

# РЕФЕРАТ

---

Выпускная квалификационная работа содержит 25 страниц, 1 рисунка, 1 таблицу, XX использованных источника.

КЛЮЧЕВОЕ СЛОВО 1, КЛЮЧЕВОЕ СЛОВО 2, ...

Краткое описание содержания работы.

# СОДЕРЖАНИЕ

---

<b>Введение .....</b>	<b>4</b>
<b>Теоретическая часть .....</b>	<b>6</b>
1.1 Сентимент-анализ текстов .....	6
1.1.1 Методы, основанные на правилах и словарях .....	8
1.2 Основная идея архитектуры нейронных сетей .....	8
1.2.1 Функции активации .....	10
1.3 Обработка текстов .....	10
1.3.1 Дистрибутивная семантика .....	11
1.3.2 Методы основанные на подсчетах .....	11
1.3.3 Распределенные представления слов .....	11
1.3.4 GloVe .....	16
1.4 Алгоритмы классического машинного обучения .....	16
1.4.1 Логистическая регрессия .....	16
1.4.2 Метод опорных векторов .....	18
<b>Практическая часть .....</b>	<b>19</b>
<b>Заключение .....</b>	<b>20</b>
<b>Список использованных источников .....</b>	<b>21</b>
<b>Приложение .....</b>	<b>22</b>

## ВВЕДЕНИЕ

---

Классификация текстов — невероятно популярная задача. Мы пользуемся текстовыми классификаторами в почтовом клиенте: он классифицирует письма и фильтрует спам. Другие приложения включают классификацию документов, обзоров и так далее.

Обычно классификация текстов используется не как самостоятельная задача, а является частью более крупного пайплайна. Например, голосовой помощник классифицирует ваше высказывание, чтобы понять, что вы хотите (например, установить будильник, заказать такси или просто поболтать) и передать ваше сообщение другой модели в зависимости от решения классификатора. Другой пример — поисковый движок: он может использовать классификаторы для определения языка запроса, чтобы предсказать его тип (например, информационный, навигационный, транзакционный) и понять хотите ли вы увидеть картинки или видео помимо документов и прочего.

Моя работа посвящена одному из приложений классификации — автоматическому определению эмоциональной оценки русскоязычных текстов. Главная особенность заключается в том, что предсказание базируется на эмоциональных моделях, предложенных Робертом Плутчиком и Полом Экманом.

Поскольку большинство датасетов для классификации содержат только одну правильную метку. А у нас многоклассовая классификация.

Задачи интеллектуальной обработки текста делятся на три типа: синтаксические, основанные на понимании смысла и третий — не просто понимание написанного, а еще и генерация нового текста. Сентимент-анализ относится ко второму типу.

# ОСНОВНАЯ ЧАСТЬ

# ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

---

В этой части мы подробнее раскроем теорию.

## 1.1. СЕНТИМЕНТ-АНАЛИЗ ТЕКСТОВ

Сентимент-анализ, как направление компьютерной лингвистики, стал очень популярен последние десятилетие. С появлением больших данных насыщенных эмоциональной составляющей возникла потребность в их обработке. Компании начали устраивать соревнования с внушительными призовыми фондами, а исследователи искать лучшую архитектуру для классификации этих данных. Так в открытом доступе появились большие размеченные датасеты с отзывами, данными из соцсетей и новостями.

Сам термин «сентимент» означает совокупность чувств и взглядов, как основа для действия или суждения; общая эмоциональная установка. Целью сентимент-анализа является выделение этих тональных компонент из текста. Рассмотрим его применение на разных уровнях [7].

Пусть есть целый документ, тогда, как правило, в нем можно выделить один субъект и один объект, а так же сентимент. Ярким примером такого уровня является отзыв. Здесь автор выступает в качестве субъекта, а предмет отзыва — в качестве объекта. Это уровень документа.

Если документ более сложный, то можно рассматривать его на уровне предложений. В результате можно определить эмоциональную окраску всего документа или предложений по отдельности. Зависит от поставленной задачи.

Также анализ бывает на уровне аспектов. Смысл его в том, что эмоциональная установка определяется не для конкретно объекта, а для его отдельных составляющих — аспектов. Например, для объекта «компьютер» можно выделить аспекты «производительность», «дизайн», «сборка» т.д., другими словами, к аспектам относится все то, к чему могут быть выражены эмоции. Данная задача очень востребована, потому что позволяет точнее определять отношение автора к объекту, а для некоторых задач это очень важный критерий.

Последний вид анализа самый сложный и проводится на уровне именованных сущностей (Named Entities). Именованная сущность — это абстрактный или физический объект, который может быть обозначен собственным именем. Сама по себе задача извлечения именованных сущностей (Named

Entity Recognition) не из простых, а вкупе с sentiment-анализом становится действительно комплексным решением.

Методы sentiment анализа можно также разделить на несколько основных направлений [7]:

- *метод основанный на правилах (rule-based)*. Здесь используются наборы правил классификации, составленные экспертом и эмоционально размеченные словари. Определенный класс присваивается в зависимости от найденных ключевых слов и их использования с другими ключевыми словами. Такой метод достаточно эффективный, но очень трудоемкий;
- *метод основанный на словарях*. Самый простой метод, основанный на подсчете sentimentных единиц содержащихся в словарях тональности. Очень сильно зависит от размера словаря и не очень точен в разрыве с правилами русского языка. Попытка получения sentimentа из текста таким способом описана в части ++ этой работы;
- *методы основанные на машинном и глубоком обучении*. (machine learning, deep learning) Наиболее популярная группа методов в sentiment-анализе, работающих на способности алгоритмов обобщать выделенные из текста признаки;
- *гибридные методы*. Позволяют одновременно использовать несколько подходов.

Разберем подробнее машинное и глубокое обучение. Суть метода в выделении признаков из текста и последующем их обобщении с помощью разнообразных алгоритмов. Для выделения признаков используют, как простые алгоритмы по типу мешок слов (Bag of Words) или TF-IDF, так и небольшие нейронные сети для генерирования эмбеддингов, например, Word2Vec ++, GloVe ++, FastText ++. Существуют и более сложные алгоритмы, которые формируют признаки на уровне предложений, к ним относятся ELMo ++, BERT (Bidirectional Encoder Representations from Transformers) ++ и др.

Чтобы обработать выделенные признаки используют разнообразные алгоритмы. К классическому обучению относятся:

- Байесовский классификатор (Naive Bayes classifier);
- дерево решений (Decision Tree);
- логистическая регрессия (Logistic Regression);
- метод опорных векторов (Support Vector Machine).

Среди алгоритмов глубокого обучения можно выделить:

-

### 1.1.1. МЕТОДЫ, ОСНОВАННЫЕ НА ПРАВИЛАХ И СЛОВАРЯХ

ИЛИ

## 1.2. ОСНОВНАЯ ИДЕЯ АРХИТЕКТУРЫ НЕЙРОННЫХ СЕТЕЙ

Чтобы объяснить работу нейронных сетей нужно определить из чего они состоят. Для этого рассмотрим простейший линейный бинарный классификатор – перцептрон Розенблата. Пространство данных разделяется на два множества гиперплоскостью, а метка класса будет ставиться в зависимости от значения линейной функции от входных признаков.

$$\text{sign}(w_0 + w_1x_1 + w_2x_2 + \dots + w_dx_d), \quad (1.1)$$

где  $x = (x_1, x_2, \dots, x_d) \in \mathbb{R}^d$ . Мы ищем такие веса  $w_0, w_1, \dots, w_d \in \mathbb{R}^d$ , чтобы  $\text{sign}$  от скалярного произведения признаков и весов  $w^\top x$  совпадал с верной меткой  $y(x) \in -1, 1$ , но для этого добавим фиктивную переменную в вектор  $x = (1, x_1, x_2, \dots, x_d) \in \mathbb{R}^{d+1}$ , чтобы размерности сохранялись.

Теперь обучим эту функцию, для этого нам нужна функция ошибки, она называется критерий Перцептрона:

$$L_P(w) = - \sum_{x \in M} y(x)(w^\top x), \quad (1.2)$$

где  $M$  — множество неверно классифицируемых примеров. В качестве оптимизатора выберем градиентный спуск. С помощью него мы минимизируем суммарное отклонение предсказаний классификатора от верных, но только в неправильную сторону. Верное предсказание никак не влияет на функцию ошибки. В результате получается кусочно-линейная функция, которая почти везде дифференцируема и этого достаточно для применения градиентного спуска. Процесс обучения выглядит так: если предсказание верное, то не делаем ничего, если классификатор ошибся, то делаем градиентный шаг.

Такая модель перцептрона линейная и результат ее работы не слишком содержателен. Чтобы из перцептронов можно было составить сеть, нужно добавить нелинейность. Такой нелинейностью будет функция активации. Они бывают разные, самая распространенная — сигмоида рис. 1.1:

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (1.3)$$

Обучить этот перцептрон также не составляет труда. Просто теперь мы будем решать задачу бинарной классификации, а функция ошибки будет cross-энтропией:

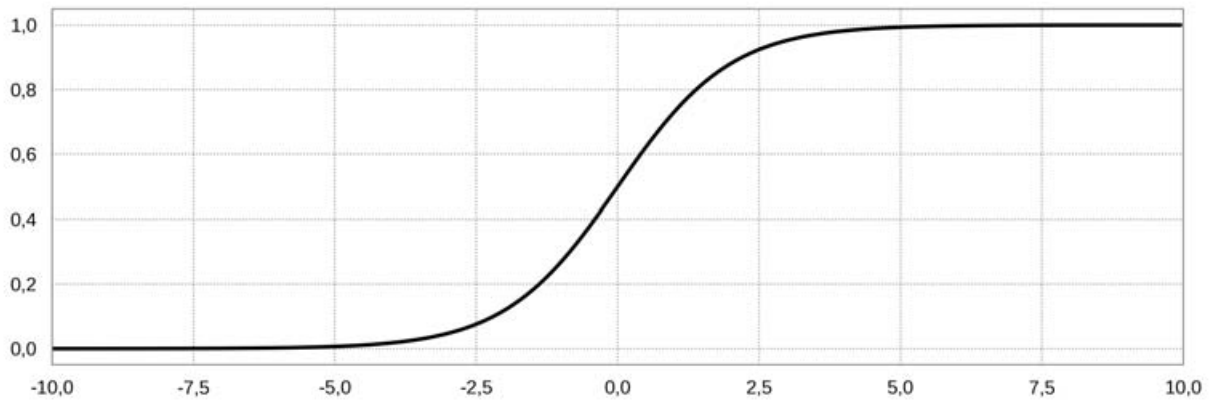


Рисунок 1.1 — Сигмоида

$$L(w) = -\frac{1}{N} \sum_{i=1}^N (y_i \log \sigma(w^\top x_i) + (1 - y_i) \log(1 - \sigma(w^\top x_i))). \quad (1.4)$$

Эта функция дифференцируема, значит мы можем сделать градиентный шаг. При этом прецептрон с сигмодой реализует логистическую регрессию. Обобщение этой модели можно найти в разделе 1.4.1.

Графическое изображение структуры перцептрона представлено на рис. 1.2, а.

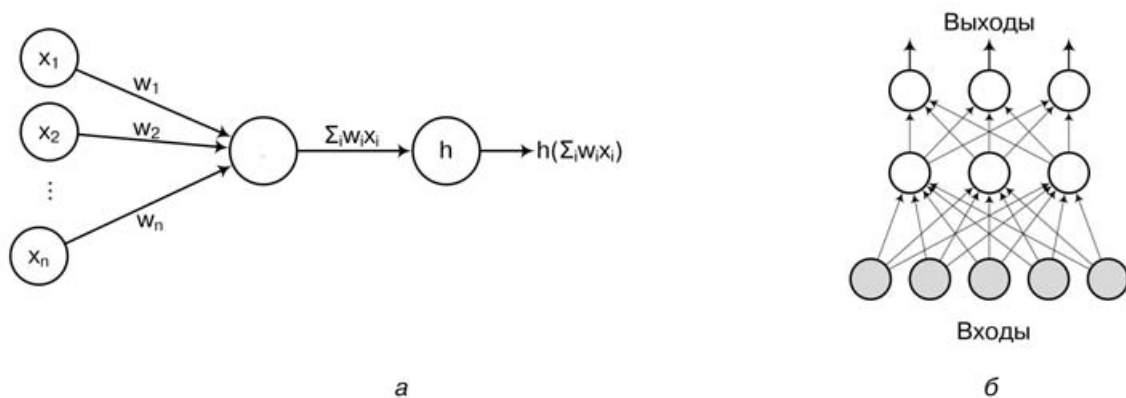


Рисунок 1.2 — (а) граф вычислений перцептрона; (б) полносвязная нейронная сеть с одним скрытым слоем.

На рис. 1.2, б изображена несложная нейронная сеть, на ее примере покажем как можно векторизовать вычисления в слое нейронов с применением функции активации.

Пусть у нас в слое  $k$  нейронов с весами  $w_1, w_2, \dots, w_k$ ,  $w_i = (w_{i1}, \dots, w_{in})^\top$ , на вход вектор  $x = (x_1, x_2, \dots, x_n)^\top$ . В результате получим выход  $y_i = f(w_i^\top x)$ , где  $f$  — функция активации. Эти вычисления можно представить в векторной форме:



$$\begin{pmatrix} y_1 \\ \dots \\ y_k \end{pmatrix} = y = f(Wx) = \begin{pmatrix} f(w_1^\top x) \\ \dots \\ f(w_k^\top x) \end{pmatrix}, \text{ где } W = \begin{pmatrix} w_1 \\ \dots \\ w_k \end{pmatrix} = \begin{pmatrix} w_{11} & \dots & w_{1n} \\ \dots & & \dots \\ w_{k1} & \dots & w_{kn} \end{pmatrix}.$$

### 1.2.1. ФУНКЦИИ АКТИВАЦИИ

Обычно в нелинейных нейронах применяется сигмоида 1.3, о которой писалось выше. Но существуют и другие функции активации, например,  $\text{ReLU} = \max(0, x)$ . Чтобы понять, как она появилась рассмотрим сумму бесконечного ряда сигмOID, каждая из которых смещена на единицу относительно предыдущей:

$$f(x) = \sigma(x + \frac{1}{2}) + \sigma(x - \frac{1}{2}) + \sigma(x - \frac{3}{2}) + \dots$$

$f(x)$  можно представить в виде интеграла:

$$\int_{\frac{1}{2}}^{\infty} \sigma(x + \frac{1}{2} - y) dy,$$

Его значение можно приуменьшить единичными прямоугольниками:

$$\begin{aligned} \sum_{i=0}^{\infty} \sigma(x + \frac{1}{2} - i) &\approx \int_{\frac{1}{2}}^{\infty} \sigma(x + \frac{1}{2} - y) dy = \\ &= \left[ -\log(1 + e^{x+\frac{1}{2}-y}) \right]_{y=\frac{1}{2}}^{y=\infty} = \log(1 + e^x), \end{aligned}$$

а это в точности интеграл от сигмOIDы:

$$\int \sigma(x) dx = \log(1 + e^x) + C.$$

Получается бесконечный ряд сигмOID гораздо более выразительная функция активации и это почти тоже самое что и  $\log(1 + e^x)$

## 1.3. ОБРАБОТКА ТЕКСТОВ

То, как модели видят данные отличается от того, как их видят люди. Например, мы легко можем понять предложение «Да будет свет», но модели не могут — им нужны векторы с признаками. Такие векторы являются представлениями слов, которые может обработать наша модель.

Самой простой формой представления слов является дискретное представление, т.е. one-hot представление: все слова представляются в виде вектора, размерность которого совпадает с числом слов словаре. Причем все компоненты кроме  $i$ -го равны нулю, а позиция, соответствующая  $i$ -му слову равна единице. Очевидно, что такой способ не самый лучший. Во-первых такое представление зависит от положения слов в словаре, а это нежелательно, потому что задает бессмысленные отношения между словами. Во-вторых размеры такого словаря растут прямо пропорционально количеству слов в нем, а это значит размерность его может достигать до сотен тысяч и работать с ними станет очень вычислительно накладно. В-третьих такое представление совершенно не учитывает значение слов, для решения этой проблемы обратимся к дистрибутивной семантике.

### 1.3.1. ДИСТРИБУТИВНАЯ СЕМАНТИКА

Чтобы зафиксировать значение слов в их векторах, нам нужно сначала определить понятие значения. Для этого возьмем несколько предложений

- возьмем несколько предложений
- построим вектора
- способы их сравнить (косинусное расстояние)

### 1.3.2. МЕТОДЫ ОСНОВАННЫЕ НА ПОДСЧЕТАХ

PPMI

LSA

### 1.3.3. РАСПРЕДЕЛЕННЫЕ ПРЕДСТАВЛЕНИЯ СЛОВ

Подход к обучению моделей распределенных представлений слов был описан в работе Йошуа Бенджи с соавторами [1], которая была продолжена в [6]. Идея подхода описанного в [1] основанна на задаче построения языковой модели, процесс обучения выглядит так:

- всем токенам из словаря  $i \in V$  ставят в соответствие вектор признаков  $w_i$  размерности  $d$  ( $w_i \in \mathbb{R}^d$ ). Стандартным значением  $d$  является 300;
- теперь можно определить вероятности для каждого токена  $i$ , что он появится в контексте  $c_1, \dots, c_n$ . Для этого определим функцию от векторов признаков  $w$ :

$$\hat{p}(i|c_1, \dots, c_n) = f(w_i, w_{c_1}, \dots, w_{c_n}; \theta), \quad (1.5)$$

- где  $w_{c_1}, \dots, w_{c_n}$  — векторы признаков токенов из контекста, а  $f$  — функция с параметрами  $\theta$ , которая принимает векторы признаков;
- максимизируя логарифм правдоподобия большого корпуса текстов можно обучить векторы признаков  $w$  и параметры  $\theta$

$$L(W, \theta) = \frac{1}{K} \sum_t \log f(w_k, w_{k-1}, \dots, w_{k-n+1}; \theta) + R(W, \theta), \quad (1.6)$$

где  $K$  размер окна контекста, а  $R(W, \theta)$  — регуляризация.

Для получения функции  $f$  можно использовать нейронную сеть. Модель word2vec строится на описании нейросетевой модели, предложенной в [1]. Она была разработана Томасом Миколовым с соавторами и опубликована в работах [3, 2], причем в двух вариациях:

- CBOW (Continuous Bag Of Words) — по контексту восстановить слово;
- skip-gram — восстановить контекст в зависимости от слова;

Архитектура word2vec представляет собой полносвязную нейронную сеть с одним скрытым слоем рис. 1.3.

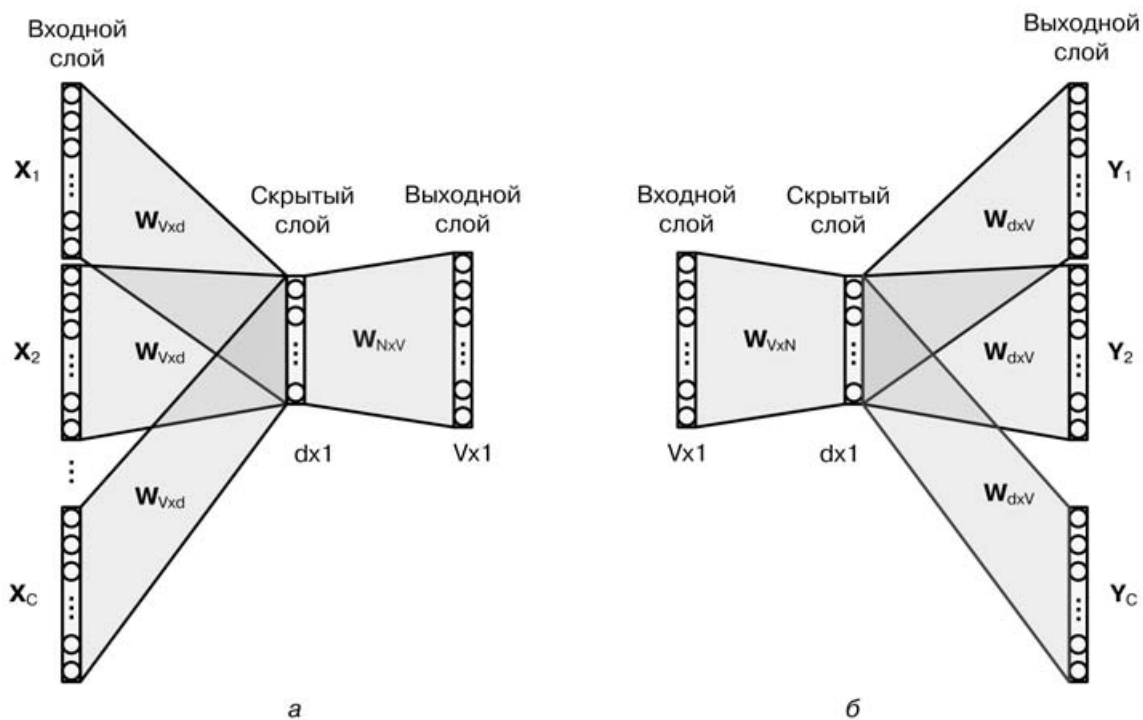


Рисунок 1.3 — (а) CBOW; (б) skip-gram.

Принцип работы модели CBOW рис. 1.3, а выглядит так:

- на вход сети подаются one-hot вектора размерности  $V$ , где  $V$  — это размер словаря;
- скрытый слой — это матрица  $W$  размерности  $V \times d$ , которая переводит наши представления слов в  $d$ -мерное пространство;

- на выходе для каждого слова в словаре берем среднее всех полученных векторов и получаем оценку  $u_j$ , где  $j = 1, \dots, V$ .

Чтобы найти апостериорное распределение модели, просто вычисляем softmax:

$$\hat{p}(i|c_1, \dots, c_n) = \frac{\exp u_j}{\sum_{j'=1}^V \exp u_{j'}}. \quad (1.7)$$

Для аппроксимации апостериорным распределением распределения данных используем loss-функцию для одного окна:

$$L = -\log p(i|c_1, \dots, c_n) = -u_j + \log \sum_{j'=1}^V \exp u_{j'}. \quad (1.8)$$

Принцип работы модели skip-gram рис. 1.3, *б* полностью противоположный. До этого мы усредняли контекст, чтобы получить среднее слово в окне, а теперь будем предсказывать слова контекста исходя из центрального слова. На выходе мы получаем  $K - 1$  мультиномиальных распределений (центральное слово не учитывается):

$$\hat{p}(c_k|i) = \frac{\exp u_{kc_k}}{\sum_{j'=1}^V \exp u_{j'}}, \quad (1.9)$$

loss-функция для окна размера  $K$  выглядит так:

$$L = -\log p(c_1, \dots, c_n|i) = -\sum_{k=1}^K u_{kc_k} + n \log \sum_{j'=1}^V \exp u_{j'}. \quad (1.10)$$

Возникает вопрос, как же обучить такую модель? Этот процесс хорошо описан в докладе Голдберга и его соавторов [4].

Подробно разберем модель skip-gram для корпуса документов  $D$ . Нашей задачей стоит нахождение оптимальных параметров модели  $\theta$ , чтобы максимизировать функцию правдоподобия:

$$L(\theta) = \prod_{i \in D} \left( \prod_{c \in C(i)} p(c|i; \theta) \right) = \prod_{(i,c) \in D} p(c|i; \theta), \quad (1.11)$$

где  $C(i)$  — множество контекстных слов внутри окна вокруг центрального слова  $i$ . Вероятность  $p(c|i; \theta)$  определяется, как softmax-функция, зависящая от всех возможных векторов контекста.

$$p(c|i; \theta) = \frac{\exp \tilde{w}_c^\top w_i}{\sum_{c'} \exp \tilde{w}_{c'}^\top w_i}, \quad (1.12)$$

где  $\tilde{w}_c$  — вектор признаков слова из контекста  $c$ , который отличается от  $w_i$ . Для каждого слова  $i$  надо обучить два вектора признаков  $w_i$  и  $\tilde{w}_i$ , в первом случае это слова выступает в качестве центрального, во втором в качестве контекстного.

Эта особенность обучения, когда мы берем два разных вектора одного и того же слова вместо одного, описана в [4]. И мотивирована тем, что слова редко встречаются в контексте себя самих. Вот, например, слово «мотивация» вряд ли можно встретить в контексте другого слова «мотивация», под это правило попадают почти все слова. Поэтому в процессе обучения модель сведет вероятности  $p(i|i, \theta)$  к нулю. А если вектора контекста и центрального слова будут равны нулю, то норма вектора  $|w_i| = w_i^\top w_i$  тоже будет равняться нулю, а это очень не желательно. Поэтому для каждого слова мы обучаем два разных вектора.

Теперь выразим максимум функции правдоподобия для всего корпуса через логарифм 1.11 и 1.12:

$$\begin{aligned} \arg \max_{\theta} \prod_{(i,c) \in D} p(c|i; \theta) &= \arg \max_{\theta} \sum_{(i,c) \in D} \log p(c|i; \theta) = \\ &= \arg \max_{\theta} \sum_{(i,c) \in D} \left( \exp \tilde{w}_c^\top w_i - \log \sum_{c'} \exp \tilde{w}_{c'}^\top w_i \right). \end{aligned} \quad (1.13)$$

Оптимизируя данную функцию мы получаем хорошее распределенное представление слов. Но для этого нужно решить сложнейшую задачу: суммировать скалярные произведения всех возможных слов и их контекста  $\sum_{c'} \tilde{w}_{c'}^\top w_i$  при том, что размер словаря может достигать миллионов.

Чтобы уменьшить количество вычислений Миколов с соавторами [2] предложили элегантный метод: negative sampling. Нам не нужно считать всю сумму  $\sum_{c'} \tilde{w}_{c'}^\top w_i$ , а только случайно выбрать несколько ее элементов в качестве отрицательных примеров (примеры в которых слово не находится в определенном контексте) и обновить только их. Т.е. теперь нам нужно посчитать только небольшую сумму  $\sum_{c' \in D'} \tilde{w}_{c'}^\top w_i$ , где  $D'$  — случайное подмножество отрицательных примеров.

По сути negative sampling — это тоже правдоподобие, но другого события. Пусть у нас есть слово  $i$  и его контекст  $c$ , наша задача максимизировать вероятность  $p((i, c) \in D; \theta)$ , параметризованную вектором  $\theta$ , т.е. правдоподобие появления пары  $(i, c)$ :

$$\arg \max_{\theta} \prod_{(i,c) \in D} p((i, c) \in D; \theta) = \arg \max_{\theta} \sum_{(i,c) \in D} \log p((i, c) \in D; \theta). \quad (1.14)$$

Выразим  $p((i, c) \in D; \theta)$  через softmax. Но так как это бинарное событие, то заменим softmax сигмодой  $\sigma(x) = \frac{1}{1+\exp(-x)}$ :

$$p((i, c) \in D; \theta) = \frac{1}{1 + \exp(-\tilde{w}_c^\top w_i)} \quad (1.15)$$

Максимизируем логарифм правдоподобия:

$$\begin{aligned} \arg \max_{\theta} \sum_{(i,c) \in D} \log p((i, c) \in D; \theta) &= \\ &= \arg \max_{\theta} \sum_{(i,c) \in D} \log \frac{1}{1 + \exp(-\tilde{w}_c^\top w_i)}. \end{aligned} \quad (1.16)$$

Из 1.16 видно, что оптимальное значение логарифма будет получено при максимальном значении скалярного произведения  $\tilde{w}_c^\top w_i$ . Сделаем равные векторы с большой нормой и можно без проблем получить правдоподобие почти равное единице. Подвох заключается в том, что модель обучается на данных для бинарной классификации, но мы рассматриваем только набор состоящий из положительных примеров. Классификатор, который всегда предсказывает «да» — плохой. Поэтому имеет смысл добавить отрицательных примеров, просто случайно выбирая слова и контекст, которых нет в данных. После того, как мы получим набор отрицательных данных, максимизация правдоподобия будет выглядеть так:

$$\arg \max_{\theta} \prod_{(i,c) \in D} p((i, c) \in D; \theta) \prod_{(i',c') \in D'} p((i', c') \notin D; \theta) \quad (1.17)$$

Выразим в 1.17 пару  $(i, c) \in D$ :

$$\begin{aligned} \arg \max_{\theta} \prod_{(i,c) \in D} p((i, c) \in D; \theta) \prod_{(i',c') \in D'} 1 - p((i', c') \in D; \theta) &= \\ = \arg \max_{\theta} \left[ \sum_{(i,c) \in D} \log p((i, c) \in D; \theta) \sum_{(i',c') \in D'} \log (1 - p((i', c') \in D; \theta)) \right] &= \\ = \arg \max_{\theta} \sum_{(i,c) \in D} \left[ \log \frac{1}{1 + \exp(-\tilde{w}_c^\top w_i)} \sum_{(i',c') \in D'} \log \frac{1}{1 + \exp(\tilde{w}_{c'}^\top w_i)} \right] &= \\ = \arg \max_{\theta} \sum_{(i,c) \in D} \left[ \log \sigma(\tilde{w}_c^\top w_i) \sum_{(i',c') \in D'} \log \sigma(-\tilde{w}_{c'}^\top w_i) \right] \end{aligned} \quad (1.18)$$

Получили формулу для negative sampling из [2]. Значит мы для каждого окна случайно берем несколько отрицательных примеров  $D'$  и делаем градиентный шаг для loss-функции:

$$L = \log \sigma(\tilde{w}_c^\top w_i) \sum_{(i,c') \in D'} \log \sigma(-\tilde{w}_{c'}^\top w_i) \quad (1.19)$$

Аналогичные рассуждения можно провести для модели CBOW.

### 1.3.4. GloVe

Модель GloVe [5] представляет собой комбинацию методов на основе подсчета и методов предсказания (например, word2vec). Название модели GloVe расшифровывается как «Global Vectors», что отражает ее идею: использование информации из всего корпуса для обучения векторов.

## 1.4. АЛГОРИТМЫ КЛАССИЧЕСКОГО МАШИННОГО ОБУЧЕНИЯ

### 1.4.1. ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Логистическая регрессия — классическая дискриминативной линейная модель классификации. Дискриминативная значит, что нас интересует  $P(y = k|x)$ , а не совместное распределение  $p(x, y)$ . Свое начало она берет из расстояния Кульбака-Лейблера. Оно задается формулой:

$$KL(P||Q) = \int \log \frac{dP}{dQ} dP, \quad (1.20)$$

, где  $P$  — истинное распределение, а  $Q$  — приближенное. Для дискретного случая:

$$KL(P||Q) = \sum_y p(y) \log \frac{p(y)}{q(y)}, \quad (1.21)$$

а если раскрыть получаем:

$$\begin{aligned} KL(P||Q) &= \sum_y p(y) \log \frac{p(y)}{q(y)} = \\ &= \sum_y p(y) \log p(y) - \sum_y p(y) \log q(y) = -H(p) + H(p, q), \end{aligned} \quad (1.22)$$

, где  $H(p)$  — энтропия распределения  $p$ , а  $H(p, q)$  — наша кросс энтропия. Из этой суммы видно, что нам нужно минимизировать  $H(p, q)$ . Для бинарной классификации loss-функция будет выглядеть так:

$$L(w) = H(p_{data}, q(w)) = -\frac{1}{N} \sum_{i=1}^N (y_i \log \hat{y}_i(w) + (1 - y_i) \log(1 - \hat{y}_i(w))) \quad (1.23)$$

где  $p_{data}$  — распределение наших данных,  $q(w)$  — апостериорное распределение,  $\hat{y}_i(w)$  — оценка вероятности при входных параметрах  $w$  и  $y_i$  — истинное предсказание. Два слагаемых мы получаем, т.к. события несовместные. Например, в тексте говорится о кошечках или о собачках, события появления кошечки или собачки несовместные, т.е.  $p(\text{кошечки}) = 1 - p(\text{собачки})$ . Если мы предсказываем кошечку (1), как абсолютный 0 или собачку (0), как 1, то ошибка будет бесконечной из первого и второго слагаемых соответственно — это не допустимо.

Перед тем, как перейти к нескольким классам, рассмотрим сначала задачу классификации с Байесовской точки зрения: определим для каждого класса  $C_k$  плотность  $p(x|C_k)$  и какие-то априорные распределения  $p(C_k)$  (пусть это будут размеры классов, т.е. мы ничего не знаем о примере, но предполагаем с какой вероятностью он относится к конкретному классу) и найдем  $p(C_k|x)$ . Для двух классов:

$$\begin{aligned} p(C_1|x) &= \frac{p(x|C_1)p(C_1)}{p(x|C_1)p(C_1) + p(x|C_2)p(C_2)} = \\ &= 1 / \frac{(p(x|C_1)p(C_1) + p(x|C_2)p(C_2))}{p(x|C_1)p(C_1)} = \\ &= 1 / (1 + \frac{p(x|C_2)p(C_2)}{p(x|C_1)p(C_1)}) = \frac{1}{1 + e^{-a}} = \sigma(a), \end{aligned} \quad (1.24)$$

где

$$a = \ln \frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)}, \quad \frac{1}{1 + e^{-a}} = \sigma(a). \quad (1.25)$$

Используя логистическую регрессию мы делаем предположение о виде аргумента сигмной функции  $a$  — это будет скалярное произведение вектора признаков на вектор данных:  $a = w_{\top} x$ . Сигмоида переводит результат вычисления этой линейной функции на отрезок  $[0; 1]$  и как результат мы получаем апостериорную вероятность первого или второго классов:

$$p(C_1|x) = y(x) = \sigma(w_{\top} x), \quad p(C_2|x) = 1 - p(C_1|x), \quad (1.26)$$

чтобы обучить эту модель мы можем просто оптимизировать правдоподобие по  $w$ .

Для набора  $x_n, t_n$ , где  $x_n$  — входы, а  $t_n \in \{0; 1\}$  — соответствующие метки классов, получается такое правдоподобие:



$$p(t|w) = \prod_{n=1}^N y_n^{t_n} (1 - y_n)^{1-t_n}, \quad \text{где } y_n = p(C_1|x_n). \quad (1.27)$$

И теперь мы, максимизируя логарифм вероятности, ищем наилучшие параметры функции правдоподобия для этого можно использовать различные оптимизаторы.

$$\text{Likelihood}(w) = -\ln p(t|w) = -\sum_{n=1}^N [t_n \ln y_n + (1 - t_n) \ln(1 - y_n)]. \quad (1.28)$$

Теперь можно легко обобщить задачу на несколько классов. Только вместо сигмоиды будем использовать softmax функцию. Для  $K$  классов получаем:

$$p(C_k|x) = \frac{p(x|C_k)p(C_k)}{\sum_{j=1}^K p(x|C_j)p(C_j)} = \frac{e^{a_k}}{\sum_{j=1}^K e^{a_j}}, \quad (1.29)$$

где количество аргументов  $a_k = \ln p(x|C_k)p(C_k)$  равняется количеству классов. Функция правдоподобия почти не изменилась. Пусть на вход метки класса подаются в формате one-hot векторов, тогда для набора векторов  $T = t_n$  функция правдоподобия выглядит следующим образом:

$$p(T|w_1, \dots, w_K) = \prod_{n=1}^N \prod_{k=1}^K p(C_k|x_n)^{t_{nk}} = \prod_{n=1}^N \prod_{k=1}^K y_{nk}^{t_{nk}}. \quad (1.30)$$

, где  $y_{nk} = y_k(x_n)$ . Опять переходим к логарифму и получаем функцию максимального правдоподобия для  $K$  классов:

$$\text{Likelihood}(w_1, \dots, w_K) = -\ln p(T|w_1, \dots, w_K) = \sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk}. \quad (1.31)$$

#### 1.4.2. МЕТОД ОПОРНЫХ ВЕКТОРОВ

## ПРАКТИЧЕСКАЯ ЧАСТЬ

---

## ЗАКЛЮЧЕНИЕ

---

## СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

---

1. *Bengio Y., Ducharme R., Vincent P.* A neural probabilistic language model : тех. отч. — 2001. — С. 1137—1155.
2. Distributed representations of words and phrases and their compositionality / Т. Mikolov [и др.] // Advances in Neural Information Processing Systems. — 2013. — Окт. — ISSN 10495258. — arXiv: 1310.4546. — URL: <http://arxiv.org/abs/1310.4546>.
3. Efficient estimation of word representations in vector space / Т. Mikolov [и др.] // 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings. — International Conference on Learning Representations, ICLR, 01.2013. — arXiv: 1301.3781. — URL: <http://ronan.collobert.com/senna/>.
4. *Goldberg Y., Levy O.* word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method : тех. отч. — 2014. — arXiv: 1402.3722. — URL: <http://arxiv.org/abs/1402.3722>.
5. *Pennington J., Socher R., Manning C. D.* GloVe: Global vectors for word representation // EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. — Association for Computational Linguistics (ACL), 2014. — С. 1532—1543. — ISBN 9781937284961. — DOI: 10.3115/v1/d14-1162. — URL: <http://nlp..>
6. A neural probabilistic structured-prediction model for transition-based dependency parsing / H. Zhou [и др.] // ACL-IJCNLP 2015 - 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, Proceedings of the Conference. Т. 1. — Association for Computational Linguistics (ACL), 2015. — С. 1213—1222. — ISBN 9781941643723. — DOI: 10.3115/v1/p15-1117. — URL: <https://www.aclweb.org/anthology/P15-1117>.
7. *Semina T. A.* Sentiment analysis: Modern approaches and existing problems. // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Серия 6: Языкознание. Реферативный журнал. — 2020. — С. 47—64.

## ПРИЛОЖЕНИЕ

---





pic example

$$\begin{aligned} F &\rightarrow x \mid y \mid (S) \\ T &\rightarrow F \mid T * F \\ S &\rightarrow T \mid S + T \end{aligned}$$

(а)



(б)

Рисунок 2.1 — (а) Арифметические выражения; (б) Look R96 2016.

Таблица 2.1 — Расчет параметров

Параметр $x_i$	Параметр $x_j$				Первый шаг		Второй шаг	
	$X_1$	$X_2$	$X_3$	$X_4$	$w_i$	$K_{\text{в}i}$	$w_i$	$K_{\text{в}i}$
$X_1$	1	1	1.5	1.5	5	0.31	19	0.32
$X_2$	1	1	1.5	1.5	5	0.31	19	0.32
$X_3$	0.5	0.5	1	0.5	2.5	0.16	9.25	0.16
$X_4$	0.5	0.5	1.5	1	3.5	0.22	12.25	0.20
Итого:					16	1	59.5	1