R}^+$ и $g_2(\mathbf{w}) : \Omega \times \Omega_2 \to \mathbb{R}^+$, где $\Omega = \mathbb{R}^n$, $\Omega_1 = \mathbb{R}^{n_1}$, $\Omega_2 = \mathbb{R}^{n_2}$, $n, n_1, n_2 \geqslant 0$ называется *корректной*, если удовлетворяет следующим требованиям.

1. $s(g_1, g_2)$ определена $\forall\ n, n_1, n_2 \geqslant 0$, если $g_1(\mathbf{w}, \mathbf{w}_1), g_2(\mathbf{w}, \mathbf{w}_2) < \infty\ \forall\ \mathbf{w} \in \Omega, \mathbf{w}_1 \in \Omega_1, \mathbf{w}_2 \in \Omega_2$,
2. $s(g_1, g_2) \in [0, 1]$,
3. $s(g_1, g_2) = s(g_2, g_1)$,
4. $s(g_1, g_2) \leqslant s(g_1, g_1)$,
5. $s(g_1, g_1) = 1$,
6. a. Если g_2 является неинформативным относительно g_1, то $s(g_1, g_2) = 1$;
   b. $\forall\ g_1 : n_1 = 0$, то есть $g_1(\mathbf{w}, \mathbf{w}_1) = g_1(\mathbf{w})$, для малоинформативной последовательности распределений $g_2^1, \ldots, g_2^k, \ldots$ выполнено

$$s(g_1, g_2^k) \to 1 \text{ при } k \to \infty$$

Теорема 3. Функция сходства порожденная расстоянием Васерштейна не является корректной в терминах Адуенко.

Доказательство.

Покажем, что требование 6 не выполняется для функции сходства порожденной расстоянием Васерштейна. Будем доказывать от противного.

Возьмем распределение $g_2 = U[-2, 2]$ неинформативное относительно $g_1 = U[-1, 1]$, тогда должно выполняться $s_{Wp}(g_1, g_2) = 1$, что эквивалентно $W_p(g_1, g_2) = 0$. g_2 является неинформативным относительно g_1, потому что $\text{supp}(g_1) = [-1, 1] = A$ и
$\exists\ \tau = 0.5 > 0, B : A \times [-\tau, \tau]^0 \subseteq B$. При этом выполнено, что g_2 равномерное распределение на множестве B.

$W_p(g_1, g_2) = 0 \Rightarrow \exists \gamma \sim \Gamma(g_1, g_2) : \mathbb{E}[\|x - y\|_p] = 0 \Leftrightarrow P(x = x_0 | y = y_0) = 1$. Последнее утверждение некорректно. Так как событие $(x = x_0 | y = x_0)$ является сужением элемента сигма алгебры $x = x_0$ на область $y = x_0$, то при $x_0 = 1.5$ событие $x = x_0$ имеет меру 0, поэтому и сужение не может иметь меру больше нуля. Это утверждение приводит к противоречию и доказывает теорему.

■

Определение 16. Назовем *δ-функцией сходства* функцию сходства, которая различает любую пару несовпадающих распределений, то есть $s_\delta(g_1, g_2) = [g1 \equiv g2]$.

Теорема 4. δ-функция сходства не является корректной в терминах Адуенко.

Доказательство.

Доказательство очевидно, потому что на несовпадающих распределениях δ-функция всегда дает значение 0, что противоречит пункту 6а в определении корректной функции сходства.
■

Определение 17. Функция $s(g_1, g_2)$, определенная на паре распределений $g_1(\mathbf{w}) : \Omega \times \times \Omega_1 \to \mathbb{R}^+$ и $g_2(\mathbf{w}) : \Omega \times \Omega_2 \to \mathbb{R}^+$, где $\Omega = \mathbb{R}^n,\ \Omega_1 = \mathbb{R}^{n_1},\ \Omega_2 = \mathbb{R}^{n_2},\ n, n_1, n_2 \geqslant 0$ называется *функцией сходства*, если удовлетворяет следующим требованиям.

1. $s(g_1, g_2)$ определена $\forall\ n, n_1, n_2 \geqslant 0$, если $g_1(\mathbf{w}, \mathbf{w}_1), g_2(\mathbf{w}, \mathbf{w}_2) < \infty\ \forall\ \mathbf{w} \in \Omega, \mathbf{w}_1 \in \Omega_1, \mathbf{w}_2 \in \Omega_2$,
2. $s(g_1, g_2) \in [0, 1]$,
3. $s(g_1, g_2) = s(g_2, g_1)$,

4. $\begin{cases} \mathrm{s}(\mathrm{g}_1, g_2) < s(g_1, g_1) & \text{если } \mathrm{g}_1 \not\equiv g_2, \\ \mathrm{s}(\mathrm{g}_1, g_1) = 1 & \text{если } \mathrm{g}_1 \equiv g_2 \end{cases}$

Замечание 3. Корректная функция сходства по Адуенко не является функцией сходства.

Доказательство.

Для доказательства достаточно рассмотреть рассмотреть пункт 6а из определения корректной функции сходства и пункт 4 из определения функции сходства. Затем применить подход из доказательства теоремы (3). ■