Capstone Project for Data Science **29 Oct, 2020**

Final Project

Instructor: Yuan Yao *Due: 23:59 Thursday 19 Nov, 2020*

1 Project Requirement

This project as a warm-up aims to explore basic techniques in machine learning.

1. Pick up ONE (or more if you like) favourite dataset below to work. If you would like to work on a different problem outside the candidates we proposed, please email course instructor about your proposal.

2. Team work: we encourage you to form small team, up to **THREE** persons per group, to work on the same problem. Each team just submit **ONE** report, *with a clear remark on each person's contribution*. The report can be in the format of either a *poster*, e.g.

   `https://github.com/yuany-pku/2017_math6380/blob/master/project1/DongLoXia_poster.pptx`

   or *technical report within 8 pages*, e.g. NIPS conference style (preferred format)

   `https://nips.cc/Conferences/2019/PaperInformation/StyleFiles`,

   with source codes such as Python (Jupyter) Notebooks with a detailed documentation.

3. For Kaggle contests, please register your team with name in the format of math4995_lastname, so that we could easily find your results on Kaggle. For example, a team with Shawn Zhu and Kate Wong would be named by math4995_Zhu_Wong.

4. In the report, show your proposed scientific questions to explore and main results with a careful analysis supporting the results toward answering your problems. If possible, you should include your Kaggle contest score or rating in the report. Remember: scientific analysis and reasoning are more important than merely the performance tables. Separate source codes may be submitted through email as a GitHub link, or a zip file.

5. Submit your report by email or paper version no later than the deadline, to the following address (wzhuai@connect.ust.hk) with a title "MATH4995: Project 2"

2 Kaggle in-class Contest: Nexperia Image Classification II

Nexperia (`https://www.nexperia.com/`) is one of the biggest Semi-conductor company in the world. They produce billions of semi-conductors every year. Meanwhile, a lot of unqualified devices are mixed with the good ones. Mass production makes it difficult for human workers to examine all of the products. Therefore, we would like to use modern machine learning methods, particularly deep learning, to help Nexperia pick out as many defect devices as possible while preserving the good ones, thus improving their yield rate.

Nexperia provided a dataset for Kaggle in-class contest that aims to classify images of semiconductor devices into two main classes, good and defect. For example, Fig. 1 shows a good example and a bad example. The Nexperia image dataset in the Kaggle contest contain 34457 train images (27420 good and 7039 bad) and 3830 test images with similar good-to-bad ratio. The key is to detect as many defect images as possible while not sacrificing too many passed ones. So on Kaggle contest, we adopt Area-Under-the-Curve (AUC) for ROC as the evaluation rule. Note that AUC values are in the range of $[0.5, 1]$, the higher, the better.

We note that this real world dataset may contain noisy labels, especially the images labeled as "good" possibly being "bad" ones in fact. We do not have ground truth on which labels are wrong, but you may pay additional attention to this issue.

Checking the following Kaggle website for more details.

- `https://www.kaggle.com/c/semi-conductor-image-classification-second-stage`

To participate the contest, you need to login your Kaggle account first, then open the following invitation link and accept the Kaggle contest rule to download the data:

`https://www.kaggle.com/t/5cbb376414c24ba5a9a9183ac73d648f`

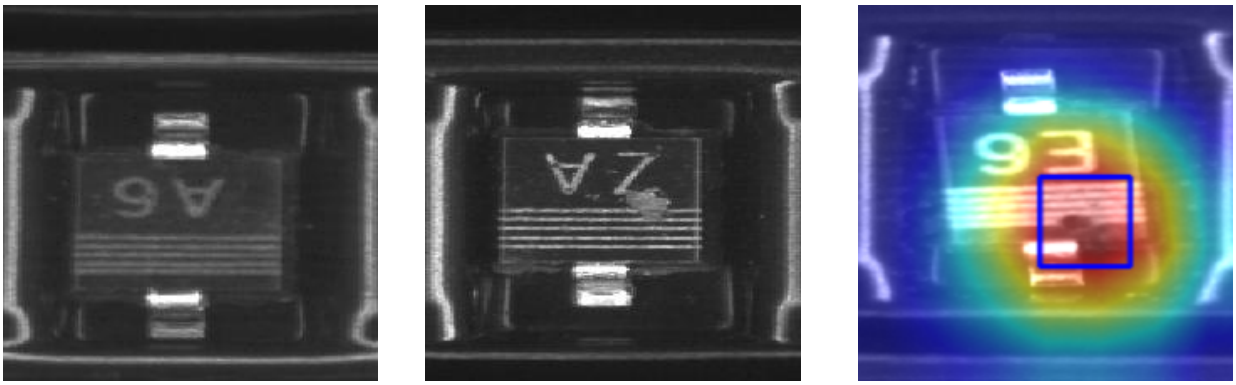

Figure 1: Examples of Nexperia image data. Left: a good example; Middle: a bad example; Right: a visualization based on heatmap

Requirements. For Kaggle contests, please register your team with name in the format of

math4995_lastname, so that we could easily find your results on Kaggle. For example, a team with Shawn Zhu and Kate Wong would be named by math4995_Zhu_Wong.

3 Kaggle Contest: Predict Survival on the Titanic

The sinking of the Titanic is one of the most infamous shipwrecks in history.

On April 15, 1912, during her maiden voyage, the widely considered "unsinkable" RMS Titanic sank after colliding with an iceberg. Unfortunately, there were not enough lifeboats for everyone onboard, resulting in the death of 1502 out of 2224 passengers and crew.

While there was some element of luck involved in surviving, it seems some groups of people were more likely to survive than others.

In this challenge, we ask you to build a predictive model that answers the question: "what sorts of people were more likely to survive?" using passenger data (*i.e.* name, age, gender, socio-economic class, etc). Visit the following website to join the Kaggle contest:

`https://www.kaggle.com/c/titanic`

Requirements. For Kaggle contests, please register your team with name in the format of math4995_lastname, so that we could easily find your results on Kaggle. For example, a team with Shawn Zhu and Kate Wong would be named by math4995_Zhu_Wong.

4 Kaggle Contest: Home Credit Default Risk

Many people struggle to get loans due to insufficient or non-existent credit histories. And, unfortunately, this population is often taken advantage of by untrustworthy lenders.

Home Credit strives to broaden financial inclusion for the unbanked population by providing a positive and safe borrowing experience. In order to make sure this underserved population has a positive loan experience, Home Credit makes use of a variety of alternative data–including telco and transactional information–to predict their clients' repayment abilities.

While Home Credit is currently using various statistical and machine learning methods to make these predictions, they're challenging Kagglers to help them unlock the full potential of their data. Doing so will ensure that clients capable of repayment are not rejected and that loans are given with a principal, maturity, and repayment calendar that will empower their clients to be successful.

Visit the following website to join the competition.

`https://www.kaggle.com/c/home-credit-default-risk/`

Requirements. For Kaggle contests, please register your team with name in the format of math4995_lastname, so that we could easily find your results on Kaggle. For example, a team with Shawn Zhu and Kate Wong would be named by math4995_Zhu_Wong.