

МОСКОВСКИЙ ФИЗИКО-ТЕХНИЧЕСКИЙ
ИНСТИТУТ (ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ)

ФАКУЛЬТЕТ УПРАВЛЕНИЯ И ПРИКЛАДНОЙ
МАТЕМАТИКИ

Сравнительный анализ линейных классификаторов

Выпускная квалификационная работа на степень бакалавра
студента 376 группы Змеева Максима Владимировича

НАУЧНЫЙ РУКОВОДИТЕЛЬ: ПРОФЕССОР РЯЗАНОВ
В.В.

Содержание

1	Введение	2
2	Постановка задачи	2
2.1	Для двух классов	2
2.2	Для любого числа классов	2
3	Предлагаемые методы решения	3
3.1	Для двух классов	3
3.1.1	Релаксационный метод	3
3.1.2	Сведение к линейному программированию	4
3.1.3	Аппроксимация гладкой функцией	5
3.2	Для любого числа классов	6
3.2.1	Метод опорных векторов (SVM)	6
3.2.2	Минимизация кросс-энтропии (Softmax)	7
4	Реализация методов	8
4.1	Реализация релаксационного метода	8
4.2	Реализация метода сведения к линейному программированию	10
4.3	Реализация аппроксимации гладкой функцией	12
4.4	Реализация метода опорных векторов (SVM)	13
4.5	Реализация метода минимизации кросс-энтропии (Softmax)	14
5	Сравнение на модельных данных	15
5.1	Первая выборка	15
5.2	Вторая выборка	16
5.3	Третья выборка	17
5.4	Четвёртая выборка	18
6	Выводы	19
6.1	Реализованы:	19
6.2	Исследованы:	19
6.3	Проведено сравнение:	19
7	Используемая литература	21

1 Введение

В наше время машинное обучение завоевало строго положительную репутацию. В данных часто есть закономерности, которые не удаётся отследить невооруженным глазом. Для этого существует много математических моделей, которые находят зависимости между данными и позволяют делать предсказания. Предсказательные модели можно использовать практически в любой сфере. Самые простые предсказания строятся на основе линейных моделей, которые отличаются от других высокой скоростью обучения, но меньшей предсказательной силой. В данной работе рассмотрены основные известные виды линейных классификаторов, к ним добавлены два менее известных подхода (релаксационный метод и метод сведения к линейному программированию). Оба подхода решают задачу поиска максимальной совместной подсистемы системы линейных неравенств. Представляет интерес сравнить их между собой и с уже проявившими себя методами. Каждый классификатор реализован в Python3 с максимальным использованием библиотеки `numpy` для ускорения вычислений. Все классификаторы сравнивались как на модельных данных, так и на настоящих данных, взятых из известных датасетов. Сравнение производилось по доле правильных ответов (классы всегда достаточно сбалансированы поэтому эта метрика вполне подходит). Также сравнивались времена работы классификаторов. Было выяснено, что сведение к линейному программированию имеет то преимущество, что позволяет сказать, можно ли с самого начала построить разделяющую гиперплоскость между объектами разных классов. Релаксационный метод проявил себя как самый быстро обучающийся, а три оставшиеся в очередной раз подтвердили свою высокую скорость и хорошую для линейной классификации точность предсказаний.

2 Постановка задачи

2.1 Для двух классов

Есть N элементов, каждый из которых принадлежит одному из двух классов. Каждый объект представляет из себя вектор размерности D . Для N_{train} элементов их классы известны:

$$y_k = \begin{cases} 1, & \text{если объект } \mathbf{x}_k \text{ принадлежит классу } 1 \\ 0, & \text{если объект } \mathbf{x}_k \text{ принадлежит классу } 0 \end{cases}.$$

Для остальных $N_{test} = N - N_{train}$ элементов (тестовой выборки) их классы нужно предсказать для тестовых N_{test} объектов.

2.2 Для любого числа классов

Имеется N объектов, каждый из которых принадлежит одному из M классов. Каждый объект имеет D числовых признаков. Для N_{train} объек-

тов(обучающей выборки) их классы известны.

$$y_k = \begin{cases} 0, & \text{если объект } \mathbf{x}_k \text{ принадлежит классу 0} \\ 1, & \text{если объект } \mathbf{x}_k \text{ принадлежит классу 1} \\ \dots & \\ M, & \text{если объект } \mathbf{x}_k \text{ принадлежит классу M} \end{cases}.$$

Для остальных $N_{test} = N - N_{train}$ элементов(тестовой выборки) их классы нужно предсказать для тестовых N_{test} объектов.

3 Предлагаемые методы решения

3.1 Для двух классов

3.1.1 Релаксационный метод

Предлагается пытаться отделять объекты классов друг от друга гиперплоскостью в D-мерном пространстве. Гиперплоскость задаётся (D+1)-мерным вектором \mathbf{w} . Объект \mathbf{x}_k считается объектом класса 1, если $w_0 + \sum_{j=1}^D x_{kj} \cdot w_j > 0$ и объектом класса 0, если $w_0 + \sum_{j=1}^D x_{kj} \cdot w_j < 0$ (вероятность, что непрерывно распределенная случайная величина будет равна ровно нулю в такой линейной комбинации, равна нулю, поэтому качество не будет ухудшено вне зависимости от того, будет ли присваиваться нулевой или первый класс объектам, на которых оба неравенства не выполняются). Все векторы \mathbf{x}_k для удобства преобразуются в (D+1)-мерные вектора вида $(1, x_{k1}, x_{k2}, \dots, x_{kD})$, чтобы можно было более кратко писать $\mathbf{x}_k^T \mathbf{w}$. Для нахождения вектора \mathbf{w} минимизируется функционал:

$$L = \frac{1}{N_{train}} \sum_{k=1}^{N_{train}} ([\mathbf{x}_k^T \mathbf{w} > 0] \cdot (1 - y_k) + [\mathbf{x}_k^T \mathbf{w} < 0] \cdot y_k) \quad (1)$$

То есть долю ошибок при классификации обучающей выборки.

Составляется система неравенств:

$$\left\{ \mathbf{x}_k^T \mathbf{w} \cdot (-1)^{y_k} \geq 0, \quad k = \overline{1, N_{train}} \right. \quad (2)$$

Теперь получается, что нужно найти такой вектор \mathbf{w} , чтобы максимальное число неравенств из (2) выполнялось. Задача сводится к поиску максимальной совместной подсистемы системы неравенств.

Далее под \mathbf{x}_k подразумевается $\mathbf{x}_k \cdot (-1)^{y_k}$ как коэффициент одного уравнения системы (2). Допустим, система совместна. Искать её решение предлагается, строя последовательность $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k, \dots$. Введено обозначение $I(\mathbf{w}_k) = \{k | \mathbf{x}_k^T \mathbf{w} < 0\}$. Далее идут разные варианты построения последовательности $\{\mathbf{w}_k\}_{k=1}^K$, а точнее выбор направления \mathbf{d}_w в $\mathbf{w}_{k+1} = \mathbf{w}_k + \theta_{k+1} \cdot \mathbf{d}_w$:

•

$$\mathbf{d}_w = \mathbf{x}_q \cdot \frac{\mathbf{x}_q^T \mathbf{w}_k}{\|\mathbf{x}_q\|^2}, \quad \text{где } q = \arg \max_{j \in I(\mathbf{w}_k)} |\mathbf{x}_j^T \mathbf{w}_k| \quad (3)$$

•

$$\mathbf{d}_w = \mathbf{z} \cdot \frac{\sum_{j \in I(\mathbf{w}_k)} (\mathbf{x}_j^T \mathbf{w}_k)^2}{\|\mathbf{z}\|^2}, \text{ где } \mathbf{z} = \sum_{j \in I(\mathbf{w}_k)} \mathbf{x}_j \cdot |\mathbf{x}_j^T \mathbf{w}_k| \quad (4)$$

•

$$\mathbf{d}_w = \mathbf{z} \cdot \frac{\sum_{j \in I(\mathbf{w}_k)} \lambda_j \cdot |\mathbf{x}_j^T \mathbf{w}_k|}{\|\mathbf{z}\|^2}, \text{ где } \mathbf{z} = \sum_{j \in I(\mathbf{w}_k)} \mathbf{x}_j \cdot \lambda_j; \lambda_j \geq 0 \quad (5)$$

λ_j находятся как решения задачи максимизации числа неравенств, которые сменяют свой знак с отрицательного на положительный.

Для таких случаев при $0 \leq \theta \leq 2$ и, если система совместна, доказано[1]:

$$\rho(\mathbf{w}_{k+1}, Q) \leq \sigma \cdot \rho(\mathbf{w}_k, Q), 0 < \sigma < 1, \text{ где } Q - \text{множество решений}$$

Вернёмся к случаю, когда система не обязательно совместна. Ясно, что тогда не будет найдено решение полной системы, потому что не абсолютно любые объекты отделяются гиперплоскостью, также всегда можно встретить выбросы. То есть нужно отбрасывать некоторые из неравенств. Предлагается это делать через каждые несколько итераций, отбрасывая те неравенства, которые чаще всего не выполняются. Таким образом, наступит момент, когда система станет совместна, и решение будет найдено.

Для увеличения скорости сходимости рекомендуется решать систему:

$$\left\{ \mathbf{x}_k^T \mathbf{w} \cdot (-1)^{y_k} \geq \varepsilon, \quad k = \overline{1, N_{train}} \right. \quad (6)$$

Итак, данный классификатор, имеет 3 гиперпараметра:

- ε - поправка к неравенствам;
- k - число шагов, после которого отбрасываются неравенства, которые чаще всего не выполнялись;
- θ - скорость обучения.

Данное число, конечно, можно расширить за счёт того, что скорость обучения можно менять по мере приближения к множеству решений, одновременно с этим можно менять и число шагов, которое делается до отбрасывания неравенств по мере обучения.

3.1.2 Сведение к линейному программированию

Задача снова сводится к поиску максимальной совместной подсистемы системы линейных неравенств. Снова те неравенства, чьи объекты принадлежат к классу 0, умножаются на -1. Вводится дополнительная переменная ε , которую нужно максимизировать в следующей задаче:

$$\left\{ \begin{array}{l} \varepsilon \rightarrow \max \\ \mathbf{x}_k^T \mathbf{w} \cdot (-1)^{y_k} > \varepsilon, \quad k = \overline{1, N_{train}} \end{array} \right. \quad (7)$$

Если система несовместна, то очевидно, что решение будет $\varepsilon = \epsilon \leq 0$. Далее предлагается отбрасывать неравенства. Отбрасывается то, для которого достигается максимум величины

$$\frac{\frac{1}{|I_0|} \cdot \left(\sum_{\substack{i \in I_0 \\ i \neq j}} w_i^T \right) w_j}{\left\| \sum_{\substack{i \in I_0 \\ i \neq j}} w_i \right\| \cdot \|w_j\|}$$

Такая процедура выполняется, пока ε не станет больше нуля.

Итак, получили классификатор с 2 гиперпараметрами:

- k - число шагов, после которого мы отбрасываем неравенства, самые отклоняющиеся от остальных невыполненных;
- $part_to_exclude$ - доля неравенств, которую мы будем исключать из системы.

3.1.3 Аппроксимация гладкой функцией

Теперь хочется ускорить процесс обучения. Для этого нужно сделать функцию потерь L непрерывно дифференцируемой. Хорошо подходит сигмоида:

$$\sigma(\mathbf{x}_k) = \frac{1}{1 + \exp(-\mathbf{x}_k^T \mathbf{w} \cdot (-1)^{y_k})} \quad (8)$$

Если объект k классифицировали как объект класса 1, то значение $\mathbf{x}_k^T \mathbf{w} > 0$. Если решение принято верно (true positive), то степень экспоненты будет положительной, значит вся дробь будет стремиться к нулю при увеличении значения $\mathbf{x}_k^T \mathbf{w}$. Если решение ошибочное (false positive), то степень экспоненты будет отрицательной, а значит вся дробь будет стремиться к единице при отдалении $\mathbf{x}_k^T \mathbf{w}$ от единицы вправо. Аналогично, если $\mathbf{x}_k^T \mathbf{w} < 0$ и выбор правильный (true negative), то вся дробь будет тем ближе к нулю, чем левее от нуля значение $\mathbf{x}_k^T \mathbf{w}$, если выбор неправильный (false negative), то вся дробь будет тем ближе к единице, чем больше значение $\mathbf{x}_k^T \mathbf{w}$. То есть при правильных ответах значение (8) близко к единице, при неправильных - к нулю.

Функция потерь - это сумма таких сигмoids для каждого объекта из обучающей выборки:

$$L = \sum_{k=1}^{N_{train}} \sigma(\mathbf{x}_k) \quad (9)$$

Подсчёт градиента сигмоиды:

$$\begin{aligned} \frac{\partial \sigma(\mathbf{x}_k)}{\partial w_j} &= \\ &= - \frac{1}{(1 + \exp(-\mathbf{x}_k^T \mathbf{w}(-1)^{y_k}))^2} \cdot \exp(-\mathbf{x}_k^T \mathbf{w}(-1)^{y_k}) \cdot (-x_{kj} \cdot (-1)^{y_k}) = \\ &= \frac{1}{1 + \exp(-\mathbf{x}_k^T \mathbf{w}(-1)^{y_k})} \cdot \frac{\exp(-\mathbf{x}_k^T \mathbf{w}(-1)^{y_k})}{1 + \exp(-\mathbf{x}_k^T \mathbf{w}(-1)^{y_k})} \cdot (-1)^{y_k+1} \cdot x_{kj} = \\ &= \sigma(\mathbf{x}_k^T \mathbf{w}) \cdot (1 - \sigma(\mathbf{x}_k^T \mathbf{w})) \cdot x_{kj} \cdot (-1)^{y_k+1}, \text{ при } k = \overline{1, N_{train}}, j = \overline{1, D+1} \end{aligned}$$

Видно, что градиент сигмоиды легко выражается через саму сигмоиду, здесь же нужно добавить производную показателя сигмоиды.

Итак, эту функцию потерь минимизируют итерационным методом, использующим аналитический подсчёт градиента, например, градиентным спуском и тем самым находят решение.

3.2 Для любого числа классов

3.2.1 Метод опорных векторов (SVM)

Теперь имеется более, чем два класса, поэтому для принятия решения недостаточно одной разделяющей гиперплоскости. Для каждого класса заводится свой вектор весов $\mathbf{w}_k, k = \overline{1, M}$. В итоге имеется матрица весов $\mathbf{W} = \|\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_M\|$. Для определения класса объекта \mathbf{x} вычисляется произведение $\mathbf{W}^T \mathbf{x}$ - вектор оценок. Выбирается тот класс, чья оценка самая высокая.

Функция потерь L_i на одном объекте выглядит следующим образом:

$$L_k = \sum_{j \neq y_k} \max(0, \mathbf{x}_k^T \mathbf{w}_j - \mathbf{x}_k^T \mathbf{w}_{y_k} + \Delta) \quad (10)$$

Также есть смысл использовать регуляризацию

$$R(W) = \sum_{i=1}^D \sum_{j=1}^M W_{i,j}^2 \quad (11)$$

Итоговая функция потерь может быть записана как

$$L = \frac{1}{N} \sum_{i=1}^N L_i + \lambda R(W)$$

Или более полно:

$$L = \frac{1}{N} \sum_{k=1}^N \sum_{j \neq y_k} \max(0, \mathbf{x}_k^T \mathbf{w}_j - \mathbf{x}_k^T \mathbf{w}_{y_k} + \Delta) + \lambda \sum_i \sum_j W_{i,j}^2 \quad (12)$$

Остаётся найти минимум по весам данной функции. Оптимизация проводится градиентным спуском. Производная функции потерь по весу верного класса выражается в виде:

$$\nabla_{\mathbf{w}_{y_k}} L_k = - \left(\sum_{j \neq y_k} \mathbf{1}(\mathbf{x}_k^T \mathbf{w}_j - \mathbf{x}_k^T \mathbf{w}_{y_k} + \Delta > 0) \right) \mathbf{x}_k \quad (13)$$

где $\mathbf{1}$ - функция-индикатор, равная единице, если условие верно, и нулю, если условие не верно. По остальным весам ($j \neq y_k$):

$$\nabla_{\mathbf{w}_j} L_k = \mathbf{1}(\mathbf{x}_k^T \mathbf{w}_j - \mathbf{x}_k^T \mathbf{w}_{y_k} + \Delta > 0) \mathbf{x}_k \quad (14)$$

Данный классификатор имеет 2 гиперпараметра:

1. Δ - сдвиг в функции потерь;
2. λ - коэффициент перед весами. Чем он больше, тем сильнее регуляризуются веса.

3.2.2 Минимизация кросс-энтропии (Softmax)

Пусть имеется некая оценочная функция $f_j(\mathbf{x}_k, \mathbf{W})$, в данном случае $f_j(\mathbf{x}_k, \mathbf{W}) = \mathbf{W}^T \mathbf{x}_k$, дающая оценку принадлежности объекта \mathbf{x}_k классу j . Функция потерь в таком случае выглядит следующим образом:

$$L_k = -\log \left(\frac{e^{f_{y_k}(\mathbf{x}_k, \mathbf{W})}}{\sum_j e^{f_j(\mathbf{x}_k, \mathbf{W})}} \right) = -f_{y_k} + \log \sum_j e^{f_j(\mathbf{x}_k, \mathbf{W})} \quad (15)$$

Кросс-энтропия между оценённым распределением q и настоящим p даётся формулой:

$$H(p, q) = -\sum_x p(x) \log q(x) \quad (16)$$

В нашем случае $q = \frac{e^{f_{y_k}(\mathbf{x}_k, \mathbf{W})}}{\sum_j e^{f_j(\mathbf{x}_k, \mathbf{W})}}$, $p = (0, 0, 0, \dots, 1, \dots, 0)$, где единице стоит на y_k месте. То есть данная функция потерь минимизирует кросс-энтропию между оценённым и истинным распределением.

Также такой выбор функции потерь можно интерпретировать как желание найти максимум логарифма функции правдоподобия для данного рапределения.

Итоговая функция потерь выглядит следующим образом:

$$L = \frac{1}{N} \sum_k L_k + R(W) \quad (17)$$

Практический совет: из-за использования экспонент в функции потерь могут возникнуть проблемы с численными подсчётами, потому что значения $e^{f_j(\mathbf{x}_k, \mathbf{W})}$ могут оказаться очень большими. Для избежания этих проблем стоит сделать поправку на константу:

$$\frac{e^{f_{y_k}(\mathbf{x}_k, \mathbf{W})}}{\sum_j e^{f_j(\mathbf{x}_k, \mathbf{W})}} = \frac{C e^{f_{y_k}(\mathbf{x}_k, \mathbf{W})}}{C \sum_j e^{f_j(\mathbf{x}_k, \mathbf{W})}} = \frac{e^{f_{y_k}(\mathbf{x}_k, \mathbf{W}) + \log C}}{\sum_j e^{f_j(\mathbf{x}_k, \mathbf{W}) + \log C}} \quad (18)$$

Для уверенности можно выбрать $\log(C) = -\max_j f_j$. Тогда все экспоненты будут давать значение меньше единицы, что даст уверенность в точности вычислений.

Остаётся подсчитать градиент функции потерь:

$$p_k = \frac{e^{f_{y_k}}}{\sum_j e^{f_j}} \quad L_k = -\log p_k$$

$$\frac{\partial L_k}{\partial f_j} = p_j - \mathbf{1}(k = y_k) \quad (19)$$

$$\frac{\partial L_k}{\partial \mathbf{w}_j} = (p_j - \mathbf{1}(k = y_k)) \mathbf{x}_k \quad (20)$$

4 Реализация методов

4.1 Реализация релаксационного метода

Везде далее `np` - библиотека `numpy`, ускоряющая и упрощающая многие вычисления. Итак, приступим к реализации: создаётся свой класс, который называется `one_weight_linear_model`, имеющий вектор весов w , число объектов N , число признаков D (включая константу), число шагов перед исключением неравенств `steps_to_exclude`, параметр θ , который называется `theta`, параметр ε : `eps` и долю исключаемых неравенств `part_to_exclude`, потому что, если исключать по одному неравенству на большом наборе данных, то слишком долго будет находиться решение. Вектор весов сначала пустой, так как размерность данных изначально неизвестна.

```
class one_weight_linear_model():
    def __init__(self, method, theta, eps, steps_to_exclude,
                 part_to_exclude):
        self.w = None
        self.D = None
        self.N = None
        self.theta = theta
        self.eps = eps
        self.part_to_exclude = part_to_exclude
        self.steps_to_exclude = steps_to_exclude
```

Далее нужно описать метод `fit(self, X, y)`, принимающий на вход матрицу объектов и вектор номеров классов. Но для начала преобразовывается X так, что для объектов класса 0 добавляется ещё один константный признак, равный единице, а объекты класса 1 также получают константный признак, а затем домножаются на -1. Таким образом матрица X - матрица коэффициентов системы линейных неравенств, для которой все неравенства должны быть больше нуля. Так же инициализируются веса модели w . Воспользовавшись простыми соображениями о том, что примерно половина весов будет больше нуля, половина - меньше нуля, w инициализируются случайными числами из нормального распределения. Чтобы отклик не был слишком большим, для большего числа размерности устанавливается меньшая дисперсия, пропорционально \sqrt{k} . Эти подготовительные действия помещаются в отдельный метод `prepare_data(self, X, y)`

```
def prepare_data(self, X, y):
    X = np.array(X)
    self.N = X.shape[0]
    X = np.concatenate((X, np.ones([self.N, 1])), axis=1)
    self.D = X.shape[1]
    self.w = np.random.normal(0, 2. / self.D, self.D)
    X = np.apply_along_axis(lambda a: a * (-1) ** (y + 1), 0, X)
    return X
```

Вызов этого метода делается первым в методе `fit`. Далее нужно приготовить словарь номеров неравенств, которые не выполняются. Для этого предлагается хранить вектор `unfulfilled` (изначально из нулей) размера n .

Если неравенство под номером i не было выполнено, то i -тый элемент вектора **unfulfilled** под номером i будет увеличен на единицу. Теперь нужно выбрать, какой шаг будет делаться. Здесь реализован второй вариант. Написана функция **calc_dw(self, X)**, которая будет высчитывать направление шага, которое будет умножаться на **theta**:

```
def calc_dw(self, X):
    negatives = X[X.dot(self.w) <= 0]
    scalar_mults = np.abs(negatives.dot(self.w))
    dw = np.sum(np.apply_along_axis(lambda a: a * scalar_mults, 0,
                                    negatives), axis=0)
    dw *= np.sum(scalar_mults ** 2) / np.sum(dw ** 2)
    return dw
```

Здесь **negatives** - те объекты, для которых неравенства не выполняются, **scalar_mults** - это матрица попарных произведений $x_{ji}w_{ki}$, где j - номер объекта, i - номер признака.

Есть один нюанс: может случиться так, что **dw** окажется очень маленьким по норме. для таких случаев написано дополнительное условие, которое бы не позволяло делить на 0 в вычислении согласно формуле (4):

```
if np.any(np.abs(dw) > 1e-10):
    dw *= np.sum(scalar_mults ** 2) / np.sum(dw ** 2)
else:
    dw *= np.sum(scalar_mults ** 2) / 1e-11
```

Здесь, если видно, что по модулю **dw** очень мало, то делится не на 0, а на очень маленькую величину.

Теперь для обучения нужно отбрасывать те неравенства, которые не выполняются чаще всего. Для этого пишется метод **delete_worst(self, X, unfulfilled)**, который на вход получает **X** - матрицу объекты-признаки и вектор **unfulfilled**, содержащий в себе частоту невыполнения каждого из неравенств:

```
def delete_worst(self, X, unfulfilled):
    for i in range(max(1, int(self.N * self.part_to_exclude))):
        num_to_del = np.argmax(unfulfilled)
        X = np.delete(X, num_to_del, 0)
        unfulfilled = np.delete(unfulfilled, num_to_del, 0)
    return X, unfulfilled
```

Видно, что исключается по крайней мере одно неравенство, которое чаще всего не выполняется.

Итак, всё готово для написания метода **fit**. Последовательность действий довольно проста: подготавливается матрица объекты-признаки методом **prepare_data**, затем идёт цикл, где на каждом шаге делается один шаг на **dw**. Каждые **steps_to_exclude** шагов отбрасывается одно или часть **part_to_exclude** неравенств, пока не будут выполнены все неравенства оставшейся системы.

```
def fit_relax(self, X, y):
    if self.method == 'relax':
```

```

X = self.prepare_data(X, y)
unfulfilled = np.zeros(self.N)
step_number = 0
while np.any(X.dot(self.w) < self.eps) > 0:
    dw = self.calc_dw(X)
    self.w += self.theta * dw
    unfulfilled += X.dot(self.w) < self.eps
    step_number += 1
    if step_number % self.steps_to_exclude == 0:
        X, unfulfilled = self.delete_worst(X, unfulfilled)

```

Также нужно написать метод, который будет предсказывать классы для объектов после обучения **predict(self, X)**:

```

def predict(self, X):
    X = np.array(X)
    return (X.dot(self.w[: -1]) + self.w[-1] >= 0).astype(np.int)

```

4.2 Реализация метода сведения к линейному программированию

Для реализации данного метода нам нужно уметь решать задачи линейного программирования. Для этого будем использовать библиотеку `python-scipy.optimize`, в которой реализована нужная нам функция **linprog**. Переписывать конструктор класса нам не нужно, потому что все нужные нам гиперпараметры уже присутствуют в том, который мы написали ранее.

Так как мы максимизируем ε , а метод **linprog** минимизирует линейную функцию $\mathbf{c}^T \mathbf{w}$, а нам хочется максимизировать ε , что равносильно минимизации $-\varepsilon$, то в нашем случае $\mathbf{c} = (\underbrace{0, 0, \dots}_{D+1 \text{ нулей}}, -1)^T$. Условия выглядят

следующим образом:

$$A_{ub} \mathbf{w} \leq \mathbf{b}_{ub}. \quad (21)$$

Нам нужно их свести к

$$X \mathbf{w} \geq \boldsymbol{\varepsilon}, \text{ где } \boldsymbol{\varepsilon} = (\underbrace{\varepsilon, \varepsilon, \dots, \varepsilon}_{N_{train} \text{ штук}})^T, \quad (22)$$

что равносильно $X \mathbf{w} - \boldsymbol{\varepsilon} \geq \mathbf{0}$ или $-X \mathbf{w} + \boldsymbol{\varepsilon} \leq \mathbf{0}$. Для упрощения записи введём новый вектор $\tilde{\mathbf{w}} = (\mathbf{w}, \varepsilon)^T$ и новую матрицу объекты-признаки $\tilde{X} = (\begin{array}{c|c} -X & \mathbf{1} \end{array})$. Тогда (22) эквивалентно $\tilde{X} \tilde{\mathbf{w}} \leq \mathbf{0}$.

Таким образом, в роли A_{ub} будет \tilde{X} , в роли \mathbf{b}_{ub} будет $(\underbrace{0, 0, \dots, 0}_{N_{train} \text{ штук}})^T$.

Итак, напомним метод **prepare_for_linprog(self, X, y)** для подготовки матрицы объекты признаки \tilde{X} , вектора \mathbf{c} и границ **bounds**. Начальный вид для \tilde{X} и \mathbf{c} описаны выше. Пойдём, как нужно задать границы для переменных. Ясно, что вектор весов \mathbf{w} не должен быть ограничен какими-то конкретными числами (хотя желательно, чтобы по модулю они не были слишком велики). Но если мы не ограничим ε сверху, то, решение вовсе не будет ограничено, а значит, не будет найдено. Причина этому проста:

допустим, что у нас есть какое-то решение $\tilde{\mathbf{w}}^*$ такое, что последняя его переменная, то есть ε , больше нуля. Тогда, так как система (21) однородна, есть решение $\tilde{\mathbf{w}}' = 2 \cdot \tilde{\mathbf{w}}^*$, у которого ε в 2 раза больше. Поэтому нам нужно ограничить ε сверху любым положительным числом, например, единицей. Ниже приведена реализация метода:

```
def prepare_for_linprog(self, X, y):
    X = self.prepare_data(X, y)
    X = np.concatenate((- X, np.ones([self.N, 1])), axis=1)
    c = np.concatenate([np.zeros(self.D), [-1]])
    bounds = [[None, None] for i in np.arange(self.D + 1)]
    bounds[-1][1] = 1
    return X, c, bounds
```

Далее нужно написать метод `exclude_with_largest_angles`, который будет исключать неравенства, вектор \mathbf{x}_k которых имеет наибольший косинус угла со средним вектором остальных невыполненных:

```
def exclude_with_largest_angles(self, X, res):
    unfulfilled = X[:, :-1].dot(res[:-1]) <= 0
    sumed = np.sum(X[unfulfilled, :-1], axis=0)
    cosines = np.ones(X.shape[0])
    for j in np.arange(len(unfulfilled)):
        if not unfulfilled[j]:
            continue
        rest_of_sum = sumed - X[j, :-1]
        cosines[j] = X[j, :-1].dot(rest_of_sum) / \
            rest_of_sum.dot(rest_of_sum) / X[j, :-1].dot(X[j, :-1])
        to_exclude = []
    for i in np.arange(max(1, self.part_to_exclude * self.N)):
        argmin = np.argmin(cosines)
        cosines[argmin] = 1
        to_exclude.append(argmin)
    X = np.delete(X, to_exclude, axis=0)
    return X
```

В первом цикле мы находим косинусы для каждого невыполненного неравенства. В следующем цикле сохраняем номера тех, которые собираемся удалять. Далее производим удаление. На число невыполненных неравенств делить не обязательно, потому что это общий коэффициент для всех косинусов, поэтому лишнее деление отброшено.

Остаётся собрать всё вместе и написать метод `fit_using_linprog(self, X, y)` и дописать метод `fit(self, X, y)`:

```
def fit_using_linprog(self, X, y):
    X, c, bounds = self.prepare_for_linprog(X, y)
    res = np.nan
    while True:
        b_ub = np.zeros(X.shape[0])
        res = op.linprog(c, A_ub=X, b_ub=b_ub, bounds=bounds).x
        if res[-1] > 0:
            break
    X = self.exclude_with_largest_angles(X, res)
```

```

        if res is not np.nan:
            self.w = res[:-1]

def fit(self, X, y):
    if self.method == 'relax':
        self.fit_using_relaxation(X, y)
    elif self.method == 'linprog':
        self.fit_using_linprog(X, y)

```

4.3 Реализация аппроксимации гладкой функцией

Для данной реализации также понадобится библиотека **scipy.optimize**, из которой мы будем использовать функцию **minimize**.

Для подготовки данных просто воспользуемся написанной **prepare_data**. Тогда все наши объекты можно будет считать за объекты класса **1**, тогда перед скалярным произведением в показателе экспоненты будет просто единичный

$$\sigma(\mathbf{x}_k^T \mathbf{w}) = \frac{1}{1 + \exp(\mathbf{x}_k^T \mathbf{w})}$$

$$\frac{\partial \sigma(\mathbf{x}_k^T \mathbf{w})}{\partial w_j} = -\sigma(\mathbf{x}_k^T \mathbf{w}) \cdot (1 - \sigma(\mathbf{x}_k^T \mathbf{w})) \cdot x_{kj}$$

Остаётся написать методы для вычисления самой функции штрафов L и для вычисления её градиента. Назовём их **calc_sigmoid_L(self, X, w)** и **calc_grad_L(self, X, w)**:

```

def calc_sigmoid_L(self, X, w):
    return (1 / (1 + np.exp(X.dot(w)))) .sum()
def calc_grad_L(self, X, w):
    s = (1 / (1 + np.exp(X.dot(w))))
    return - X.T.dot(s * (1 - s))

```

И остаётся просто собрать всё вместе в методе **fit_approx(self, X, y)**:

```

def fit_approx(self, X, y):
    X = self.prepare_data(X, y)
    self.w = op.minimize(lambda w: self.calc_sigmoid_L(X, w),
        self.w, method='BFGS', jac=lambda w: self.calc_grad_L(X, w)).x

```

Метод **fit** теперь выглядит следующим образом:

```

def fit(self, X, y):
    if self.method == 'relax':
        self.fit_using_relaxation(X, y)
    elif self.method == 'linprog':
        self.fit_using_linprog(X, y)
    elif self.method == 'approx':
        self.fit_approx(X, y)
    else:
        print('No such method', self.method)

```

4.4 Реализация метода опорных векторов (SVM)

Создан отдельный класс `many_weight_linear_model`. Для SVM требуется реализовать подсчёт функции потерь и её градиента. Снова нужна подготовка данных. Реализация `prepare_svm(self, X, y)`:

```
def prepare_data(self, X, y):
    X = np.array(X)
    y = np.array(y).astype(int)
    X = np.concatenate((X, np.ones([X.shape[0], 1])), axis=1)
    self.N = X.shape[0]
    self.D = X.shape[1]
    self.M = np.unique(y).shape[0]
    self.W = np.random.randn(X.shape[1], np.max(y) + 1) /
        np.sqrt(X.shape[1] / 2)
    self.t = 0
    self.m = np.zeros(self.W.shape)
    self.v = np.zeros(self.W.shape)
    return X, y
```

В этом классификаторе и следующем для минимизации градиентного спуска используется метод минимизации *adam*. Для его реализации нужно уметь подсчитывать градиент функции потерь. Функция, вычисляющая градиент и сами потери, названа `loss_and_grad_svm(self, X, y)`.

```
def loss_and_grad_svm(self, X, y):
    scores = X.dot(self.W)
    right_scores = scores[np.arange(self.N), y].reshape([self.N, 1])
    margins = np.maximum(0, scores - right_scores + self.delta)
    margins[np.arange(self.N), y] = 0
    loss = np.sum(margins) / self.N + self.lmda * np.sum(self.W * self.W)

    margins = margins > 0
    to_minus = np.sum(margins, axis=1)
    coefs_0 = ((margins[:, 0] > 0) - to_minus * (y ==
        0)).reshape([self.N, 1])
    derivatives = np.sum(X * coefs_0, axis=0)
    derivatives = derivatives.reshape([1, self.D])
    for i in np.arange(1, self.M):
        coefs_i = ((margins[:, i] > 0) - to_minus * (y ==
            i)).reshape([self.N, 1])
        derivatives = np.concatenate((derivatives, [np.sum(X * coefs_i,
            axis=0)]))
    dW = derivatives.T / self.N + 2 * self.lmda * self.W

    return loss, dW
```

Отдельно реализован метод `adam(self, X, y)`.

```
def adam(self, X, y):
    loss, dW = self.loss_and_grad(X, y)
    self.t += 1
    self.m = self.m * self.beta_1 + (1 - self.beta_1) * dW
    self.v = self.v * self.beta_2 + (1 - self.beta_2) * dW * dW
```

```

m = self.m / (1 - self.beta_1 ** self.t)
v = self.v / (1 - self.beta_2 ** self.t)
self.W -= self.learning_rate * self.m / (np.sqrt(self.v) + self.eps)

```

Метод **loss_and_grad** - обобщение того же для svm.

```

def loss_and_grad(self, X, y):
    if self.method == 'softmax':
        return self.loss_and_grad_softmax(X, y)
    else:
        return self.loss_and_grad_svm(X, y)

```

Для простоты проводится просто 1000 итераций сдвига весов без анализа изменения функции потерь.

```

def fit(self, X, y):
    X, y = self.prepare_data(X, y)
    for step in range(1000):
        self.adam(X, y)

```

Функция **predict** принимает новый вид:

```

def predict(self, X):
    X = np.concatenate((X, np.ones([X.shape[0],1])), axis=1)
    scores = X.dot(self.W)
    return np.argmax(scores, axis=1)

```

4.5 Реализация метода минимизации кросс-энтропии (Softmax)

Всё, что нужно для реализации - написать метод **loss_and_grad_softmax(self, X, y)**. Вся остальная часть заложена в уже реализованных **fit** и **adam**.

```

def loss_and_grad_softmax(self, X, y):
    scores = X.dot(self.W)
    scores -= np.max(scores, axis=1, keepdims=True)
    exp_scores = np.exp(scores)
    probs = exp_scores / np.sum(exp_scores, axis=1, keepdims=True)

    correct_logprobs = - np.log(probs[np.arange(X.shape[0]), y])
    loss = np.sum(correct_logprobs) + self.lmda * np.sum(self.W * self.W)

    dscores = probs
    dscores[np.arange(X.shape[0]), y] -= 1
    dscores /= X.shape[0]
    dW = np.dot(X.T, dscores) + self.lmda * 2 * self.W

    return loss, dW

```

5 Сравнение на модельных данных

Для начала стоит сравнить классификаторы на модельных данных. Были сгенерированы разные виды выборок:

- Выборка, имеющая 20 признаков. Для каждого признака $\mathbb{E}[x^1] = \Delta + \mathbb{E}[x^2]$, где верхний индекс - номер выборки. Объекты сгенерированы из нормального распределения, где $\mathbb{E}[x^1] = -0.5, \mathbb{E}[x^2] = 0.5, \mathbb{D}[x^i] = \sigma^2$, значение дисперсии изменяется в разных экспериментах, $\Delta = 1$.
- Датасет аналогичен первому, но есть три признака, данные для которого сгенерированы из одного распределения. Здесь тест покажет, могут ли классификаторы определять значимость признаков.
- Выборка, имеющая 2000 признаков. Объекты сгенерированы из нормального распределения, где $\mathbb{E}[x^1] = -0.5, \mathbb{E}[x^2] = 0.5, \mathbb{D}[x^i] = \sigma^2$, значение дисперсии изменяется в разных экспериментах, $\Delta = 1$.
- **breast cancer wisconsin dataset** - датасет по раковым больным в Висконсине.

Значение σ менялось от 0.5 до 2, приложен график количества правильных ответов для каждого из классификаторов для разных дисперсий.

5.1 Первая выборка

В обучающей выборке - 400 объектов, в тестовой - 200. Проводилось по 5 тестов на одну дисперсию. Дисперсия менялась в диапазоне от 0.7 до 2. Параметры для тестирования:

```
sigmas = np.linspace(0.7, 2, 14)
m1 = -0.5
m2 = 0.5
train_size = 400
test_size = 200
tests_num = 5
results = np.zeros([len(sigmas), 5])

lin_m = mm.one_weight_linear_model(method='linprog')
rel_m = mm.one_weight_linear_model(method='relax')
log_m = mm.one_weight_linear_model(method='approx')
soft_m = mm.many_weight_linear_model(method='softmax')
svn_m = mm.many_weight_linear_model(method='svm')
models = [lin_m, rel_m, log_m, soft_m, svn_m]
```

Приведён график 1, на котором видно, как менялась доля верных предсказаний для каждого классификатора.

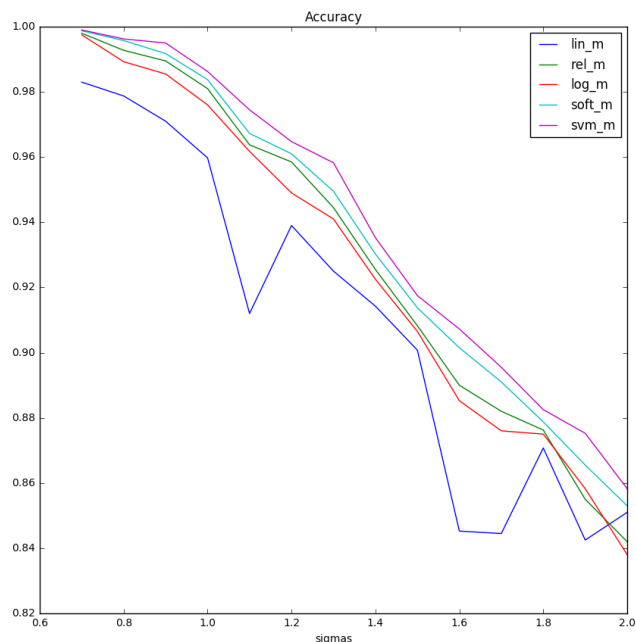


Рис. 1: График доли верных предсказаний для классификаторов на 1 выборке

На тестах оказалось, что при увеличении числа признаков довольно сильно начинает проигрывать по верности работы классификатор, отнормированный на решении задачи линейного программирования и логистическая регрессия.

Также его точно сильно флуктуирует. Отбросим его для следующих датасетов с большим числом признаков.

5.2 Вторая выборка

Пример генерации выборки:

```
X1 = np.concatenate((np.random.normal(loc=0.5, scale=sigmas[i],
size=[int(train_size / 2), 17]),
                    np.random.normal(loc=-0.5, scale=sigmas[i],
size=[train_size - int(train_size / 2), 17])))
X1 = np.concatenate((X1, np.random.normal(size=[train_size, 3])), axis=1)
y1 = np.concatenate((np.ones(int(train_size / 2)), np.zeros(train_size -
int(train_size / 2))), axis=0)
```

Тесты были проведены полностью аналогичные первым. В результате сформирован график точностей классификаторов 2. Видно, что все классификаторы работают с примерно одинаковой точностью.

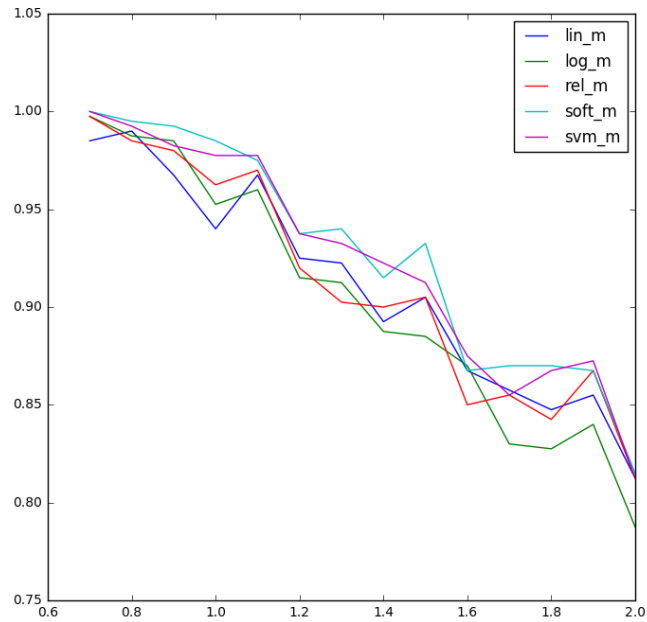


Рис. 2: График доли верных предсказаний для классификаторов на 2 выборке

5.3 Третья выборка

На данных с такой же дисперсией при большом числе признаков предсказания были верны на всех значениях, поэтому все значения **sigmas** были увеличены в 10 раз. График с результатами указан на рисунке 3.

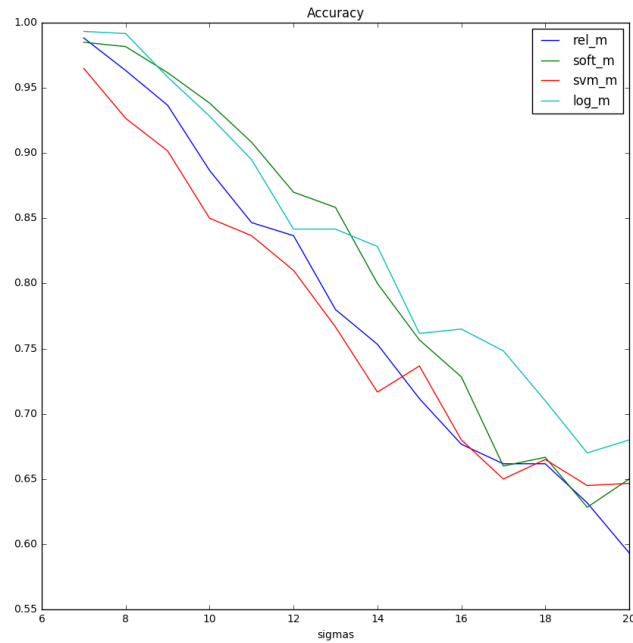


Рис. 3: График доли верных предсказаний для классификаторов на 3 выборке

5.4 Четвёртая выборка

Датасет содержит 30 признаков, все они числовые, нет нужны в каких-либо преобразованиях, кроме центрирования и деления на стандартное отклонение. Было выбрано 25% случайных объектов под тестовую выборку, остальные использовались как обучающая выборка.

Точность предсказаний указана на графике 4. Видно, что все классификаторы, кроме первого имеют примерно одинаковые результаты.

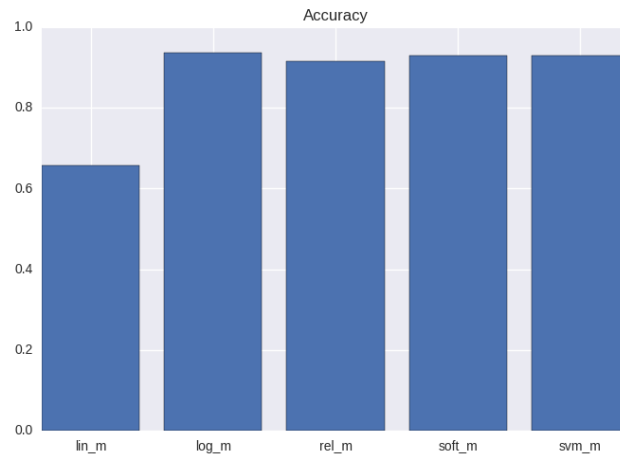


Рис. 4: График доли верных предсказаний для классификаторов на 4 выборке

6 Выводы

6.1 Реализованы:

- Релаксационный метод
- Метод сведения к линейному программированию
- Метод аппроксимации гладкой функцией
- Метод опорных векторов
- Метод минимизации кросс-энтропии

6.2 Исследованы:

- качества классификаторов на различных выборках
- скорости работы на различных выборках

6.3 Проведено сравнение:

Классификатор, сводящий задачу к линейному программированию продемонстрировал себя чуть хуже остальных по доле верных предсказаний и был самым медленным. Однако его плюс в том, что если система заранее разделима, то мы сразу получим $\varepsilon \geq 0$, в случае $\varepsilon = 0$ нам нужно убедиться, что веса модели не занулились. Неравенство $|\mathbf{w}| > 0$ говорит о том, что есть нормаль к двум полуплоскостям, в которых находятся объекты классов. То есть сведение задачи к линейному программированию позволяет сказать, есть ли между начальными данными разделяющая гиперплоскость.

Аппроксимация гладкой функцией , чего и следовало ожидать, проявила себя хорошо и в плане точности предсказаний, и в плане времени работы.

Релаксационный метод также оказался достаточно хорош и работал быстрее остальных. Можно заключить, что он вполне жизнеспособен при условии, что есть уверенность, что и обучающая, и тренировочная выборка достаточно схожи. Основной плюс данного классификатора в высокой скорости работы.

Метод опорных векторов также проявил себя вполне успешно. Является разумным выбором в случае классификации с большим числом классов, если речь идёт о линейной классификации.

Минимизация кросс-энтропии , что подтверждается статьями, ведёт себя примерно на одном уровне с методом опорных векторов.

7 Используемая литература

Список литературы

- [1] The relaxation method for linear inequalities, Canad. J. Math. 6(1954), 382-392.
- [2] CS231n Convolutional Neural Networks for Visual Recognition <http://cs231n.github.io/linear-classify/>
- [3] Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989.
- [4] Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. — М.: Наука, 1974. — 416 с.
- [5] Вапник В. Н. Восстановление зависимостей по эмпирическим данным. — М.: Наука, 1979. — 448 с.
- [6] Дуда Р., Харт П. Распознавание образов и анализ сцен. — М.: Мир, 1976.
- [7] Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning, 2nd edition. — Springer, 2009. — 533 p.
- [8] Langford J. Tutorial on Practical Prediction Theory for Classification. — 2005. — 28 P.
- [9] McAllester D. Simplified PAC-Bayesian Margin Bounds. — 2003.
- [10] Анисимов Б.В., Курганов В.Д., Злобин В.К. Распознавание и цифровая обработка изображений. - М.: Высшая школа, 1983. - 295 с.
- [11] Горелик А.Л., Скрипкин В.А. Методы распознавания. - М.: Высшая школа, 1984. - 208с.
- [12] Дуда Р., Харт П. Распознавание образов и анализ сцен: Пер. с англ. - М.: Мир, 1976. - 512 с.
- [13] Ту Дж., Гонсалес Р. Принципы распознавания образов: Пер. с англ. - М.: Мир, 1978. - 412с.
- [14] Фомин Я.А., Тарловский Г.Р. Статистическая теория распознавания образов. — М.: Радио и связь, 1986. — 264с.
- [15] Фукунага К. Введение в статистическую теорию распознавания образов: Пер. с англ. - М.: Наука, 1979. - 368с.
Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. — Springer, 2001.
- [16] Hebb D. The organization of behavior. — New York: Wiley, 1949.
- [17] LeCun Y., Bottou L., Orr G. B., Muller K.-R. Efficient BackProp // Neural Networks: tricks of the trade. — Springer, 1998.

- [18] McCulloch W. S., Pitts W. A logical calculus of ideas immanent in nervous activity // Bulletin of Mathematical Biophysics. — 1943. — no. 5. — Pp. 115–133.
- [19] Mercer J. Functions of positive and negative type and their connection with the theory of integral equations // Philos. Trans. Roy. Soc. London. — 1909. — Vol. A, no. 209. — Pp. 415–446.
- [20] Novikoff A. B. J. On convergence proofs on perceptrons // Proceedings of the Symposium on the Mathematical Theory of Automata. — Vol. 12. — Polytechnic Institute of Brooklyn, 1962. — Pp. 615–622