

# CSCI933: Machine Learning - Algorithms and Applications

## Assignment #3

Due date: 23:59 (Beijing Time), 19<sup>th</sup> May, 2020

### 1 Introduction

The goal of this assignment is to give students an opportunity to work in a group of **3 or 4**, and collaboratively explore and report on an interesting topic in machine learning that has not been discussed in the lectures. The group of 3 or 4 has been specified because of the size of the topic and the length of the report expected from the group. A well-written 10-page term paper is the expected outcome of this assignment.

### 2 Semi-supervised Learning

In real life applications it is possible that there is insufficient labelled data to successfully train a “supervised learner”. However, usually there is large amount of unlabeled data. This scenario gives rise to the semi-supervised learning paradigm in which both labelled and unlabeled data are used in the design of the learner.

Consider the following problem. As the world population increases the problem of food security becomes a matter of significant concern for many nations in the world. A large quantity of food crops are lost to diseases. This is especially the case with smallholder farmers<sup>1</sup> who may not have access to knowledge about the signs/symptoms of various food crop diseases. It is also known that the onset of most of these diseases can be recognized from the changes to the leaves of the plant. What if these farmers can be provided with smart phone-based tools that can recognize the diseases based on the images captured from the phone cameras? It may be possible for the farmers to report the onset of the disease to expert plant scientists or agricultural extension officers before the disease took a foothold on the farm or spread to surrounding farms.

In this problem, there are usually insufficient images of the various diseased plant leaves to train a classifier/recognizer with sufficient accuracy for production deployment. This assignment requires you to study the semi-supervised learning paradigm by referring to the resources provided with this specification and gain some level of understanding. You may consult additional resources if required. Based on your understanding of the technique you are required to design an appropriate semi-supervised image-based classifier/recognizer for food crop diseases. **You are not required to implement your classifier/recognizer.**

---

<sup>1</sup> Smallholder farmers are those farmers owning small-based plots of land on which they grow subsistence crops and one or two cash crops relying almost exclusively on family labour.

### 3 Resources

Some papers and a book have been compiled as initial resource for this assignment. You are expected to find some other papers that may help you complete your paper. See the accompanying “zip” file.

### 4 Structure of term paper

The term paper you submit **must be no more than 12 pages and no less than 10 pages** in length (**excluding** references). This includes the writing, images, graphs and tables. You are allowed to use graphs, images, equations and tables from other publications as long as the source is properly acknowledged and there are segments of text in your writing that refer to, and describe them. Your whole paper must be properly referenced, i.e. in-text referencing. **Work without proper referencing amounts to plagiarism and will be awarded zero mark.** You must cite at least five (5) relevant references in your paper. Your paper must contain no less than ten (10) mathematical equations related to the topic of semi-supervised learning and surrounding concepts. You must describe the equations to demonstrate that you understand them. Each symbol must be defined/described.

Your paper should be in 11- or 12-point font and one-and-a-half line spacing, with 1.5 cm margin around all edges.

Use the following headings which also indicate the allocated marks in your paper:

#### 1. Heading/title - 5 marks

Give your paper a good title that is meaningful and indicative of the content. The names of the members of your group must appear as authors of the paper along with their student numbers and e-mail.

#### 2. Introduction - 10 marks

You should describe semi-supervised learning and its place in the machine learning “landscape”. Describe the various forms of semi-supervised learning and their theoretical basis. How does it work? What are its limitations? What are the applications or the kind of problem it can solve?

#### 3. Background theory - 30 marks

Provide a theoretical development that demonstrates your understanding of the machinery of semi-supervised learning. You may want to devote separate sub-section to each type of semi-supervised learning and how it is trained.

#### 4. Application (Food crop disease classification/recognition) - 30 marks

In this section you are to devise a solution to the problem of crop disease classification/recognition using one of the methods of semi-supervised learning methods you described in the previous section. Describe how your chosen method will be applied to this problem. Describe the method of preparing the data. You may need to search for a crop disease leaf image database on the WWW.

## 5. Conclusion - 10 marks

You must derive some conclusion based on the understanding you have gained from the materials you have studied. What do you see as the future direction of research in semi-supervised learning?

## 6. References - 15 marks

Your references must follow the IEEE style. If you do not know about this style, see the example below. The in-text referencing should follow the numbered citation style. Note the marks allocated to referencing. This is to encourage you to be diligent in your use of in-text referencing and accurate recording of the sources of your references.

# 5 What to submit

You are to submit your group term paper in PDF format through the Moodle site on or before the deadline. Only one submission per group is required. Specify the ID number, name, and the individual effort (in percentage) of your team members in the report.

## References

- [1] Q. Iqbal and J. K. Aggarwal, "Feature integration, multi-image queries and relevance feedback in image retrieval," in *Proceedings, 6th International Conference on Visual Information Systems(VISUAL 2003)*, Miami, Florida, Sep. 24 - 26 2003, pp. 467 – 474.
- [2] X.-Y. Jing, H.-S. Wong, and D. Zhang, "Face recognition based on Fisherface approach," *Pattern Recognition*, vol. 39, no. 4, pp. 707 – 710, April 2006.
- [3] A. Papoulis, *Probability, Random Variables and Stochastic Processes*, 3rd ed. New York: McGrawHill International Edition, 1991.
- [4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge Massachusetts: The MIT Press, 2016.