

НАЗВАНИЕ УЧРЕЖДЕНИЯ, В КОТОРОМ ВЫПОЛНЯЛАСЬ ДАННАЯ  
ДИССЕРТАЦИОННАЯ РАБОТА

На правах рукописи

УДК **xxx.xxx**

Фамилия Имя Отчество автора

НАЗВАНИЕ ДИССЕРТАЦИОННОЙ РАБОТЫ

Специальность **XX.XX.XX** —

«**Название специальности**»

Диссертация на соискание учёной степени  
кандидата **физико-математических наук**

Научный руководитель:

**уч. степень, уч. звание**

**Фамилия Имя Отчество**

Город — 20XX

## Оглавление

	Стр.
<b>Введение</b> . . . . .	4
<b>Глава 1. Text classification</b> . . . . .	6
1.1 Relevance of the problem . . . . .	6
1.2 Statement of classification problem . . . . .	6
1.3 Short review of existing mathematical models, which can be used to solve the classification problem . . . . .	8
<b>Глава 2. Mathematical models and algorithms for text         classification</b> . . . . .	11
2.1 Words representations . . . . .	11
2.1.1 Bag-of-Words Approach . . . . .	11
2.1.2 TF-IDF Approach . . . . .	13
2.1.3 Embeddings . . . . .	14
<b>Глава 3. Вёрстка таблиц</b> . . . . .	19
3.1 Таблица обыкновенная . . . . .	19
<b>Заключение</b> . . . . .	20
<b>Список литературы</b> . . . . .	21
<b>Список рисунков</b> . . . . .	22
<b>Список таблиц</b> . . . . .	23
<b>Приложение А. Примеры вставки листингов программного кода</b>	24
<b>Приложение Б. Очень длинное название второго приложения,         в котором продемонстрирована работа с         длинными таблицами</b> . . . . .	30
Б.1 Подраздел приложения . . . . .	30
Б.2 Ещё один подраздел приложения . . . . .	32

Б.3	Очередной подраздел приложения . . . . .	36
Б.4	И ещё один подраздел приложения . . . . .	36

## Введение

Обзор, введение в тему, обозначение места данной работы в мировых исследованиях и т. п.

**Целью** данной работы является . . .

Для достижения поставленной цели необходимо было решить следующие задачи:

1. Исследовать, разработать, вычислить и т. д. и т. п.
2. Исследовать, разработать, вычислить и т. д. и т. п.
3. Исследовать, разработать, вычислить и т. д. и т. п.
4. Исследовать, разработать, вычислить и т. д. и т. п.

**Основные положения, выносимые на защиту:**

1. Первое положение
2. Второе положение
3. Третье положение
4. Четвертое положение

**Научная новизна:**

1. Впервые . . .
2. Впервые . . .
3. Было выполнено оригинальное исследование . . .

**Научная и практическая значимость . . .**

**Степень достоверности** полученных результатов обеспечивается . . . Результаты находятся в соответствии с результатами, полученными другими авторами.

**Апробация работы.** Основные результаты работы докладывались на: перечисление основных конференций, симпозиумов и т. п.

**Личный вклад.** Автор принимал активное участие . . .

**Публикации.** Основные результаты по теме диссертации изложены в XX печатных изданиях [?; ?; ?; ?], X из которых изданы в журналах, рекомендованных ВАК [?; ?], XX — в тезисах докладов [?; ?].

**Объем и структура работы.** Диссертация состоит из введения, четырёх глав, заключения и двух приложений. Полный объём диссертации состав-

ляет 36 страниц с 5 рисунками и 3 таблицами. Список литературы содержит 9 наименований.

## Глава 1. Text classification

### 1.1 Relevance of the problem

Nowadays, retail e-commerce sales are quickly expanding. A large online e-commerce websites serve millions of users' requests per day. Therefore it necessary to make the process of registrations and purchases as much convenient and fast as possible. For many classifieds platform such as Amazon or Avito users who would like to create a new advertisement must to fulfill compulsory fields: title, description, price and category. Choosing category can be a tricky moment, because in most cases users have a choice more then from three hundreds categories. Therefore, the problem of advertisement automatic category prediction is very important in terms to save moderators' time and as a result decrease number of necessary moderators to process them. The effective algorithm which would work with text data, have a high accuracy and an appropriate speed are in high demand.

### 1.2 Statement of classification problem

Classification problem - the problem of identifying to which category a new observation belongs. The basic example can be situation when you receive a new email and algorithm automatically decides whether it belongs to social network, promotions or business letters.

In text classification, we are given a description  $d \in \mathbb{X}$  of a document, where  $\mathbb{X}$  is the document space ; and a fixed set of classes  $\mathbb{C} = \{c_1, c_2, \dots, c_J\}$ . Classes are also called categories or labels . Typically, the document space  $\mathbb{X}$  is some type of high-dimensional space, and the classes are human defined for the needs of an application, as in the examples China and documents that talk about multicore computer chips above. We are given a training set  $\mathbb{D}$  of labeled documents  $d$ , where  $d \in \mathbb{X} \times \mathbb{C}$ . For example:

$$\langle d, c \rangle = \langle \text{Beijing joins the World Trade Organization, } \textit{China} \rangle \quad (1.1)$$

for the one-sentence document Beijing joins the World Trade Organization and the class (or label) China. Using a learning method or learning algorithm, we then wish to learn a classifier or classification function  $\gamma$  that maps documents to classes:

$$\gamma : \mathbb{X} \rightarrow \mathbb{C} \quad (1.2)$$

This type of learning is called supervised learning because a supervisor (the human who defines the classes and labels training documents) serves as a teacher directing the learning process. We denote the supervised learning method by  $\Gamma$  and write  $\Gamma(\mathbb{D}) = \gamma$ . The learning method  $\Gamma$  takes the training set  $\mathbb{D}$  as input and returns the learned classification function  $\gamma$ .

The classes in text classification often have some interesting structure such as the hierarchy in Figure 1.1. There are two instances each of region categories, industry categories, and subject area categories. A hierarchy can be an important aid in solving a classification problem. Our goal in text classification is high accuracy on test data or new data - for example, the newswire articles that we will encounter tomorrow morning in the multicore chip example. It is easy to achieve high accuracy on the training set (e.g., we can simply memorize the labels). But high accuracy on the training set in general does not mean that the classifier will work well on new data in an application. When we use the training set to learn a classifier for test data, we make the assumption that training data and test data are similar or from the same distribution. [2, p.256-257]

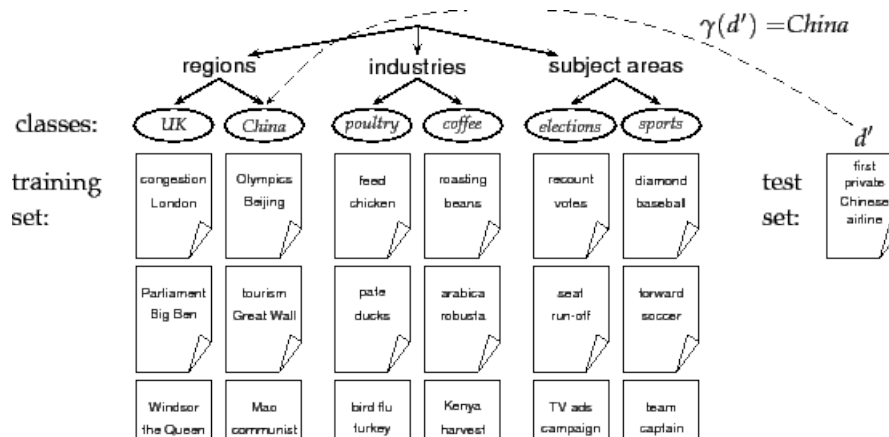


Figure 1.1 — Classes, training set, and test set in text classification.

### 1.3 Short review of existing mathematical models, which can be used to solve the classification problem

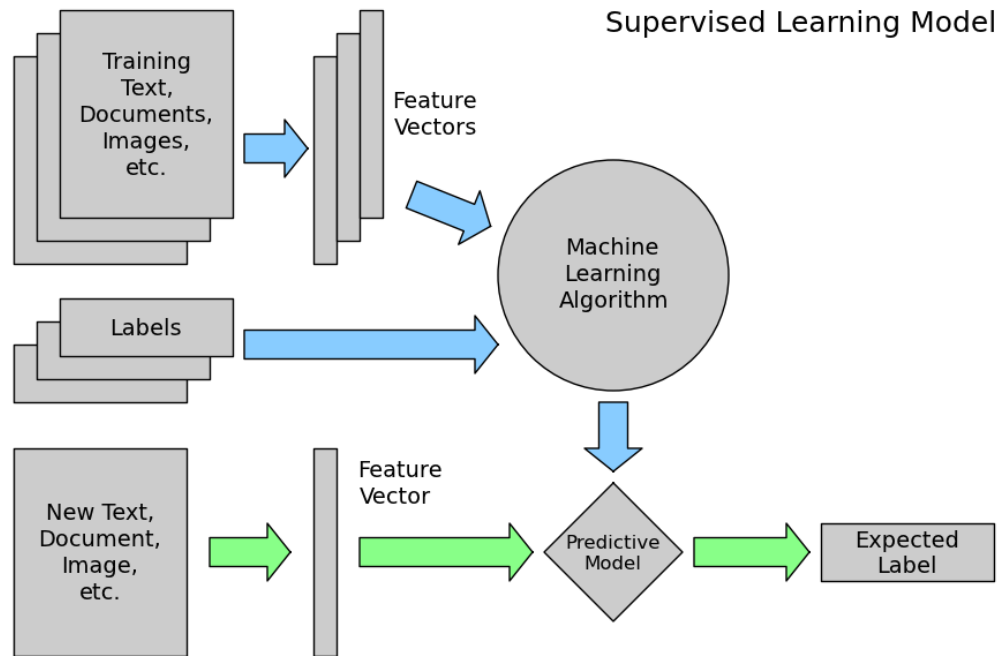


Figure 1.2 — Supervised learning work flow.

Supervised learning - the machine learning task of inferring a function from labeled training data. The training data consist of a set of training examples. Between inputs and reference outputs there may be some dependence, but it is unknown. On the basis of this data, it is necessary to restore the dependence. In order to measure the accuracy a quality function can be introduced. [3, p.7] The diagram of the supervised learning process is presented in Figure 1.2

**Here are some of the most important supervised learning algorithms:**

1. Naive Bayes . [6]. [7]
2. Logistic Regression . [5]
3. Support Vector Machines (SVMs) . [8]
4. Decision Trees and Random Forests . [2]
5. Neural networks . [2]



1. **Naive Bayes** - a probabilistic learning method. The probability of a document  $d$  being in class  $c$  is computed as

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c) \quad (1.3)$$

where  $P(t_k|c)$  is the conditional probability of term  $t_k$  occurring in a document of class  $c$ . We interpret  $P(t_k|c)$  as a measure of how much evidence  $t_k$  contributes that  $c$  is the correct class.  $P(c)$  is the prior probability of a document occurring in class  $c$ . If a document's terms do not provide clear evidence for one class versus another, we choose the one that has a higher prior probability.  $\langle t_1, t_2, \dots, t_{n_d} \rangle$  are the tokens in  $d$  that are part of the vocabulary we use for classification and  $n_d$  is the number of such tokens in  $d$ .

In text classification, our goal is to find the best class for the document. The best class in NB classification is the most likely or maximum a posteriori ( MAP ) class  $c_{map}$ :

$$c_{map} = \underset{c \in \mathbb{C}}{\operatorname{argmax}} \hat{P}(c|d) = \underset{c \in \mathbb{C}}{\operatorname{argmax}} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c). \quad (1.4)$$

For the priors this estimate is

$$\hat{P}(c) = \frac{N_c}{N}, \quad (1.5)$$

where  $N_c$  is the number of documents in class  $c$  and  $N$  is the total number of documents.

We estimate the conditional probability  $\hat{P}(t|c)$  as the relative frequency of term  $t$  in documents belonging to class  $c$ :

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'} + 1)}, \quad (1.6)$$

where  $T_{ct}$  is the number of occurrences of  $t$  in training documents from class  $c$ , including multiple occurrences of a term in a document  $T_{ct}$  is a count of occurrences in all positions  $k$  in the documents in the training set,  $V$  - vocabulary To eliminate zeros, we use add-one or Laplace smoothing, which simply adds one to each count. [2, p.258-260]

```

TRAINMULTINOMIALNB( $\mathbf{C}, \mathbf{ID}$ )
1   $V \leftarrow \text{EXTRACTVOCABULARY}(\mathbf{ID})$ 
2   $N \leftarrow \text{COUNTDOCS}(\mathbf{ID})$ 
3  for each  $c \in \mathbf{C}$ 
4  do  $N_c \leftarrow \text{COUNTDOCSINCLASS}(\mathbf{ID}, c)$ 
5      $\text{prior}[c] \leftarrow N_c / N$ 
6      $\text{text}_c \leftarrow \text{CONCATENATETEXTOFALLDOCSINCLASS}(\mathbf{ID}, c)$ 
7     for each  $t \in V$ 
8     do  $T_{ct} \leftarrow \text{COUNTTOKENSOFTERM}(\text{text}_c, t)$ 
9     for each  $t \in V$ 
10    do  $\text{condprob}[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{c'} (T_{c't}+1)}$ 
11 return  $V, \text{prior}, \text{condprob}$ 

APPLYMULTINOMIALNB( $\mathbf{C}, V, \text{prior}, \text{condprob}, d$ )
1   $W \leftarrow \text{EXTRACTTOKENSFROMDOC}(V, d)$ 
2  for each  $c \in \mathbf{C}$ 
3  do  $\text{score}[c] \leftarrow \log \text{prior}[c]$ 
4     for each  $t \in W$ 
5     do  $\text{score}[c] += \log \text{condprob}[t][c]$ 
6  return  $\arg \max_{c \in \mathbf{C}} \text{score}[c]$ 

```

Figure 1.3 — Naive Bayes algorithm (multinomial model): Training and testing.

**2.Logistic Regression.** For now, we will focus on the binary classification problem in which  $y$  can take on only two values, 0 and 1

## Глава 2. Mathematical models and algorithms for text classification

### 2.1 Words representations

In supervised learning domain, to perform classification tasks, usually our goal is to find a parametrized model, best in its class:

$$A(X, \hat{w}) : A(X, \hat{w}) \simeq f(X) \Leftrightarrow A(X, \hat{w}) = \arg \min_w \|A(X, w) - f(X)\| \quad (2.1)$$

Where  $X \in R^{n \times m}$  - feature matrix ( $n$  observations with  $m$  features),  $w \in R^m$  - vector of model parameters,  $\hat{w}$  - "best" model parameters. However, as a candidate for  $X$  - all that we have is raw text input, algorithms can not use it as it is. In order to apply machine learning on textual data, firstly content should be transformed into specific numerical format, in another words it is necessary to form feature vectors. In Natural Language Processing automated feature extraction may be achieved in many ways.[тут вставить список литературы]

#### 2.1.1 Bag-of-Words Approach

Bag-of-words - an unordered set of words, with their exact position ignored. [1, p.641],

In bag-of-words approach we work under the following assumptions:

- The text can be analyzed without taking into account the word/token order.
- It is only necessary to know which words/tokens the text consists of and how many times.

Formally, there is a collection of texts  $T_1, T_2, \dots, T_n$ . Unique tokens  $w_1, w_2, \dots, w_m$  are extracted to form a dictionary. Thus, each text  $T_i$  is represented by feature vector  $F_j = \{x_{ij}, j \in [1, m]\}$ , where  $x_{ij}$  corresponds to number of occurrences of word  $w_j$  in text  $T_i$ .

Example: Our corpus represented by 2 texts: ["The sun is yellow "The sky is blue"]

Our tokens are simple unigrams, therefore there are 6 unique words: the, sun, is, yellow, sky, blue. Then, given corpus is mapped to feature vectors:  $T_1 = (1,1,1,1,0,0)$ ,  $T_2 = (1,0,1,0,1,1)$

Table 2.1

Feature vector						
Text	the	sun	is	yellow	sky	blue
$T_1$	1	1	1	1	0	0
$T_2$	1	0	1	0	1	1

Benefits:

- Despite its simplicity, demonstrate good results.
- Fast preprocessing.
- Built-in in many scientific/NLP libraries

Drawbacks:

- Huge corpus usually leads to huge vocabulary size.
- Not memory-efficient: if we have corpus with 20 thousand texts then this textual corpus might spawn a dictionary with around 100 thousand elements. Thus, storing feature vectors as an array of type int32 would require  $20000 \times 100000 \times 4$  bytes 8GB in RAM.
- A bag of words is an orderless representation: throwing out spatial relationships between features leads to the fact that simplified model cannot let us to distinguish between sentences, built from the same words while having opposite meanings: "This paintings don't feel like ugly - buy them!"(positive) and "This paintings feel like ugly - don't buy them!"(negative)

In order to capture dependencies between words **N-grams** technique can be used. N-gram is a sequence of  $N$  basic tokens, which can be defined in different ways.

1. **Word n-grams - catches more semantics :**

- unigrams: "The sun is yellow."  $\rightarrow$  ['The', 'sun', 'is' ...]
- bigrams: "The sun is yellow."  $\rightarrow$  ['The sun', 'sun is' ...]
- 3-grams: "The sun is yellow."  $\rightarrow$  ['The sun is ', 'sun is yellow']

2. **Character n-grams - helps to deal with misspelling:**

- 4-grams: "The sun is yellow."  $\rightarrow$  ['The ', 'sun ', 'is y', 'ello' ...]

### 3. Skip-n-gram - sequence of $N$ basic tokens, having distance of $\leq K$ tokens between them

– 1-skip-2-grams:: "The sun is yellow."  $\rightarrow$  ['The is', 'sun yellow ']

In TF-IDF approach (term frequency - inverse document frequency), in addition to usual BoW-model, the following augmentation is made:

#### 2.1.2 TF-IDF Approach

Instead of just counting up the overlapping words, the algorithms applies a weight to each overlapping word. The TF weight measures how many times the word occurs in particular document while the IDF weight measures how many different documents a word occurs in and is thus a way of discounting function words. Since function words like the, of, etc., occur in many documents, their IDF is very low, while the IDF content words is high. [1, p.647] Formaly it can be defined:

$$\begin{cases} TF(w,T) = n_{Tw} \\ IDF(w,T) = \log \frac{N}{n_w} \end{cases} \implies TF-IDF(w,T) = n_{Tw} \log \frac{N}{n_w} \quad \forall w \in W \quad (2.2)$$

where  $T$  corresponds to current document (text),

$w$  - selected word in document  $T$ ,

$n_{Tw}$  - number of occurences of  $w$  in text  $T$ ,

$n_w$  - number of documents, containing word  $w$ ,

$N$  - total number of documents in a corpus.

$$\lim_{n_w \rightarrow N} TF-IDF(w,T) = 0 \quad (2.3)$$

### 2.1.3 Embeddings

Core idea: A word's meaning is given by the words that frequently appear close-by.

тут вставить информацию про работы проведенные в данной области.  
<https://arxiv.org/pdf/1301.3781.pdf> <https://arxiv.org/pdf/1310.4546.pdf>

#### 1. Skip-gram model

To begin with key definitions of softmax 2.4 and sigmoid 2.5 functions,

$$\text{softmax}(\mathbf{x})_i = \frac{e^{x_i}}{\sum_{j=1} e^{x_j}} \quad (2.4)$$

$$\text{softmax} = \sigma(z) = \frac{1}{1 + e^{-z}}. \quad (2.5)$$

The gradient of sigmoid function is follows:

$$\sigma'(z) = \sigma(z)(1 - \sigma(z)) \quad (2.6)$$

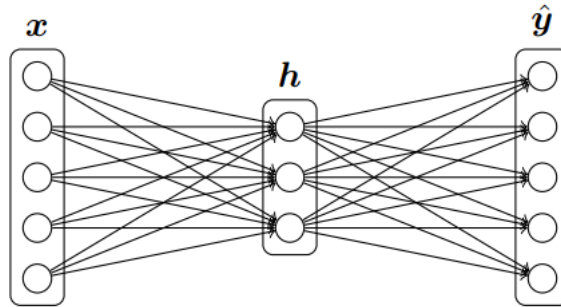


Figure 2.1 — Neural Network

where  $x$  is one-hot input vector,  $h$  - hidden layer,  $y$  is the one-hot label vector, and  $\hat{y}$  is the predicted probability vector for all classes. The neural network employs sigmoid activation function for the hidden layer, and softmax for the output layer and cross entropy cost 2.16 is used.

$$\text{CE}(\mathbf{y}, \hat{\mathbf{y}}) = - \sum_i y_i \log \hat{y}_i \quad (2.7)$$

Now, we will compute the gradient of cross entropy:

$$\frac{\partial(\text{CE})}{\partial \hat{y}_i} = -\frac{y_j}{\hat{y}_i} \quad (2.8)$$

That leads,

$$\frac{\partial(\text{CE})}{\partial \theta_k} = \frac{\partial(\text{CE})}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial \theta_k} = -\frac{y_j}{\hat{y}_i} \frac{\partial \hat{y}_i}{\partial \theta_k} \quad (2.9)$$

Calculating the partial derivative of  $\hat{y}_i$  (where  $i = k$ )

$$\begin{aligned} \frac{\partial \hat{y}_i}{\partial \theta_i} &= \frac{\partial}{\partial \theta_i} \left( \frac{e^{\theta_i}}{\sum_{j=1} e^{\theta_j}} \right) = \\ &= \frac{e^{\theta_i}}{\sum_{j=1} e^{\theta_j}} - \left( \frac{e^{\theta_i}}{\sum_{j=1} e^{\theta_j}} \right)^2 = \\ &= \hat{y}_i \cdot (1 - \hat{y}_i) \end{aligned} \quad (2.10)$$

and (where  $i \neq k$ ),

$$\begin{aligned} \frac{\partial \hat{y}_i}{\partial \theta_k} &= \frac{\partial}{\partial \theta_k} \left( \frac{e^{\theta_i}}{\sum_{j=1} e^{\theta_j}} \right) = \\ &= - \left( \frac{e^{\theta_i} e^{\theta_k}}{\sum_{j=1} e^{\theta_j}} \right) = -\hat{y}_i \hat{y}_k \end{aligned} \quad (2.11)$$

After combination of equations 2.8, 2.10, 2.11,

$$\frac{\partial(\text{CE})}{\partial \theta_k} = \begin{cases} -y_j(1 - \hat{y}_k) & \text{for } i = k \\ y_j \hat{y}_k & \text{for } i \neq k \end{cases} \quad (2.12)$$

$y_j$  should be non-zero,  $k = j$  and  $y_j = 1$ , leads to,

$$\frac{\partial(\text{CE})}{\partial \theta_j} = \begin{cases} (\hat{y}_j - 1) & \text{for } i = j \\ \hat{y}_j & \text{for } i \neq j \end{cases} \quad (2.13)$$

Which is equivalent to,

$$\frac{\partial(\text{CE})}{\partial \boldsymbol{\theta}} = \hat{\mathbf{y}} - \mathbf{y} \quad (2.14)$$

Forward propagation is as follows:

$$\mathbf{h} = \text{sigmoid}(\mathbf{x}\mathbf{W}_1 + \mathbf{b}_1) \quad (2.15)$$

$$\hat{\mathbf{y}} = \text{softmax}(\mathbf{h}\mathbf{W}_2 + \mathbf{b}_2) \quad (2.16)$$

where  $\mathbf{W}_i$  and  $\mathbf{b}_i$  ( $i \in \{1,2\}$ ) are the weights and biases, respectively of the two layers.

To optimize weights for each layer of neural network the back propagation algorithm is used. Therefore, it is necessary to calculate the gradients for each layer. In order to simplify the notation used to solve the problem, define the following terms:

$$\begin{aligned} \mathbf{z}_1 &\equiv \mathbf{x}\mathbf{W}_1 + \mathbf{b}_1 \\ \mathbf{z}_2 &\equiv \mathbf{h}\mathbf{W}_2 + \mathbf{b}_2 \end{aligned} \quad (2.17)$$

Starting with the results from 2.7:

$$\frac{\partial J}{\partial \mathbf{z}_2} = \hat{\mathbf{y}} - \mathbf{y} \quad (2.18)$$

and

$$\frac{\partial \mathbf{z}_2}{\partial \mathbf{h}} = \mathbf{W}_2^\top \quad (2.19)$$

Sigmoid ( $\sigma$ ) derivative 2.6:

$$\frac{\partial \mathbf{h}}{\partial \mathbf{z}_1} \equiv \sigma'(\mathbf{z}_1) \quad (2.20)$$

Combining these, and using  $\cdot$  to denote element-wise product:

$$\frac{\partial J}{\partial z_i} = (\hat{\mathbf{y}} - \mathbf{y})\mathbf{W}_2^\top \cdot \sigma'(\mathbf{z}_1) \quad (2.21)$$

Finally, using the results from Equation 2.19:

$$\frac{\partial J}{\partial \mathbf{W}^{(1)}} = (\hat{\mathbf{y}} - \mathbf{y})\mathbf{W}_2^\top \cdot \sigma'(\mathbf{z}_1) \cdot \mathbf{X}^\top \quad (2.22)$$

$$\frac{\partial J}{\partial \mathbf{W}^{(2)}} = (\hat{\mathbf{y}} - \mathbf{y})\mathbf{h}^\top \quad (2.23)$$

We have everything to update our weights:

Now, turn definitely to skip-gram model:

Assume you are given a predicted word vector  $\mathbf{v}_c$  corresponding to the center word  $c$  for `skipgram`, and word prediction is made with the `softmax` function found in



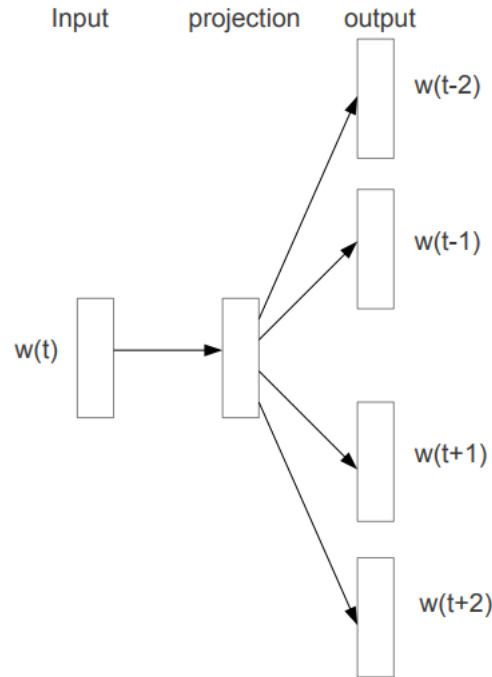


Figure 2.2 — The Skip-gram model architecture. The training objective is to learn word vector representations that are good at predicting the nearby words.

word2vec models

$$\hat{y}_o = p(\mathbf{o} | \mathbf{c}) = \frac{\exp(\mathbf{u}_o^\top \mathbf{v}_c)}{\sum_{j=1}^{|W|} \exp(\mathbf{u}_j^\top \mathbf{v}_c)} \quad (2.24)$$

where  $w$  denotes the  $w$ -th word and  $\mathbf{u}_w$  ( $w = 1, \dots, |W|$ ) are the ‘output’ word vectors for all words in the vocabulary ( $v'_w$  in the lecture notes). Assume the cross entropy cost is applied to this prediction and word  $o$  is the expected word (the  $o$ -th element of the one-hot label vector is one), derive the gradients with respect to  $\mathbf{v}_c$ .

**Hint:** It will be helpful to use notation from question 2. For instance, letting  $\hat{\mathbf{y}}$  be the vector of the softmax prediction for every word,  $\mathbf{y}$  as the expected word vector, and the loss function

$$J_{\text{softmax-CE}}(\mathbf{o}, \mathbf{v}_c, \mathbf{U}) = CE(\mathbf{y}, \hat{\mathbf{y}}) \quad (2.25)$$

where  $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{|W|}]$  is the matrix of all the output vectors. *Make sure you state the orientation of your vectors and matrices.*

---

Applying cross-entropy cost to the above softmax probability defined above:

$$J = -\log p = -\mathbf{u}_o^\top \mathbf{v}_c + \log \sum_{j=1}^{|V|} \exp(\mathbf{u}_j^\top \mathbf{v}_c) \quad (2.26)$$

Let  $z_j = \mathbf{u}_j^\top \mathbf{v}_c$ , and  $\delta_j^i$  be the indicator function (Kronecker delta), then

$$\frac{\partial J}{\partial z_k} = -\delta_k^i + \frac{\exp(\mathbf{u}_i^\top \mathbf{v}_c)}{\sum_{j=1}^{|V|} \exp(\mathbf{u}_j^\top \mathbf{v}_c)} \quad (2.27)$$

Now, using the chain rule, we can calculate,

$$\frac{\partial J}{\partial \mathbf{v}_c} = \frac{\partial J}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{v}_c} \quad (2.28)$$

$$= \sum_{j=1}^{|V|} \mathbf{u}_j^\top \left( \frac{e^{z_j}}{\sum_{k=1}^{|V|} e^{z_k}} - 1 \right) \quad (2.29)$$

$$= \sum_{k=1}^{|V|} \mathbf{P}(\mathbf{u}_j | \mathbf{v}_c) \mathbf{u}_j - \mathbf{u}_j \quad (2.30)$$

As in the previous problem, derive gradients for the ‘output’ word vectors  $\mathbf{u}_w$ ’s (including  $\mathbf{u}_o$ )

---

This follows immediately from the chain rule:

$$\frac{\partial J}{\partial \mathbf{u}_j} = \frac{\partial J}{\partial \mathbf{z}} \frac{\partial \mathbf{z}}{\partial \mathbf{u}_j} \quad (2.31)$$

$$= \mathbf{v}_c \left( \frac{\exp(\mathbf{u}_0^\top \mathbf{v}_c)}{\sum_{j=1}^{|V|} \exp(\mathbf{u}_j^\top \mathbf{v}_c)} - \delta_j^0 \right) \quad (2.32)$$

## Глава 3. Вёрстка таблиц

### 3.1 Таблица обыкновенная

## Заключение

Основные результаты работы заключаются в следующем.

1. На основе анализа ...
2. Численные исследования показали, что ...
3. Математическое моделирование показало ...
4. Для выполнения поставленных задач был создан ...

И какая-нибудь заключающая фраза.

## Список литературы

1. Jurafsky D. Speech and Language Processing / D. Jurafsky, M. James H. – New Jersey, 2008. – 1031 с. – (Pearson). – (ISBN: 0131873210). 2
2. Manning C. An Introduction to Information Retrieval / C. Manning, P. Raghavan, H. Schütze. – Cambridge, England: Online edition, 2009. – 544 p. – (Cambridge University Press). (ISBN: 0521865719). 3
3. Mohri M. Foundations of Machine Learning / M. Mohri, A. Rostamizadeh, A. Talwalkar. – Cambridge, Massachusetts, 2012. – 412 с. – (The MIT Press). – (ISBN: 9780262018258). 4
4. Géron A. Hands-On Machine Learning with Scikit- Learn and TensorFlow / Aurélien Géron. – Gravenstein Highway North, Sebastopol, CA, 2017. – 751 с. – (O'Reilly Media). – (ISBN: 8009989938). 5
5. Yang, Yiming, and Xin Liu. 1999. A re-examination of text categorization methods. In Proc. SIGIR, pp. 42–49. ACM Press. 2
6. McCallum, Andrew, and Kamal Nigam. 1998. A comparison of event models for Naive Bayes text classification. In AAAI/ICML Workshop on Learning for Text Categorization, pp. 41–48.
7. Rennie, Jason D., Lawrence Shih, Jaime Teevan, and David R. Karger. 2003. Tackling the poor assumptions of naive Bayes text classifiers. In Proc. ICML, pp. 616–623.
8. Tong, Simon, and Daphne Koller. 2001. Support vector machine active learning with applications to text classification. JMLR 2:45–66
9. Zavrel, Jakub, Peter Berck, and Willem Lavrijssen. 2000. Information extraction by text classification: Corpus mining for features. In Workshop Information Extraction Meets Corpus Linguistics. URL: [www.cnts.ua.ac.be/Publications/2000/ZBL00](http://www.cnts.ua.ac.be/Publications/2000/ZBL00). Held in conjunction with LREC-2000.

## Список рисунков

1.1	Classes, training set, and test set in text classification. . . . .	7
1.2	Supervised learning work flow. . . . .	8
1.3	Naive Bayes algorithm (multinomial model): Training and testing. . .	10
2.1	Neural Network . . . . .	14
2.2	The Skip-gram model architecture. The training objective is to learn word vector representations that are good at predicting the nearby words. . . . .	17

## Список таблиц

2.1	Feature vector . . . . .	12
Б.2	Наименование таблицы средней длины . . . . .	32

## Приложение А

### Примеры вставки листингов программного кода

Для крупных листингов есть два способа. Первый красивый, но в нём могут быть проблемы с поддержкой кириллицы (у вас может встречаться в комментариях и печатаемых сообщениях), он представлен на листинге [А.1](#).

Листинг А.1

Программа “Hello, world” на C++

```

5  #include <iostream>
    using namespace std;

    int main() //кириллица в комментариях при xelatex и lualatex и
        мееет проблемы с пробелами
    {
        cout << "Hello, world" << endl; //latin letters in
            commentaries
        system("pause");
        return 0;
10 }

```

Второй не такой красивый, но без ограничений (см. листинг [А.2](#)).

Листинг А.2

Программа “Hello, world” без подсветки

```

#include <iostream>
using namespace std;

int main() //кириллица в комментариях
{
    cout << "Привет, мир" << endl;
}

```

Можно использовать первый для вставки небольших фрагментов внутри текста, а второй для вставки полного кода в приложении, если таковое имеется.



Если нужно вставить совсем короткий пример кода (одна или две строки), то выделение линейками и нумерация может смотреться чересчур громоздко. В таких случаях можно использовать окружения `lstlisting` или `Verb` без `ListingEnv`. Приведём такой пример с указанием языка программирования, отличного от заданного по умолчанию:

```
| fibs = 0 : 1 : zipWith (+) fibs (tail fibs)
```

Такое решение — со вставкой нумерованных листингов покрупнее и вставок без выделения для маленьких фрагментов — выбрано, например, в книге Эндрю Таненбаума и Тодда Остина по архитектуре

Наконец, для оформления идентификаторов внутри строк (функция `main` и тому подобное) используется `lstinline` или, самое простое, моноширинный текст (`\texttt`).

Пример A.3, иллюстрирующий подключение переопределённого языка. Может быть полезным, если подсветка кода работает криво. Без дополнительного окружения, с подписью и ссылкой, реализованной встроенным средством.

Листинг A.3

Пример листинга с подписью собственными средствами

```
## Caching the Inverse of a Matrix

## Matrix inversion is usually a costly computation and there
  may be some
5 ## benefit to caching the inverse of a matrix rather than
  compute it repeatedly
## This is a pair of functions that cache the inverse of a
  matrix.

## makeCacheMatrix creates a special "matrix" object that can
  cache its inverse

10 makeCacheMatrix <- function(x = matrix()) {#кириллица в коммента
  риях при xelatex b luaLatex имеет проблемы с пробелами
    i <- NULL
    set <- function(y) {
      x <- y
      i <- NULL
15  }
    get <- function() x
    setSolved <- function(solve) i <- solve
```

```

    getSolved <- function() i
    list(set = set, get = get,
20   setSolved = setSolved,
    getSolved = getSolved)
  }

25
  ## cacheSolve computes the inverse of the special "matrix"
    returned by
  ## makeCacheMatrix above. If the inverse has already been
    calculated (and the
  ## matrix has not changed), then the cachesolve should retrieve
    the inverse from
  ## the cache.

30
cacheSolve <- function(x, ...) {
  ## Return a matrix that is the inverse of 'x'
  i <- x$getSolved()
  if(!is.null(i)) {
35     message("getting cached data")
    return(i)
  }
  data <- x$get()
  i <- solve(data, ...)
40  x$setSolved(i)
  i
}

```

Листинг A.4 подгружается из внешнего файла. Приходится загружать без окружения дополнительного. Иначе по страницам не переносится.

Листинг A.4

Листинг из внешнего файла

```

# Analysis of data on Course Project at Getting and Cleaning
  data course of Data Science track at Coursera.

# Part 1. Merges the training and the test sets to create one
  data set.
5 # 3. Uses descriptive activity names to name the activities in
  the data set

```

```

# 4. Appropriately labels the data set with descriptive variable
      names.

if (!file.exists("UCI HAR Dataset")) {
  stop("You need 'UCI HAR Dataset' folder full of data")
10 }

library(plyr) # for mapvalues

15
#getting common data
features <- read.csv("UCI HAR Dataset/features.txt", sep=" ",
  header = FALSE,
                        colClasses = c("numeric", "character"))
activity_labels <- read.csv("UCI HAR Dataset/activity_labels.txt
  ", sep="",
20
                        header = FALSE, colClasses = c("
                            numeric", "character"))

#getting train set data
subject_train <- read.csv("UCI HAR Dataset/train/subject_train.
  txt",
                        header = FALSE, colClasses = "numeric",
                        col.names="Subject")
25 y_train <- read.csv("UCI HAR Dataset/train/y_train.txt", header
  = FALSE,
                        colClasses = "numeric")
x_train <- read.csv("UCI HAR Dataset/train/X_train.txt", sep="",
  header = FALSE,
                        colClasses = "numeric", col.names=features$V2
                        , check.names = FALSE)

30 activity_train <- as.data.frame(mapvalues(y_train$V1, from =
  activity_labels$V1,
                        to = activity_labels$
                        V2))

names(activity_train) <- "Activity"
35

```

```

#getting test set data
subject_test <- read.csv("UCI HAR Dataset/test/subject_test.txt"
,
                        header = FALSE,colClasses = "numeric",
                        col.names="Subject")
y_test <- read.csv("UCI HAR Dataset/test/y_test.txt", header =
FALSE,
40      colClasses = "numeric")
x_test <- read.csv("UCI HAR Dataset/test/X_test.txt",sep="",
header = FALSE,
                        colClasses = "numeric",col.names=features$V2,
                        check.names = FALSE)

activity_test <- as.data.frame(mapvalues(y_test$V1, from =
45      activity_labels$V1,
                                to = activity_labels$V2
                                ))
names(activity_test) <- "Activity"

# Forming full dataframe
50 data_train <- cbind(x_train,subject_train,activity_train)
data_test <- cbind(x_test,subject_test,activity_test)
data <- rbind(data_train, data_test)

# Cleaning memory
55 rm(features, activity_labels, subject_train, y_train, x_train,
    activity_train,
    subject_test, y_test, x_test, activity_test, data_train, data
    _test)

# Part 2. Extracts only the measurements on the mean and
    standard deviation for each measurement.
60
cols2match <- grep("(mean|std)",names(data))

# Excluded gravityMean, tBodyAccMean, tBodyAccJerkMean,
    tBodyGyroMean,
# tBodyGyroJerkMean, as these represent derivations of angle
    data, as

```

```

65 | # opposed to the original feature vector.

    # Subsetting data frame, also moving last columns to be first
    Subsetted_data_frame <- data[,c(562, 563, cols2match)]

70 | # Part 5. From the data set in step 4, creates a second,
    independent tidy data set
    # with the average of each variable for each activity and each
    subject.

    library(dplyr) # for %>% and summarise_each

75 | tidydata <- Subsetted_data_frame %>% group_by(Subject,Activity)
    %>%
        summarise_each(funs(mean))

    write.table(tidydata, "tidydata.txt", row.names=FALSE)

```

## Приложение Б

Очень длинное название второго приложения, в котором продемонстрирована работа с длинными таблицами

### Б.1 Подраздел приложения

Вот размещается длинная таблица:

Параметр	Умолч.	Тип	Описание
&INP			
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора

продолжение следует

[illegible]

(продолжение)			
Параметр	Умолч.	Тип	Описание
			экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ )
			1: генерация белого шума
			2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ )
			1: генерация белого шума
			2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ )
			1: генерация белого шума
			2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ )
			1: генерация белого шума
			2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс

## Б.2 Ещё один подраздел приложения

Нужно больше подразделов приложения!

Пример длинной таблицы с записью продолжения по ГОСТ 2.105

Table Б.2

Наименование таблицы средней длины

Параметр	Умолч.	Тип	Описание
&INP			
kick	1	int	0: инициализация без шума ( $p_s = const$ )
			1: генерация белого шума
			2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ )
			1: генерация белого шума



Продолжение таблицы Б.2

Параметр	Умолч.	Тип	Описание
			2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ )
			1: генерация белого шума
			2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ )
			1: генерация белого шума
			2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ )
			1: генерация белого шума
			2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ )
			1: генерация белого шума
			2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ )
			1: генерация белого шума
			2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ )
			1: генерация белого шума
			2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ )

Продолжение таблицы Б.2

Параметр	Умолч.	Тип	Описание
			1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
&SURFPAR			
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора

## Продолжение таблицы Б.2

Параметр	Умолч.	Тип	Описание
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс
kick	1	int	0: инициализация без шума ( $p_s = const$ ) 1: генерация белого шума 2: генерация белого шума симметрично относительно экватора
mars	0	int	1: инициализация модели для планеты Марс

### **Б.3 Очередной подраздел приложения**

Нужно больше подразделов приложения!

### **Б.4 И ещё один подраздел приложения**

Нужно больше подразделов приложения!