Team «123»

|  |
| --- |
| 14.11.2016 |

Digit recognition

Business Understanding

Contents

[1 Document summary 2](#_Toc466424581)

[1.1 Document purpose 2](#_Toc466424582)

[1.2 Document scope 2](#_Toc466424583)

[1.3 Versions 2](#_Toc466424584)

[1.4 Revisions 2](#_Toc466424585)

[2 Business Objectives Definition 3](#_Toc466424586)

[2.1 Background of project 3](#_Toc466424587)

[2.1.1 Organizational Structure 3](#_Toc466424588)

[2.1.2 Problem Area Description 3](#_Toc466424589)

[2.1.3 Current Solution Description 4](#_Toc466424590)

[2.2 Success criteria 4](#_Toc466424591)

[2.3 Situation assessment 4](#_Toc466424592)

[2.3.1 Hardware resources 4](#_Toc466424593)

[2.3.2 Data sources description 4](#_Toc466424594)

[2.4 Requirements, Assumptions, and Constraints 5](#_Toc466424595)

[2.4.1 Requirements 5](#_Toc466424596)

[2.4.2 Assumptions 5](#_Toc466424597)

[2.4.3 Constraints 5](#_Toc466424598)

[2.5 Risks and contingencies 5](#_Toc466424599)

[2.6 Terminology 5](#_Toc466424600)

[3 Data mining goals 5](#_Toc466424601)

[3.1 Data mining goals 5](#_Toc466424602)

[3.2 Data Mining Success Criteria 6](#_Toc466424603)

Document summary

## Document purpose

Document is intended to reflect the main points of the Digit Recognition project and the results of the considered problem. The problem of Digit Recognition is solved using data mining tools.

Document contains some information about Business Objectives Definitionand Data mining goals. The main purpose of this project is getting important basic skills in working with tools of the language R, with referring to the information about different machine learning methods for handwritten digits recognition. Moreover, the goal was set to realize the basic methods and to compare an accuracy of these methods.

## Document scope

Document describes requirements and workaround of the study project Data Mining Tools 2016.Results of the documenting process are to be discussed at the lessons.

## Versions

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Version | Editor | Description of version |
| 2016-11-07 | 0.1 | E.N.Pavlovskiy | Document template |
| 2016-11-09 | 1.0 | A.Kozhukhar, E.Barakhovskaia, A.Galaktionova | First revision |
| 2016-11-14 | 1.1 | A.Kozhukhar, E.Barakhovskaia, A.Galaktionova | Correction of the report |

## Revisions

|  |  |  |  |
| --- | --- | --- | --- |
| Date | Version | Approved by | Decision |
| 2016-11-09 | 1.0 | E.N.Pavlovskiy | Correct the document:1. Correct success criteria (BU and DM)2. Add risks3. Change Requirements4. Change data mining goals 5. Add project plan |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Business Objectives Definition

## Background of project

### Organizational Structure

The project group «123» consists of three members (one manager/documentation developer, two developers/problem analizers).

### Problem Area Description

Annually a lot of high-school graduates in Russia pass the exam called USE (Unified State Examination). To pass the exam graduates must fill out a special blank which is processed by the computer, **fig. 1**. Nevertheless nowadays there still is the problem of unreliable machine check, so checking isoften duplicated by person. It is very important to improve the quality of machine checking. The important improvements are:

* reducing the cost and processing time of examination forms
* increasing the accuracy of processing the results
* minimizing manual inspection after machining one.

Thereby the number of errors caused by human factor (tiredness, inattention, negligence) will be reduced. Moreover, the highly relevant in nowadays factor, such as possibility of prejudgment will be eliminated.



**Figure 1**.Different examples of exam blank of the USE in Russia.

### Current Solution Description

At present time checking forms is carried out in two stages. The first step is machine recognition, the second – human checking by eyes for reliability. So, checking time is increased by several stages checking, that makes all process very labor-intensive. Taking into account a huge quantity of information, this main problem creates a tangible cost to the exam organizers.

## Success criteria

It is important to clearly define the criteria for business success for our data mining project before proceeding further.

**Objective**. If number of appeals, which connect with machine-checking part of blank, decreases and after adaptation of method for blank checking is accelerated, the project will be considered as successful.

**Subjective**. Increasing of checking accuracy will make procedure of the exams more comfortable for high-school graduates and improve their mental health.

If there is no success in business part of the project then the organization of USE will break the contract with us. Our company will have to pay a penalty, which is written in contract. Moreover, the company will lose authority and prestige.

## Situation assessment

**Data**. The data files train.csv and test.csv contain gray-scale images of hand-drawn digits, from zero through nine.

Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255, inclusive.

The training data set, (train.csv), has 785 columns. The first column, called "label", is the digit that was drawn by the user. The rest of the columns contain the pixel-values of the associated image. Each pixel column in the training set has a name like pixels, where x is an integer between 0 and 783, inclusive. To locate this pixel on the image, suppose that we have decomposed x as x = i \* 28 + j, where i and j are integers between 0 and 27, inclusive. Then pixels is located on row i and column j of a 28 x 28 matrix, (indexing by zero).

### Hardware resources

To support Digit recognition project our team need only one laptop at list with 4 Gb of RAM, free storage – 1Gb.

### Data sources description

The data for this project was taken from the MNIST dataset. The MNIST ("Modified National Institute of Standards and Technology") dataset is a classic within the Machine Learning community that has been extensively studied. Set of data is located in website [www.kaggle.com](http://www.kaggle.com).

The data stored such way: Each image is 28 pixels in height and 28 pixels in width, for a total of 784 pixels in total. Each pixel has a single pixel-value associated with it, indicating the lightness or darkness of that pixel, with higher numbers meaning darker. This pixel-value is an integer between 0 and 255, inclusive.The training data set, (train.csv), has 785 columns. The first column, called "label", is the digit that was drawn by the user. The rest of the columns contain the pixel-values of the associated image.

There are no security issues preventing access to required data.The training data set is free. In the real problem data is not free, but commonality of writing of Arabic numerals provides learning of the model. Model recognizes generally accepted writing of Arabic numerals.

There are two files train.csv (73.22 mb) and test.csv(48.75 mb).

## Requirements, Assumptions, and Constraints

### Requirements

The creation of team – 2016-10-26.

The formulation the problem statement and starting a development of the project - October-November 2016.

Project modeling using program language R, receiving and outputting the results in the required format – 2016-11-4.

Submit results with www.kaggle.com and analyzing, making conclusions – 2016-11-7.

Preparing the report 2016-11-14.

### Assumptions

Input data is black and white image, which contain hand-written Arabic numeral with correct orientation.

### Constraints

Data of USE is encrypted, because privacy is very important in this problem. Access to this data is limited. Dato for building model and testing is free access, for example data is available on www.kaggle.com.

## Risks and contingencies

Types of risks include:

* Scheduling. - if project requires more time we have planned, we will lose prestige and clients, the project will be totally failed. One way to avoid schedule-problem is competent time organization of the team work by manager, adding a control over the work and writing regular reports to the boss (time-manager).
* Financial - if the project sponsor encounters budgetary problems, we will construct simpler project, which will be able to decrease the cost of working.
* Data – if the data for building model are of poor quality or coverage, we will find other source of the data.
* Results – if the initial results are less dramatic than expected, we will try other methods.

Contingencies:

In every project there are contingencies with high probability. For example, cancellations of the project for some reasons, USE format changes, unexpected payments for data, testing results, etc.

Nobody is saved. Despite a variety of unexpected occasions, our company always has alternative ways to suggest a solution to the problem and to complete the project with high quality.

## Terminology

**Accuracy** – is defined by formula$A=G/T∙100\%,$ where A – accuracy, G – number of guessed digits out of all test digits, T – number of all digits.

**k-Nearest Neighbors algorithm (k-NN**) is a non-parametric method used for classification and regression.

**Support vector machines (SVMs, also support vector networks)** are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.

**Cross-validation** is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set

**Machine learning** is the subfield of computer science that "gives computers the ability to learn without being explicitly programmed”

**Principal component analysis (PCA)** is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

Data mining goals

## Data mining goals

The main purpose of this project is to get important basic skills in working with tools of the language R, with referring to the information about different machine learning methods for handwritten digits recognition. Moreover, the goal was set to realize the basic methods and to compare an accuracy of these methods.

## Data Mining Success Criteria

The success criteria is that accuracy should be more than 90%. This value of accuracy has been chosen by strictly requirements of customer. 90% of accuracy is a value of a high probability of correct blank checking, in the same time such result really can be achieved in data mining analysis.

Project plan

|  |  |
| --- | --- |
| Problem | **Due date** |
| Business understanding | 08.11.2016 |
| Data understanding | 11.11.2016 |
| Data Preparation | 11.11.2016 |
| Modeling | 11.11.2016 |
| Evaluation | 12.11.2016 |

Project outcome

## Calculating algorithm

1. Preparing of the data for creating a model

In the available samples were removed features which had dispersion of the values closed to zero. This step was carried out using *nearZeroVar* function, which is part of the package *caret.*

1. . Application of principal component analysis using function *princomp*.

Principal component analysis let us to decrease a dimension of data without losing lots of information.

1. Parameters searching

Parameters of learning procedure were found with function *trainControl*. Two methods were used, see table 1.

1. Model developing based on transformed data.

Predicted model was created using two methods:

* **k-Nearest Neighbors algorithm (k-NN**).
* **Support vector machines (SVM)**.

For realization of these methods in language R is special functions: **knn** fork-Nearest Neighbors algorithm and **ksvm** for support vector machines.

## 5.2 Testing results

The calculating results are given in table 1.

*Table 1*. Accuracy of the results for different models

|  |  |
| --- | --- |
| **Model** | **Results on www.kaggle.com** |
| knn | 65% of training set | 10-fold cross-validation | 87.914 % |
| knn | 100% of training set | 10-fold cross-validation | 88.514 % |
| knn | 100% of training set | adaptive cross-validation | 88.386 % |
| ksvm | 100% of training set | adaptive cross-validation | 97.786 % |

## **5.3** Analysis of obtained results

After analyzing, there can be made some conclusions:

1. 100% of training set gives more correct result than 65%, but accuracy of result increases only by hundredth parts.
2. 10-fold cross-validation gives higher accuracy than adaptive cross-validation. However, this improvement is insignificantly, only thousandth parts of accuracy.
3. Calculations shows that support vector machines method with **ksvm** is significantly better than k-Nearest Neighbors algorithm. support vector machines method with adaptive cross-validation has high accuracy97.786 % (results from www.kaggle.com)

Conclusion

There are a range of methods for realization tasks of data visualization. Methods have the differences in such important parameters for calculation as performance, accuracy, efficiency and difficulty.

Two methods were tested. K-nearest neighbors algorithm and support vector machine method were chosen for researching and realization on language R, because both methods are widespread not only in data visualization field, but in many other fields. Moreover, both methods can be called as basic. That let investigate the main points of algorithm.

Basing on the considered problem, it is possible to make the *main conclusion*: there may be different approaches to solving this type of problems, but we have to remember that each approach has distinctive properties that need to be taken into account to obtain the best results.

Our team “123” achieved 461 place on public leaderboard on [www.kaggle.com](http://www.kaggle.com) on 2016.11.14.