«Московский физико-технический институт (национальный исследовательский университет)»
Физтех-школа прикладной математики и информатики
Кафедра «Интеллектуальные системы»

Шокоров Вячеслав Александрович

Оценка параметров вероятностной модели в задаче доменной адаптации

03.03.01 – Прикладные математика и физика

Выпускная квалификационная работа бакалавра

Научный руководитель:
д.ф.-м.н. Стрижов Вадим Викторович

Москва
2021

Аннотация

Методы классического машинного обучения хорошо работают, когда обучающие и тестовые данные взяты из одного и того же распределения. Однако во многих ситуациях у нас могут быть только обучающие данные для исходного домена, задача заключается в обучении модели, которая будет хорошо работать на целевом домене с другим распределением. Какими функциями сходства распределения и каким методом мы можем адаптировать модель, обученный в исходном домене для использования на целевом домене? Интуитивно понятно, что хорошее представление признаков является решающим фактором успеха адаптации домена.

Предлагаемый подход вдохновлен теорией адаптации домена, предполагающей, что для достижения эффективной трансформации домена, функция преобразования должна делать домены неразличимыми, это достигается максимизацией функции сходства. Подход реализует эту идею в контексте архитектур нейронных сетей.

Ключевые слова: *доменная адаптация, нейронная сеть, GAN, WGAN, функция сходства Адуенко. А. А.*

Содержание

1 Введение

1.1 Введение

В данной раб решается задача доменной адаптации. Цель 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 сгенерировать без значительных затрат и которые будут иметь хорошую, качественную разметку, а целевой домен — фотографии пользователей. Тогда задача доменной адаптации заключается в обучении модели на синтетических данных, которая будет хорошо работать с «реальными» объектами.

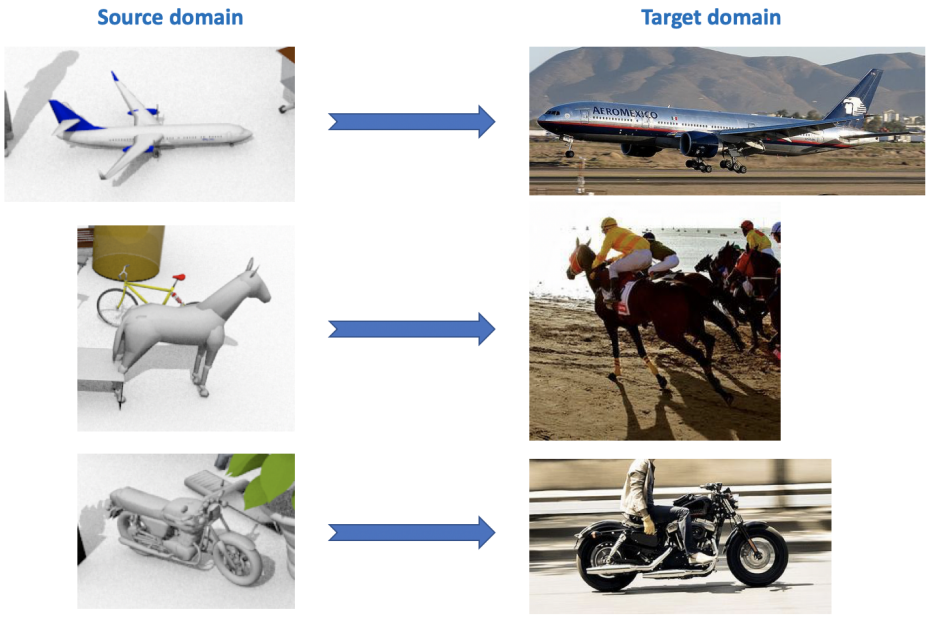


Рис. 1: Примеры задачи применения адаптации доменов. Изображения взяты из датасета VisDA, который используется в конкурсе Visual Domain Adaptation Challenge.

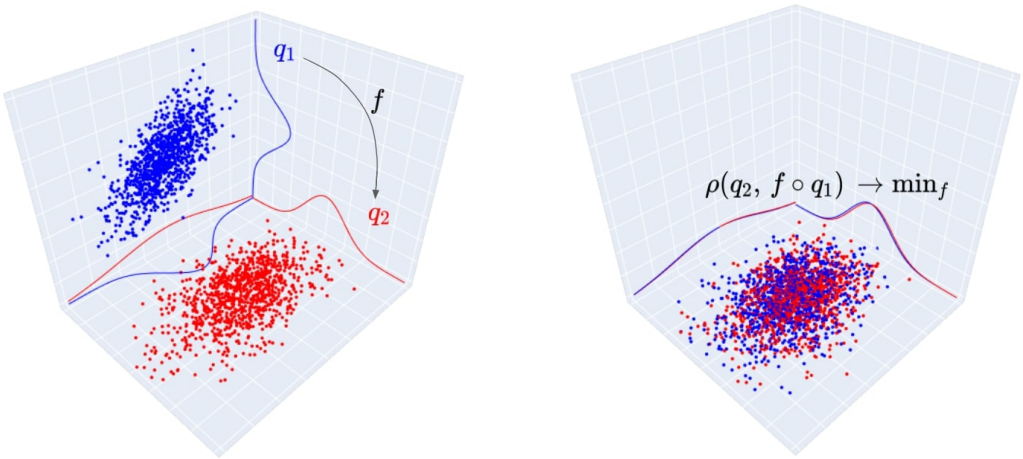
Данную задачу предлагается решить через построение функции преобразования одного домена в другой, максимизируя функцию сходства. Мы выдвигаем и проверяем на вычислительном эксперименте гипотезу о том, что, если распределения доменов схожи, то и веса моделей обученных на них будут совпадать. Это утверждение является мотивацией подхода.

Исследуется проблема построения и анализа вероятностного пространства параметров этого преобразования. Проблема усложнена тем, что домены могут принадлежать непересекающимся или слабо пересекающихся пространствам. Очевидно, что в таком случае, без использования функции преобразования, классические подходы

машинного обучения, когда обучается модель на исходном домене и применяется к целевому домену, не имеет смысла. Случай частично-ортогональных признаковых пространств доменов возможен, например, когда рассматриваются товары в магазине. Есть наблюдаемые параметры этих объектов (для карандашей - цвет грифеля и его мягкость, для книг — 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). Подходы обучения без учителя также дают недостаточное качество, так как не используют информацию, которую можно получить из первого домена.

Определение 1 *Доменом называется априорное распределение признакового описания объектов.*

Определение 2 *Функция f называется функцией преобразования домена $\mathcal{D}_1$ в домен $\mathcal{D}_2$, если* $f : \text{supp}(\mathcal{D}_1) \to \text{supp}(\mathcal{D}_2)$



(a) исходная выборка (b) преобразованная выборка

Рис. 2: Упрощенный пример-визуализация применения предлагаемого подхода. (a) Два домена, первый (целевой, синий) описывается признаками из пространства z-x, второй (исходный, красный) из пространства x-y. Для каждого домена задано распределение. (b) для первого домена применена функция преобразования. Параметры преобразования находятся при решении задачи минимизации функции сходства, которая сравнивает второй домен и преобразованный первый.

1.2 Обзор литературы

Существует большой набор методов доменной адаптации. Данная работа посвящена применению Adversarial-Based подходов. Эти подходы используют состязательную функцию ошибки, которая впервые появилась в GAN'ах.

Ключевой особенностью методов из этого семейства является обучение нейронной сети с инвариантным по отношению к исходному и целевому доменам векторным представлением. Тогда обученную на размеченном исходном домене сеть можно будет использовать на целевом домене, в идеале — практически без потери качества классификации.

В работе [1] предлагается применение трех моделей: основной сети, с помощью которой получается векторное представление , "головы", отвечающей за классификацию на исходном домене и "головы", которая обучается отличать данные из исходного домен от целевого. Особенностью данного подхода является применение Gradient reversal layer при обратном распространении ошибки в обучении для "головы", отвечающей за домены. Этот добавочный слой умножает проходящий через него градиент на негативную константу, увеличивая функцию ошибки связанную с доменном. Этим добиваются того, что распределения векторных представлений на обоих доменах становятся близки.

В подходе под названием Adversarial Discriminative Domain Adaptation (ADDA), описанном в [2], применяется разделение сети для исходного домена и сети для целевого домена. Суть ADDA заключается в том, что мы сначала обучаем хороший классификатор на размеченном исходном домене, а затем с помощью состязательного-обучения адаптируем так, чтобы векторные представления классификатора на обоих доменах были близки.

2 Теоретическая часть

2.1 Постановка задачи

В качестве функции преобразования будем брать нейронную сеть с параметрами θ_f.

Определение 3 *Оптимальной функцией преобразования домена $\mathcal{D}_1$ в домен $\mathcal{D}_2$ относительно функции сходства s назовем функцию $f_s(\cdot, \hat{\theta}_f)_s$:*

$$\hat{\theta}_f = \arg\max_{\theta_f} s\Big(\mathcal{D}_2, p(f(x, \theta_f), x \sim \mathcal{D}_1)\Big) \tag{1}$$

Для краткости будем обозначать $\hat{f}_s(\cdot) = f_s(\cdot, \hat{\theta}_f)_s$.

2.1.1 Постановка задачи для функции предложенной Адуенко А. А.

В работе [6] Адуенко А. А. ввел функцию сходства:

Определение 4 *Назовем функцией сходства s_0 пары распределений $g_1 : \mathbb{R}^n \to \mathbb{R}^+, g_2 : \mathbb{R}^n \to \mathbb{R}^+$, определенных на одном пространстве, функцию вида*

$$s_0(g_1, g_2) = \frac{\int g_1(\mathbf{x})g_2(\mathbf{x})d\mathbf{x}}{\max_{\mathbf{b}\in\mathbb{R}^n} \int g_1(\mathbf{z})g_2(\mathbf{z}-\mathbf{b})d\mathbf{z}}$$

Перед началом решения задачи и поиском параметры функции преобразования необходимо проверить, что функция Адуенко является критерием совпадения распределений, т.е. верно ли, что:

$$s_0(g_1, g_2) = 1 \overset{?}{\Leftrightarrow} g_1 \equiv g_2$$

Либо в для последовательности распределений $\{g_2^k\}_{k=1}^{\infty}$:

$$s_0(g_1, g_2^k) \xrightarrow{k\to\infty} 1 \overset{?}{\Leftrightarrow} g_2^k \xrightarrow{k\to\infty} g_1. \tag{2}$$

Теорема 1 *Пусть дана пара распределений $g_1(\mathbf{x}) : \mathbb{R}^{n_1} \to \mathbb{R}^+$, $g_2(\mathbf{x}) : \mathbb{R}^{n_2} \to \mathbb{R}^+$, где $n_1, n_2 > 0$. Тогда для некоторой последовательности параметрических линейных преобразований $\{f_\theta^k\}_{k=1}^{\infty}$, такой что $\|g_1 - f_\theta^k \circ g_2\| \to 0$, верно:*

$$s_0(g_1, f_\theta^k \circ g_2) \to 1$$

Лемма 2 *Пусть дано распределение $g_1(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^+$ и последовательность распределений $g_2^k(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}^+$. Тогда, если $\|g_1 - g_2^k\| \to 0$, верно:*

$$s_0(g_1, f_\theta^k \circ g_2) \to 1$$

Доказательство.

Зададим $g_\Delta^k = g_2^k - g_1$, тогда из $\|g_\Delta^k\| \to 0$ следует, что $\int g_\Delta^k \to 0$.

$$\begin{aligned} s_0(g_1, g_1 + g_\Delta^k) = \frac{\int g_1(\mathbf{x})(g_1 + g_\Delta^k)(\mathbf{x})d\mathbf{x}}{\max_{\mathbf{b}} \int g_1(\mathbf{z})(g_1 + g_\Delta^k)(\mathbf{z}+\mathbf{b})d\mathbf{z}} = \\ \frac{\int g_1^2(\mathbf{x})d\mathbf{x} + \int g_1(\mathbf{x})g_\Delta^k(\mathbf{x})d\mathbf{x}}{\max_{\mathbf{b}} \left[\int g_1(\mathbf{z})g_1(\mathbf{z}+\mathbf{b})d\mathbf{z} + \int g_1(\mathbf{z})g_\Delta^k(\mathbf{z}+\mathbf{b})d\mathbf{z} \right]} \geq \\ \frac{\int g_1^2(\mathbf{x})d\mathbf{x} + \int g_1(\mathbf{x})g_\Delta^k(\mathbf{x})d\mathbf{x}}{\max_{\mathbf{b}} \int g_1(\mathbf{z})g_1(\mathbf{z}+\mathbf{b})d\mathbf{z} + \max_{\mathbf{b}'} \int g_1(\mathbf{z})g_\Delta^k(\mathbf{z}+\mathbf{b}')d\mathbf{z}} \end{aligned} \tag{3}$$

Рассмотрим знаменатель полученного выражения. Из неравенства Коши-Буняковского:

$$\int g_1(\mathbf{z})g_1(\mathbf{z}-\mathbf{b})d\mathbf{z} \leq \sqrt{\int g_1^2(\mathbf{z})d\mathbf{z}}\sqrt{\int g_1^2(\mathbf{x}-\mathbf{b})d\mathbf{x}} = \int g_1^2(\mathbf{x})d\mathbf{x},$$

Причем при $\mathbf{b}=\mathbf{0}$ неравенство обращается в равенство. Подставляя данное выражение в (3), получаем:

$$s_0(g_1, g_1+g_\Delta^k) \geq \frac{\int g_1^2(\mathbf{x})d\mathbf{x} + \int g_1(\mathbf{x})g_\Delta^k(\mathbf{x})d\mathbf{x}}{\int g_1^2(\mathbf{z})d\mathbf{z} + \max_{\mathbf{b}} \int g_1(\mathbf{z})g_\Delta^k(\mathbf{z}+\mathbf{b})d\mathbf{z}} \to 1, \text{ при } k\to\infty$$

Откуда следует доказываемое утверждение.

■

Отметим, что теорема 1 является частным случаем леммы 2.

Теорема 3 *Пусть дана пара распределений* $g_1(\mathbf{x}) : \mathbb{R}^{n_1} \to \mathbb{R}^+$, $g_2(\mathbf{x}) : \mathbb{R}^{n_2} \to \mathbb{R}^+$, *где* $n_1, n_2 > 0$. *Тогда существует некоторая последовательность параметрических линейных преобразований* $\{f_\theta^k\}_{k=1}^\infty$, *такая что* $s_0(g_1, f_\theta^k \circ g_2) \to 1$, *и для нее не верно:*

$$\|g_1 - f_\theta^k \circ g_2\| \to 0$$

Лемма 4 *Для линейного преобразование* $f_\theta^k(x)$, *которое имеет вид* $x/k+b$, *где* $k \in \mathbb{N}_+, \mathbf{b} = \arg\max_{\mathbf{x}} g_2(\mathbf{x}),\ \mathbf{b} \in \mathbb{R}^{n_2}$ *Выполняются следующие свойства:*

$$\forall a > 0, f_\theta^k \circ g_2(\cdot)|_A \to U(A), \text{ где } A = \{\mathbf{x} : \|\mathbf{x}\| \leq a\} \tag{4}$$

$$\exists 0 \leq B < \infty,\ \exists k_0 : \forall k \geq k_0 \sup_{\{\mathbf{x}:\|\mathbf{x}\|\geq B\}} f_\theta^k \circ g_2(\mathbf{x}) \leq \sup_{\{\mathbf{x}:\|\mathbf{x}\|\leq B\}} f_\theta^k \circ g_2(\mathbf{x}), \tag{5}$$

где в свойстве (4) $f_\theta^k \circ g_2(\cdot)|_A$ *есть сужение* $f_\theta^k \circ g_2(\cdot)$ *на множество* A, *то есть*

$$g_2(\cdot)|_A(\mathbf{x}) = \begin{cases} 0, \text{ если } \mathbf{x} \notin A \\ \frac{g_2(\mathbf{x})}{\int_A g_2(\mathbf{z})d\mathbf{z}}, \mathbf{x} \in A \end{cases}$$

Сходимость в свойстве (4) *понимается равномерная, то есть*

$$g_2(\cdot) \to U(A) \Leftrightarrow \sup_{\mathbf{x}\in A} \left| g_2(\mathbf{x}) - 1/|A| \right| \to 0, \text{ где } k \to \infty.$$

Доказательство.

С учетом того, что интеграл функции распределения по $\mathbb{R}^{n_2}$ должен равняться 1, получаем, что последовательность распределений имеет вид:

$$f_\theta^k \circ g_2(\mathbf{x}) = g_2\Big(\frac{\mathbf{x}}{k} + \mathbf{b}\Big)\Big/k$$

Проверим выполнимость свойства (4). Так как $\mathbf{b}$ является точкой максимума неотрицательной и ненулевой функции, следовательно $g_2(\mathbf{b}) > 0$. Зафиксируем произвольные $\delta \in (0, g_2(\mathbf{b})), a > 0$

Тогда существует w_δ такая, что для всех $\mathbf{x}$ лежащих внутри окрестности нуля, т.е. $\|\mathbf{x}\| < w_\delta$ верно $g_2(\mathbf{b}) - \delta \le g_2(\mathbf{x} + \mathbf{b}) \le g_2(\mathbf{x})$. Отсюда, для $\forall k \ge \lceil \frac{a}{w_\delta} \rceil$ и $\forall \mathbf{x} \in A$ верно:

$$\frac{1}{k}\Big(g_2(\mathbf{b}) - \delta\Big) \le \frac{1}{k} g_2\Big(\frac{\mathbf{x}}{k} + \mathbf{b}\Big) \le \frac{1}{k} g_2(\mathbf{b}) \tag{6}$$

Из чего верно:

$$\frac{1}{k}\Big(g_2(\mathbf{b}) - \delta\Big)|A| \le \int_A \frac{1}{k} g_2\Big(\frac{\mathbf{x}}{k} + \mathbf{b}\Big) \le \frac{1}{k} g_2(\mathbf{b})|A| \tag{7}$$

Далее объединяя (6) и (7) получаем:

$$\frac{\frac{1}{k}\Big(g_2(\mathbf{b}) - \delta\Big)}{\frac{1}{k} g_2(\mathbf{b})|A|} \le \frac{1}{k} g_2\Big(\frac{\mathbf{x}}{k} + \mathbf{b}\Big)\Big|_A = \frac{\frac{1}{k} g_2\Big(\frac{\mathbf{x}}{k} + \mathbf{b}\Big)}{\int_A \frac{1}{k} g_2\Big(\frac{\mathbf{x}}{k} + \mathbf{b}\Big)} \le \frac{\frac{1}{k} g_2(\mathbf{b})}{\frac{1}{k}\Big(g_2(\mathbf{b}) - \delta\Big)|A|}$$

Если упростить и вычесть $1/|A|$ из каждой части получим:

$$\frac{-\delta}{g_2(\mathbf{b})|A|} \le \frac{1}{k} g_2\Big(\frac{\mathbf{x}}{k} + \mathbf{b}\Big)\Big|_A - \frac{1}{|A|} \le \frac{\delta}{(g_2(\mathbf{b}) - \delta)|A|} \tag{8}$$

Тогда в силу произвольности δ получаем равномерную сходимость сужения элементов последовательности распределений на множество A к $U(A)$, тем самым доказываем выполнения первого свойства (4).

Выполнение второго свойства очевидно, так как 0 - точка максимума и увеличение k сдвигает ее.

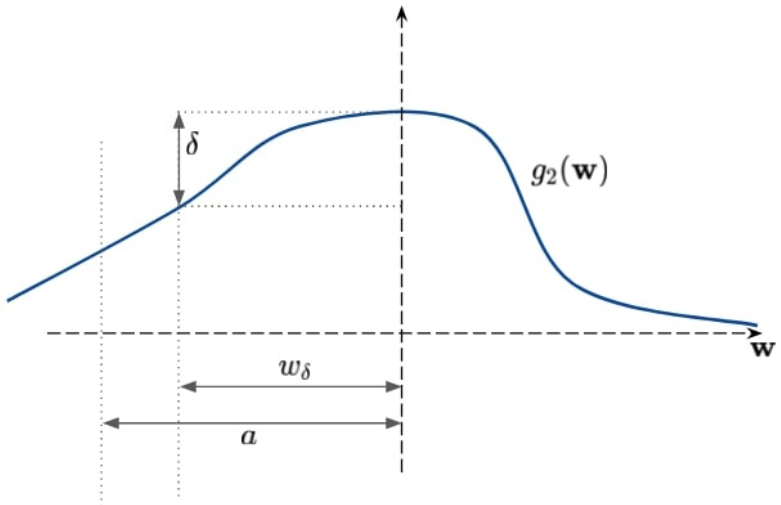


Рис. 3: Поясняющий рисунок к доказательству леммы. При увеличении k получается, что распределение "растягивается"от нуля, а окрестность, в которой значение функции отличается только на максимального значение, постепенно выходит за A.

■

Доказательство. Теоремы 3.

Для доказательства теоремы построим последовательность преобразований описанную в Лемме 4. Обозначим $g_2\left(\frac{\mathbf{x}}{k}+\mathbf{b}\right)$ как $g_2^k(\mathbf{x})$. Таким образом требуется показать, что

$$\frac{\int g_1(\mathbf{x})g_2^k(\mathbf{x})d\mathbf{x}}{\max_{\mathbf{b}}\int g_1(\mathbf{z})g_2^k(\mathbf{z}-\mathbf{b})d\mathbf{z}} \to 1 \text{ при } k\to\infty$$

Обозначим $Q_a = \{\mathbf{x} : \|\mathbf{x}\| \geq a\}, R_a = \{\mathbf{x} : \|\mathbf{x}\| \leq a\}$. Из свойства (5) в лемме 4 имеем:

$$\exists 0 \leq B < \infty,\ \exists k_0 : \forall k \geq k_0\ \sup_{Q_B} g_2^k(\mathbf{x}) \leq \sup_{R_B} g_2^k(\mathbf{x}).$$

Зафиксируем произвольное $\epsilon > 0$. Определим B_ϵ так, что

$$\int_{\{\mathbf{x}:\|\mathbf{x}\|\geq B_\epsilon\}} g_1(\mathbf{z})d\mathbf{z} < \epsilon.$$

Определим $\tilde{B} = \max(B, B_\epsilon)$. Зафиксируем также $\delta > 0$. Из свойства (4) в лемме 4 имеем:

$$\exists k_\delta : \forall k \geq k_\delta\ \frac{\sup_{R_{\tilde{B}}} g_2^k(\mathbf{x})}{\inf_{R_{\tilde{B}}} g_2^k(\mathbf{x})} \leq 1+\delta.$$

Определим $\tilde{k} = \max(k_\delta, k_0)$. Тогда для $k \geq \tilde{k}$ имеем

$$\int g_1(\mathbf{x}) g_2^k(\mathbf{x}) d\mathbf{x} \geq \int_{\{\mathbf{x}:\|\mathbf{x}\|\leq B\}} g_1(\mathbf{x}) g_2^k(\mathbf{x}) d\mathbf{x} \geq$$
$$\inf_{\{\mathbf{x}:\|\mathbf{x}\|\leq B\}} g_2^k(\mathbf{x}) \int_{\{\mathbf{z}:\|\mathbf{z}\|\leq B\}} g_1(\mathbf{z}) d\mathbf{z} \geq (1-\epsilon) \inf_{R_B} g_2^k(\mathbf{x}) \geq$$
$$(1-\epsilon)(1-\delta) \sup_{R_B} g_2^k(\mathbf{x}). \quad (9)$$

Аналогично для знаменателя выражения для s_0 с учетом свойства (5)

$$\forall \mathbf{b} \int g_1(\mathbf{z}) g_2^k(\mathbf{z}-\mathbf{b}) d\mathbf{z} \leq \sup_{\mathbf{z}} g_2^k(\mathbf{z}) = \sup_{R_B} g_2^k(\mathbf{z}). \quad (10)$$

Тогда из (9) и (10) получаем:

$$\forall k \geq \tilde{k} : s_0(g_1, g_2^k) \geq (1-\epsilon)(1-\delta).$$

С учетом произвольности выбора ϵ, δ получаем требуемое. Так как такое линейное преобразование "растягивает"распределение, т.е.

$$\forall \mathbf{x}\ g_2^k(\mathbf{x}) \to 0, \text{ при } k \to \infty$$

Что эквивалентно тому, что:

$$\|g_2^k\| \to 0, \text{ при } k \to \infty$$

Поэтому из условия $\|g_1 - g_2^k\| \to 0$ следует, что $g_1 \equiv 0$, что в общем случае неверно. Тем самым доказываем теорему.

■

Из теоремы 1 и теоремы 3 следует, что функция Адуенко А. А. является необходимой, но недостаточной в качестве меры совпадения распределений. Т.е. не выполняется (2). Из-за чего данная функция не используется для нахождения параметров преобразования.

2.1.2 Постановка задачи для дивергенция Кульбака-Лейблера

Теперь рассмотрим дивергенцию Кульбака-Лейблера в качестве функции сходства. Для двух функций распределения g_1, g_2:

$$D_{KL}(g_1, g_2) = \int g_1(\mathbf{x}) \log \frac{g_1(\mathbf{x})}{g_2(\mathbf{x})} d\mathbf{x} = \mathbb{E}_{\mathbf{x}\sim g_1} \log g_1(\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim g_1} \log g_2(\mathbf{x}) \tag{11}$$

D_{KL} достигает минимума при совпадении распределений g_1 и g_2. Так же видно, что дивергенция Кульбака-Лейблера асимметрична. В тех случаях, когда $g_1(\mathbf{x})$ близко к нулю, а $g_2(\mathbf{x})$ значительно отличается от нуля, то получается, что g_2 оказывает малое влияние. Это может привести к ошибочным результатам, когда мы просто хотим измерить сходство между двумя одинаково важными распределениями.

Определение 5 *Назовем функцией сходства порожденную метрикой D:*

$$s_D(g_1, g_2) = \exp(-D(g_1, g_2)) \tag{12}$$

Например, функция сходства порожденная дивергенцией Кульбака-Лейблера:

$$s_{KL}(g_1, g_2) = \exp(-D_{KL}(g_1, g_2)) \tag{13}$$

Отметим, что с учетом (13) задачу максимизации (1) можно переписать в следующем виде:

$$\hat{\theta}_f = \arg\min_{\theta_f} D_{KL}(g_1, f \circ g_2)$$

Для решения данной задачи, когда отсутствует явно заданное распределение, а есть только набор значений полученных сэмплированием из него, предлагается применять подход генеративно-состязательных сетей (GAN) описанных в [7]. GAN состоит из двух моделей:

- Дискриминатор D оценивает вероятность получения данной выборки из исходного домена данных. Он работает как критик и оптимизирован для того, чтобы отличать из какого домена берутся данные.

- Генератор, в нашем случае это функция преобразования, переводит целевой домен в пространство признакового описания исходного домена. Он обучен копировать распределение исходного домена, или, другими словами, генератор пытается обмануть дискриминатор.

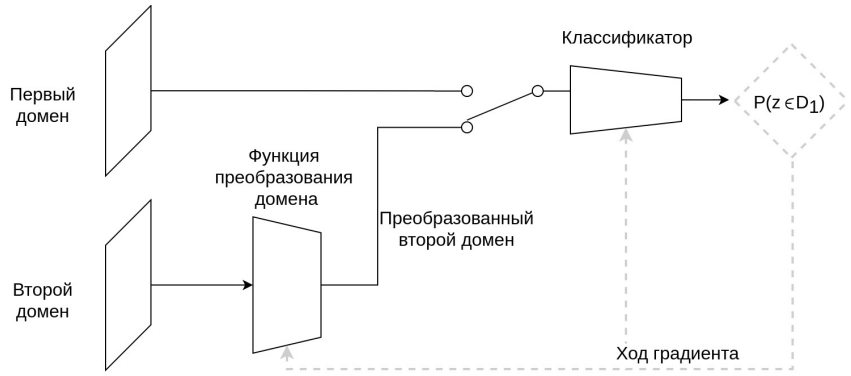


Рис. 4: Предлагаемая архитектура модели для решения задачи нахождения параметров функции преобразования оптимальной относительно дивергенцией Кульбака-Лейблера

С одной стороны, мы хотим, чтобы решения дискриминатора D по данным из первого домена были точны, максимизируя $\mathbb{E}_{\mathbf{x}\sim g_1}[\log D(\mathbf{x})]$. Между тем, учитывая преобразованную выборку второго домена $f(\mathbf{z}), \mathbf{z} \sim g_2$, ожидается, что дискриминатор покажет вероятность $D(f(\mathbf{z}))$, близкую к нулю, максимизируя $\mathbb{E}_{\mathbf{z}\sim g_2}[\log(1 - D(f(\mathbf{z})))]$.

С другой стороны, генератор обучен увеличивать вероятность того, что D даст высокую вероятность для преобразованного целевого домена, таким образом, чтобы минимизировать $\mathbb{E}_{\mathbf{z}\sim g_2}[\log(1 - D(f(\mathbf{z})))]$.

При объединении обоих аспектов вместе получается, что D и f играют в минимаксную задачу, в которой мы должны оптимизировать следующую функцию потерь:

$$L_{KL}(D, f) = \mathbb{E}_{\mathbf{x}\sim g_1}[\log D(\mathbf{z})] + \mathbb{E}_{\mathbf{z}\sim g_2}[\log(1 - D(f(\mathbf{z})))] = \\ \mathbb{E}_{\mathbf{x}\sim g_1}[\log D(\mathbf{x})] + \mathbb{E}_{\mathbf{x}\sim f\circ g_2}[\log(1 - D(\mathbf{x})]$$

где член $\mathbb{E}_{\mathbf{x}\sim g_1}[\log D(\mathbf{x})]$ не влияет на G во время обучения.

Итоговая оптимизационная задача:

$$\min_f \max_D L_{KL}(D, f) \tag{14}$$

2.1.3 Постановка задачи для расстояния Васерштейна

Теперь рассмотрим расстояние Васерштейна в качестве функции сходства. В общем случае расстояние Васерштейна имеет вид:

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

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

Для оценки параметров функции преобразования будет использоваться расстояние Васерштейна-1 или Earth-Mover (EM) distance:

$$W(g_1, g_2) = \inf_{\gamma \in \Gamma(\mu,v)} \mathbb{E}_{(x,y)\sim\gamma} |x - y| \tag{15}$$

W также достигает минимума при совпадении распределений g_1 и g_2.

Из (12) получаем, что функция сходства порожденная расстоянием Васерштейна:

$$s_W(g_1, g_2) = \exp(-W(g_1, g_2)) \tag{16}$$

Отметим, что с учетом (16) задачу максимизации (1) можно переписать в следующем виде:

$$\hat{\theta}_f = \arg\min_{\theta_f} W(g_1, f \circ g_2)$$

Выражение (15) достаточно трудно вычислять, но с учетом, двойственности Канторовича-Рубинштейна [8], которая говорит нам о том, что

$$W(g_1, g_2) = \sup_{\|D\|_L \leq 1} \mathbb{E}_{\mathbf{x}\sim g_1}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{z}\sim g_2}[D(\mathbf{z})], \tag{17}$$

где супремум берется по всеми 1-липшицевым функциям $D : \mathbb{R}^n \to \mathbb{R}$. Причем, если мы заменим $\|D\|_L \leq 1$ на $\|D\|_L \leq K$, т.е. рассмотрим K-липшицевы функции для некоторой константы K, то получим $K \cdot W(g_1, g_2)$. Поэтому, если существует параметризованное семейство из функций $\{D_w\}_{w\in W}$, все из которых являются K-липшицевыми для некоторого K, то задача переписывается к задаче максимизации:

$$\max_{w\in W} \mathbb{E}_{\mathbf{x}\sim g_1}[D(\mathbf{x})] - \mathbb{E}_{\mathbf{z}\sim g_2}[D(\mathbf{z}] \tag{18}$$

Если супремум в (17) достигается для некоторого $w \in W$, то и максимум в (18) будет достигнут, его значение будет известно с точностью до мультипликативной константы. Однако важно, что данная константа не влияет на точку максимум, поэтому значения аргументов, в которых достигаются супремум и максимум будут совпадать.

Таким образом получается оптимизационная задача функции потерь:

$$L_W(f) = \max_{w \in W} \mathbb{E}_{\mathbf{x} \sim g_1}[D_w(\mathbf{x})] - \mathbb{E}_{\mathbf{z} \sim g_2}[D(f(\mathbf{z}))]$$

Для нахождения параметров функции преобразования одного домена в другой оптимальной относительно расстояния Васерштейна решается оптимизационная задача:

$$\min_f L_W(f) \tag{19}$$

3 Результаты экспериментов

Цель эксперимента заключается в том, что мы хотим проверить верность гипотезы сформулированной в введении, и в том, что хотим определить, какая функция сходства позволяет достигать лучшие результаты. Хотим проверить верность утверждения, что полученная функция преобразования домена сохраняет инвариантность на классах и значениях целевых переменных. Для проведения эксперимента берется два различных домена и задача регрессии на первом из них. После этого находится оптимальная функция преобразования целевого домена в исходный и применяются два предлагаемых далее метода оценки качества функции преобразования.

Пусть функция преобразования сохраняет инвариантность на значениях целевых переменных, тогда:

1. среднеквадратическое отклонение предсказываемого значения от правильного ответа на преобразованном целевом домене должно быть схоже с соответствующей величиной на исходном домене.

2. можно статистически проверить гипотезу о равенстве весов в задачи линейной регрессии. Ожидается, что эта гипотезу будет неотвержима.

В работе [6] рассматривается задача различения моделей обученных на непересекающихся множествах. В ней выводится распределение значения функции сходства s_0 между апостериорными распределениями весов для пары линейных моделей в условиях истинности гипотезы H_0 о совпадении весов моделей.

Пусть $\mathbf{X}_1$ и $\mathbf{X}_2$ есть выборки. Пусть тогда целевые переменные:

$$\mathbf{y}_1 = \mathbf{X}_1\mathbf{w}_1 + \epsilon_1, \epsilon_1 \sim \mathcal{N}(0, \sigma_1^2\mathbf{I}), \mathbf{w}_1 \sim p_1(\mathbf{w}_1)$$
$$\mathbf{y}_2 = \mathbf{X}_2\mathbf{w}_2 + \epsilon_2, \epsilon_2 \sim \mathcal{N}(0, \sigma_2^2\mathbf{I}), \mathbf{w}_2 \sim p_2(\mathbf{w}_2)$$

Считаем далее, что $p_1(\mathbf{w}_1)$ и $p_2(\mathbf{w}_2)$ есть нормальные распределения, то есть

$$p_1(\mathbf{w}_1) = \mathcal{N}(\mathbf{w}_1|\mathbf{v}_1, \mathbf{\Sigma}_1^{-1}), p_2(\mathbf{w}_2) = \mathcal{N}(\mathbf{w}_2|\mathbf{v}_2, \mathbf{\Sigma}_2^{-1}).$$

Получаем, что функции совместного правдоподобия имеют вид

$$p(\mathbf{y}_k, \mathbf{w}_k|\mathbf{X}_k) = p(\mathbf{y}_k|\mathbf{X}_k, \mathbf{w}_k)p(\mathbf{w}_k) = \mathcal{N}(\mathbf{y}_k - \mathbf{X}_k\mathbf{w}_k|\mathbf{0}, \sigma_k^2\mathbf{I})\mathcal{N}(\mathbf{w}_k|\mathbf{v}_k, \mathbf{\Sigma}_k^{-1}),\ k = 1, 2.$$

Пользуясь формулой Байеса, получаем для апостериорного распределения параметров $\mathbf{w}_1$ и $\mathbf{w}_2$

$$p(\mathbf{w}_k|\mathbf{X}_k, \mathbf{y}_k) = \frac{p(\mathbf{y}_k, \mathbf{w}_k|\mathbf{X}_k)}{p(\mathbf{y}_k|\mathbf{X}_k)} \propto p(\mathbf{y}_k, \mathbf{w}_k|\mathbf{X}_k),\ k = 1, 2,$$

откуда $p(\mathbf{w}_k|\mathbf{X}_k, \mathbf{y}_k) = \mathcal{N}(\mathbf{w}_k|\hat{\mathbf{w}}_k, \tilde{\mathbf{\Sigma}}_k^{-1})$, где

$$\hat{\mathbf{w}}_k = \left(\mathbf{\Sigma}_k + \frac{1}{\sigma_k^2}\mathbf{X}_k^T\mathbf{X}_k\right)^{-1}\left(\mathbf{\Sigma}_k\mathbf{v}_k + \frac{1}{\sigma_k^2}\mathbf{X}_k^T\mathbf{y}_k\right)$$

есть оценка максимума апостериорной вероятности, а

$$\tilde{\mathbf{\Sigma}}_k = \mathbf{\Sigma}_k + \frac{1}{\sigma_k^2}\mathbf{X}_k^T\mathbf{X}_k,\ k = 1, 2$$

Тогда для s_0 для пары апостериорных распределений имеем:

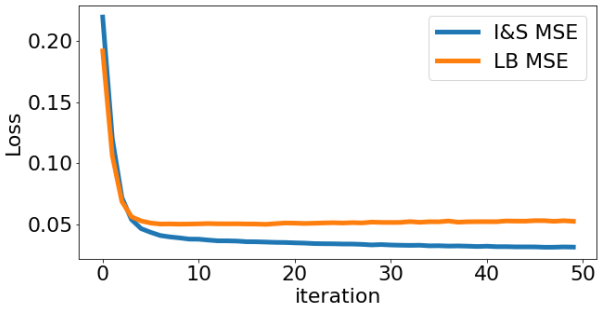
$$-2\log s_0 = (\hat{\mathbf{w}}_2 - \hat{\mathbf{w}}_1)^T\left(\left(\mathbf{\Sigma}_1 + \frac{1}{\sigma_1^2}\mathbf{X}_1^T\mathbf{X}_1\right)^{-1} + \left(\mathbf{\Sigma}_2 + \frac{1}{\sigma_2^2}\mathbf{X}_2^T\mathbf{X}_2\right)^{-1}\right)^{-1}(\hat{\mathbf{w}}_2 - \hat{\mathbf{w}}_1).$$

И также выводится, что $-2\log s_0 \xrightarrow{d} \chi^2(n)$.

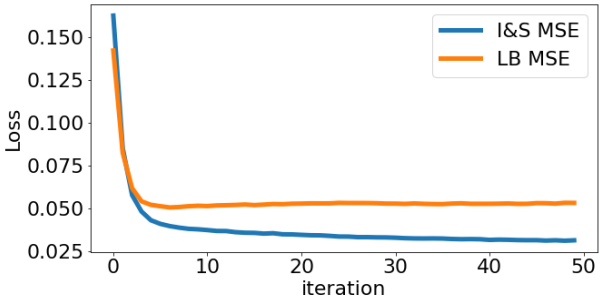
3.1 Вычислительный эксперимент для отзывов с сайта Amazon

Вычислительный эксперимент проводился на данных об отзывах о товарах с сайта Amazon. Домену соответствовала категория, и строилось векторное представление для каждого отзыва. Также каждому отзыву соответствует оценка, которую поставил пользователь. В качестве различных доменов были взяты категории "Industrial and Scientific"(I&S) и "Luxury Beauty"(LB)

Находились две функции преобразования домена LB в I&S по дивергенции Кульбака-Лейблера и по метрики Васерштейна. Вследствие двух теорем выше функция сходства Адуенко в вычислительном эксперименте не участвовала. Для каждой из построенных функций преобразования обучалась модель регрессии оценки пользователя.



(a)



(b)

Рис. 5: Среднеквадратическое отклонение предсказываемого значения от правильного ответа для (a) дивергенции Кульбака-Лейблера и (b) для расстояния Васерштейна, результаты весьма схожи.

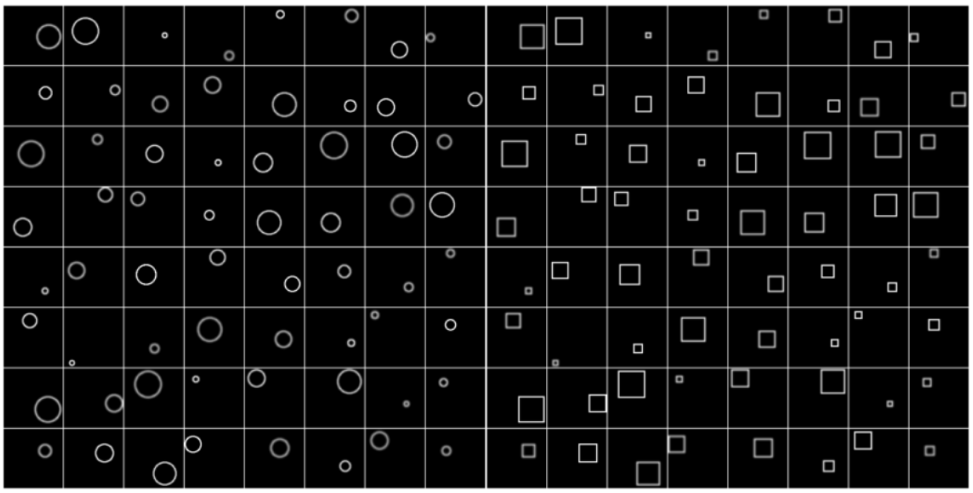
	Дивергенция Кульбака-Лейблера	Расстояние Васерштейна
p-value	0.1311	0.3412

Таблица 1: Статистическая проверка гипотезы о равенстве весов в задаче линейной регрессии на домене I&S и преобразованном домене LB для различных функций преобразования оптимальных относительно дивергенции Кульбака-Лейблера и расстояния Васерштейна соответственно.

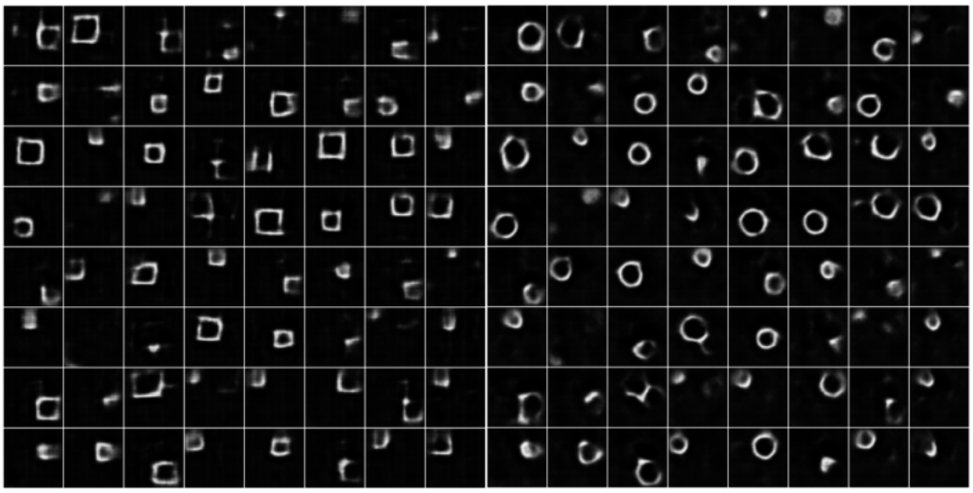
3.2 Вычислительный эксперимент для бинаризованных изображений фигур

Данный вычислительный эксперимент проводился на бинаризованных изображениях. Были взяты два домена — изображения с различными фигурами. Первому соответствовали квадраты, второму круги. Каждой фигуре можно задать в соответствие координаты центра фигуры и ее радиус.

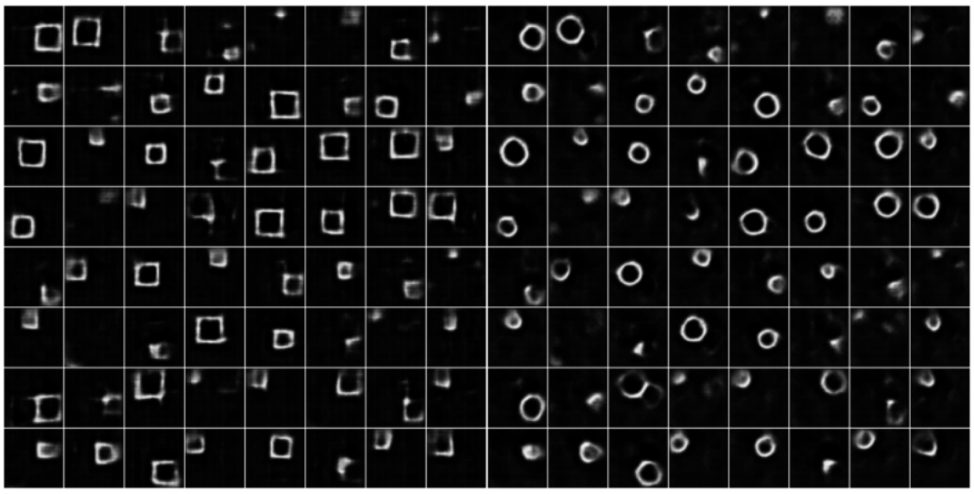
Как и раньше, находились функции преобразования доменов по дивергенции Кульбака-Лейблера и по метрики Васерштейна. Для каждой из построенных функций преобразования обучалась модель регрессии радиуса фигуры.



(a)

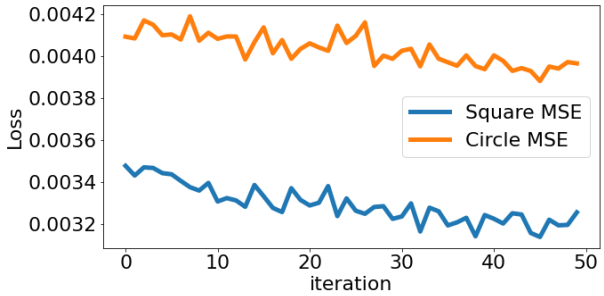


(b)

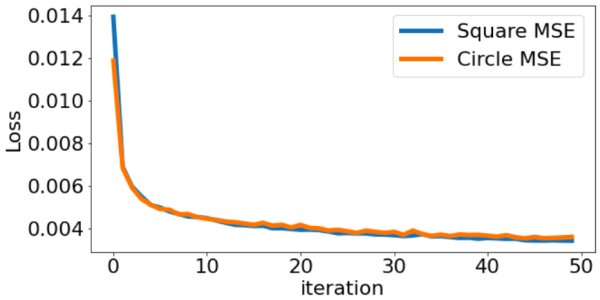


(c)

Рис. 6: (a) оригинальные изображения. Изображения преобразованные с помощью (b) дивергенции Кульбака-Лейблера или (c) расстояния Васерштейна. Важно отметить, что при применении функции преобразования фигура меняла форму, но не положение или размер.



(a)



(b)

Рис. 7: Среднеквадратическое отклонение предсказываемого значения от правильного ответа для (a) дивергенции Кульбака-Лейблера и (b) для расстояния Васерштейна.

	Дивергенция Кульбака-Лейблера	Расстояние Васерштейна
p-value квадрат в круг	0.1682	0.2994
p-value круг в квадрат	0.1778	0.2620

Таблица 2: Статистическая проверка гипотезы о равенстве весов в задаче линейной регрессии для бинаризованных картинок фигур.

4 Заключение

В работе предложен метод решения задачи доменной адаптации через оценку параметров функции преобразования домена. Теоретически доказано, что функция сходства предложенная Адуенко А. А. не является достаточным условием совпадения распределений. По этой причине она не может использоваться для оценки параметров функции преобразования. Также описан алгоритм нахождения функции преобразования, оптимальной относительно дивергенции Кульбака-Лейблера и расстояния Васерштейна.

В ходе вычислительного эксперимента подтверждена гипотеза о том, что функция преобразования сохраняет инвариантность на классах и значениях целевых переменных. Также было установлено, что расстояние Васерштейна позволяет строить более качественную функцию преобразования. Хотя достигаемое среднеквадратическое отклонение ведет себя одинаково, но второй тест, который заключается в статистической проверке равенства весов моделей линейной регрессии показывает преимущество использования расстояния Васерштейна.

Список литературы

[1] *Yaroslav Ganin and Evgeniya Ustinova and Hana Ajakan and Pascal Germain and Hugo Larochelle and François Laviolette and Mario Marchand and Victor Lempitsky*. Domain-Adversarial Training of Neural Networks, 2016

[2] *Eric Tzeng and Judy Hoffman and Kate Saenko and Trevor Darrell*. Adversarial Discriminative Domain Adaptation, 2017

[3] *Issam Laradji and Reza Babanezhad*. M-ADDA: Unsupervised Domain Adaptation with Deep Metric Learning, 2018

[4] *Kuniaki Saito and Kohei Watanabe and Yoshitaka Ushiku and Tatsuya Harada*. Maximum Classifier Discrepancy for Unsupervised Domain Adaptation, 2018

[5] *Rui Shu and Hung H. Bui and Hirokazu Narui and Stefano Ermon*. A DIRT-T Approach to Unsupervised Domain Adaptation, 2018

[6] *А.А. Адуенко*. Выбор мультимоделей в задачах классификации, 2017

[7] *Ian J. Goodfellow and Jean Pouget-Abadie and Mehdi Mirza and Bing Xu and David Warde-Farley and Sherjil Ozair and Aaron Courville and Yoshua Bengio*. Generative Adversarial Networks, 2014

[8] *C´edric Villani*. Optimal Transport: Old and New. Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.