CS 6375.502

Group Project: Digit Recognizer

Project Idea: Active Kaggle competition

Team Members:

Jiaxi Zhao

Yuchao Hu

Yijia Yu

**Introduction**

Nowadays, image recognition is still a significant part of machine learning. Of course, digit recognition that is a track of image recognition also have critically effect on daily life. For example, digit recognition is indispensable algorithm when camera correctly identify digits in the gift card.

In this situation, camera must precisely read the digits in the gift card. In our group Project, we totally selected three methods to solve this problem, which are naive Bayes classifier, SVM and XG boost. The naive Bayes is basic approach to address this problem.

In summary, our result shows accuracy rate of naive Bayes classifier is lowest and other two methods holds higher accuracy rate than naive Bayes classifier. This phenomenon is reasonable because naive Bayes classifier is available only if certain assumptions are satisfied.

**Problem Definition and Algorithm**

This Project focus on exactly recognize digits from a dataset of handwritten images that is from MNIST (Modified National Institute of Standards and Technology). In dataset, all images are 28 pixels in height and 28 pixels in width, which are totally 784 pixels for each handwritten image. Generally, each pixel has an integer value, higher value means darker and lower value means lighter. Even though this pixel-value is number between 0 and 255, we just need to know each pixel-value is 0 or otherwise. Because we don’t need to know what color is for each pixel when we do digit recognition.

For input dataset, input dataset has 785 columns that include one label and 784 pixels. The label is the digit that was handwritten by user. Each row is a handwritten image from user. According to above information, training dataset should look like Table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| label | Pixel0 | Pixel1 | Pixel2 | etc |
| 8 | 58 | 177 | 253 | etc |
| 5 | 0 | 122 | 0 | etc |

Table 1

For output dataset, output dataset has only 2 columns that include one Image ID and one Predicted label. Image ID should match with test dataset and Predicted label is digit we predict. According to above information, output dataset should look like Table 2.

|  |  |
| --- | --- |
| ImageID | Lable |
| 1 | 8 |
| 2 | 5 |

Table 2

Table 3 shows dataset details.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data size | Training samples | Test samples | Attributes | Label | Data distribution |
| 15MB (compressed) | 42000\*785 | 28000\*784 | Gray value of 784 pixels | Digit 0 to 9 | Even distribution |

Table 3

In real life, most paper work is still dealing with handwritten rather than electric form or paperless. However, we need to transfer paper work into electric form in certain situation. Specially, digit recognition is widely applied. For example, we save or withdraw money using paper check. Not only ATM must accurately read digits in the check but also ATM distinguish date and payment amount. Above process require high accuracy rate. In conclusion, digit recognition plays a significant role in our daily life.

Algorithm Definition

[Describe in reasonable detail the algorithm you are using to address this problem. A psuedocode description of the algorithm you are using is frequently useful. Trace through a concrete example, showing how your algorithm processes this example. The example should be complex enough to illustrate all of the important aspects of the problem but simple enough to be easily understood. If possible, an intuitively meaningful example is better than one with meaningless symbols.]

**Experimental Evaluation**

[3.1 Methodology

What are criteria you are using to evaluate your method? What specific hypotheses does your experiment test? Describe the experimental methodology that you used. What are the dependent and independent variables? What is the training/test data that was used, and why is it realistic or interesting? Exactly what performance data did you collect and how are you presenting and analyzing it? Comparisons to competing methods that address the same problem are particularly useful.   
  
3.2 Results

Present the quantitative results of your experiments. Graphical data presentation such as graphs and histograms are frequently better than tables. What are the basic differences revealed in the data. Are they statistically significant?   
  
3.3 Discussion

Is your hypothesis supported? What conclusions do the results support about the strengths and weaknesses of your method compared to other methods? How can the results be explained in terms of the underlying properties of the algorithm and/or the data.]

**Future Work**

According to above information, there are three major shortcomings of our methods.

First, we must calculate conditional possibility under independent assumption between the features otherwise we can’t do classification based on naive Bayes’ theorem. However, independent assumption is the key assumption. We can’t enhance or improve this shortcoming otherwise we will ruin statistics theorem.

Secondly, we did preprocess for dataset and this preprocess is one kind of simplify. In fact, this action normally has effect on dataset. Furthermore, original pixel-value is between 0 and 255 and adjusted pixel-value should be 0 or 1. Adjustment probably reduce effect of outliers, which lead up to prediction unfitted for original dataset. Improvement relatively is simply, we just don’t preprocess to original pixel-value. Then classification will keep outliers and prediction is fitted to original dataset.

Thirdly, XG boost and SVM methods have higher accuracy rate than naive Bayes classification. According to model complexity, this result is reasonable because the complex model usually has high accuracy and overfitting problem. To avoid overfitting, we should simplify model and add penalty term to model.

**Conclusion**

[Briefly summarize the important results and conclusions presented in the paper. What are the most important points illustrated by your work? How will your results improve future research and applications in the area?]