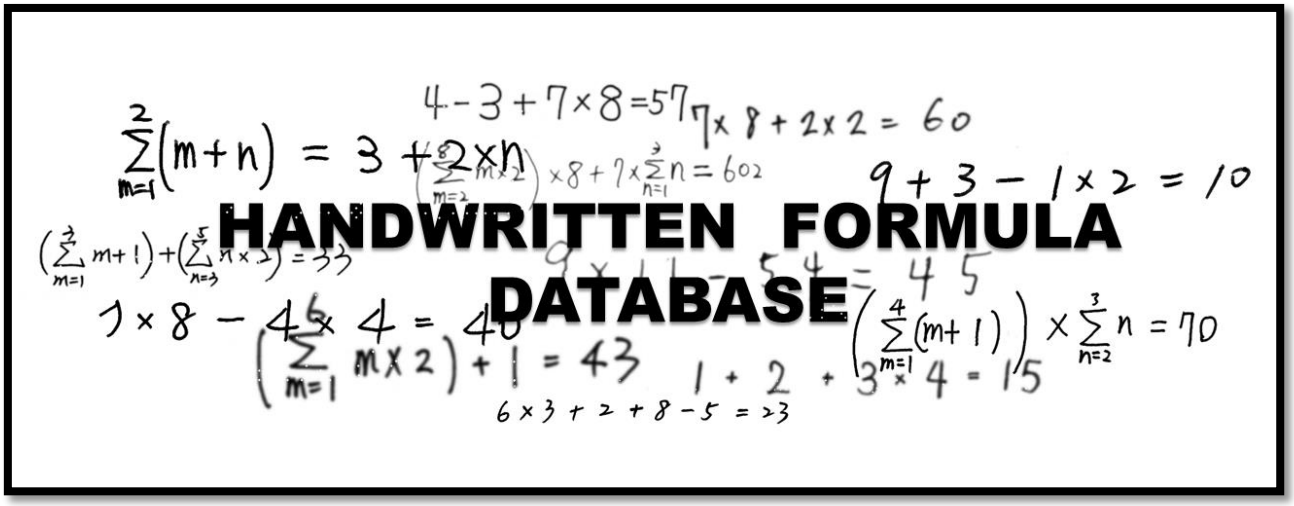
Machine Learning 2016

Final Project：Handwritten Formula Competition

Check points: **2016.06.07**
Presentation: **2016.06.22-06.24**

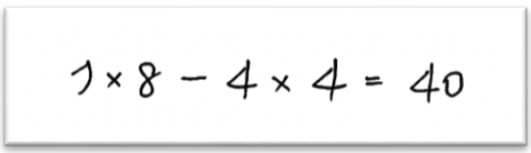
You have received the invitation to the **Handwritten Formula Competition (HFC)**, which is the topic of your final project. The content of HFC will be explained later. Please help yourself find methods to deal with the issues in the progress instead of asking what to do next. Part of your score will be decided directly from the performance of the competition, so do not underestimate it!



◆ Handwritten Formula Competition

The main goal of this challenge is to recognize handwritten formulas and compute the answers. It is fundamentally a supervised learning problem in that a training set of labelled images is provided. The difficulty of recognizing formulas can be regarded as five levels:

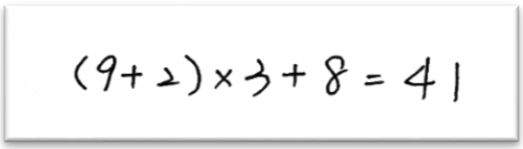
- **Level 1**
  Notations in use：0, 1, 2, 3, 4, 5, 6, 7, 8, 9, +, −, ×, =
  Example：



- **Level 2**

  Notations in use：$0, 1, 2, 3, 4, 5, 6, 7, 8, 9, +, -, \times, =, (,)$

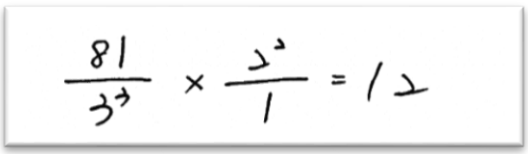
  Example：



- **Level 3**

  Notations in use：$0, 1, 2, 3, 4, 5, 6, 7, 8, 9, +, -, \times, =, (,)$

  (Annotation：Fractions and indices start to appear)



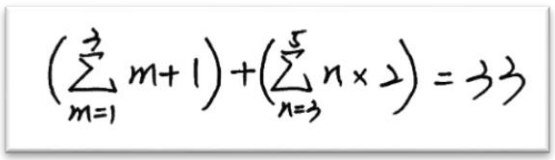
  Example：

- **Level 4**

  Notations in use：$0, 1, 2, 3, 4, 5, 6, 7, 8, 9, +, -, \times, =, (,), m, n, \sum$

  (Annotation：m and n are algebras)

  Example：

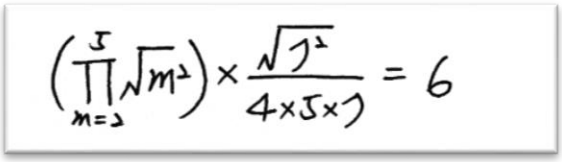


- **Level 5**

  Notations in use：$0, 1, 2, 3, 4, 5, 6, 7, 8, 9, +, -, \times, =, (,), m, n, \sum, \prod, \sqrt{}$

  (Annotation：m and n are algebras)

  Example：



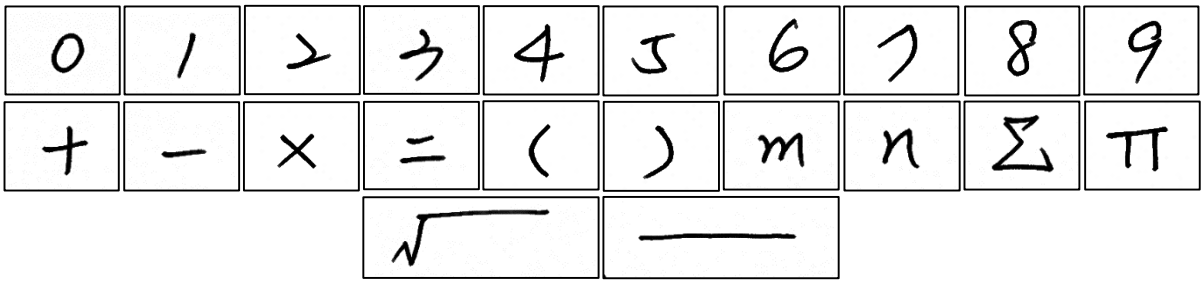
There will be 150 formulas for validation and 250 for test. The performance of the algorithm will be evaluated by its test accuracy.

◆ Handwritten Formula Database

Handwritten Formula Database is a handwritten image database which contains 22 classes of numbers and mathematical notations and hundreds of handwritten formulas.

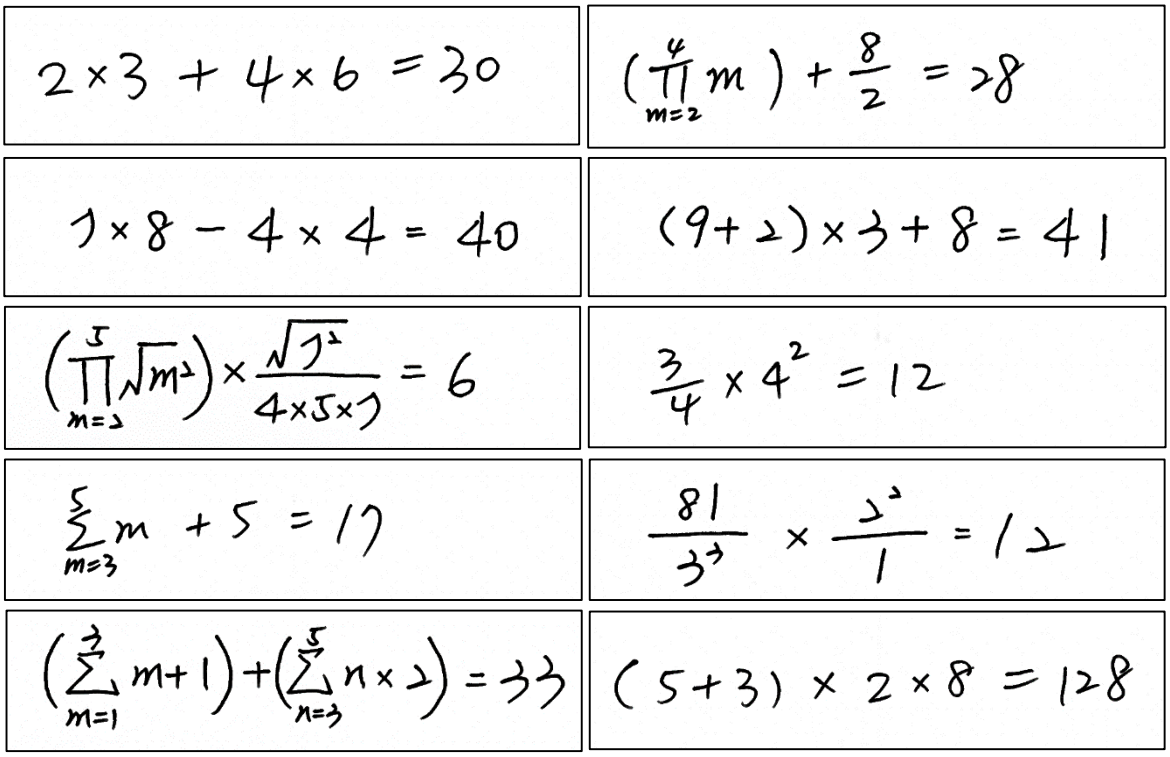
● Handwritten mathematical notations

These mathematical notations are basic elements of the handwritten formulas. Once they can be recognized well, the whole sentence won't be difficult to understand.



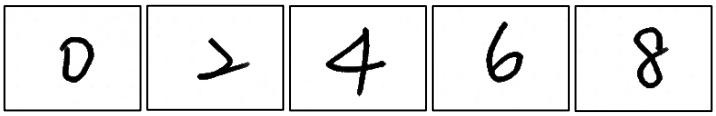
● Handwritten formulas

Like we've introduced before, these formulas have five level of difficulty to recognize. Each formula has a target value, which equals to the answer of the formula.

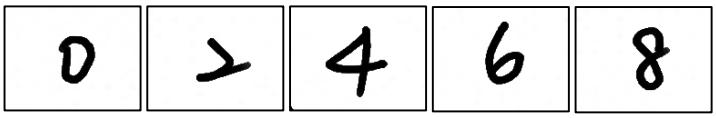


There are four kinds of the training data and test data. Please choose the one you prefer:

- White and normal digits



- White and bold digits

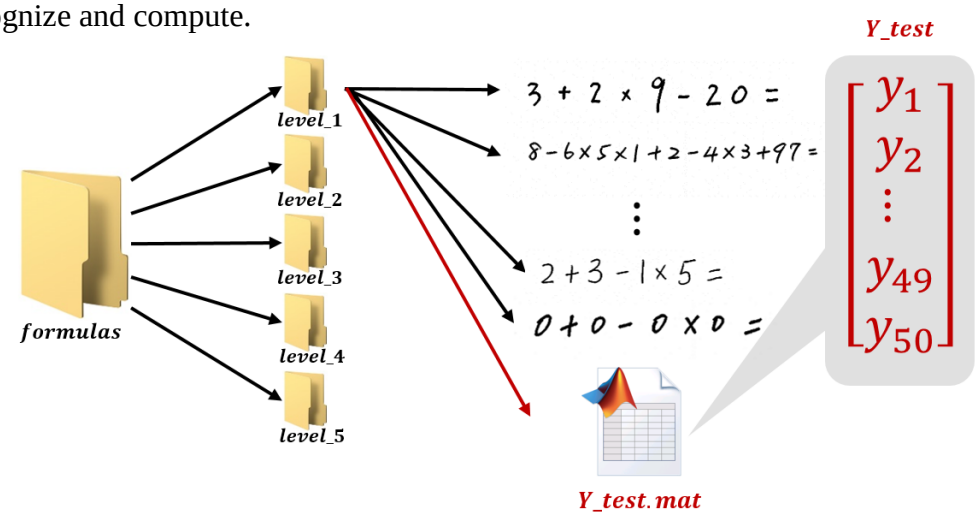


- Black and normal digits

- Black and bold digits

◆ Tasks

Given a set of handwritten formula images, please write an algorithm to recognize these images and try to compute the **correct answers** which equals to the image labels. For the demonstration, you'll receive a folder named "formulas" which contains five sub-folders "level_1" to "level_5". Each sub-folder consists of 50 handwritten formula images needed to be recognize and compute.



Please save your answer **Y_test** into the **Y_test.mat** file in every sub-folder.

◆ Checkpoint

You will need to hand in a 3-page proposal which explains your model structure and how you are going to realize the ideas. After reading it, TAs will give you some advices. Please don't forget that this proposal accounts for about 3-5% of your final score.

What should be uploaded for the checkpoint?

- ☐ A 3-page **proposal** for explaining your ideas.

What should be uploaded for the presentation?

- ☐ A runnable **DEMO code** for test data.
- ☐ A 10-page **slide** for presentation.
- ☐ Your source code with comments.
- ☐ The **ReadMe.txt** file which describes how to run your program.
- ☐ Your **report** in the format of .pdf or .doc.

Reminders:

- ☐ **One or Two members** a group for this homework. Hand one report each group.
- ☐ Please run the programs on your laptop for the demonstration.
- ☐ Each group needs to make a **15-minute presentation** to the professor and TAs.
- ☐ **DO NOT STEAL ANY IDEAS WITHOUT BEING AUTHORIZED!!!**