Math 4432: Statistical Machine Learning **May 8, 2018**

Project 3: Final

Instructor: Yuan Yao *Due: May 22 11:59pm, 2018*

Requirement

1. Pick up ONE (or more if you like) favorite challenges *below*. If you would like to work on a different problem outside the candidates we proposed, please email course instructors about your proposal. Brave hearts for explorations will be encouraged!

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 Python (Jupyter) Notebooks with a detailed documentation, a *poster* such as

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

   or a *technical report within 8 pages*, e.g. NIPS conference style

   `https://nips.cc/Conferences/2016/PaperInformation/StyleFiles`

3. In the report, show your proposed scientific questions to explore and main results with a careful analysis supporting the results toward answering your problems. Remember: scientific analysis and reasoning are more important than merely the performance tables. Separate source codes may be submitted through email as a .zip file, GitHub link, or as an appendix if it is not large. There is no restriction on the programming languages to use, but R or Python are recommended.

4. Submit your report by email or paper version no later than the deadline, to the following address (statml.hw@gmail.com) with Title: Math4432: Project 3. Late submissions may consume grades.

1 Nexperia Predictive Maintenance Contest

Refer to the introduction by Mr. Gijs Bruining:

`https://github.com/yuany-pku/2018_math4432/blob/master/slides/Nexperia.pdf`

Kaggle in-class contests are at the following website:

`https://www.kaggle.com/c/nexperia-predictive-maintenance`

where three contests are available, depending the data to use for predictions. The number of days before the predictive dates is called the *observation window* (OW) and the number of predictive dates is called the *predictive window*. As a warm-up, the mini-contest is to exploit 2 days observation window (OW=2) with prediction window (PW=1). Two additional full-contests exploits different combinations of OW (=1,2,4,8,16, all in as features) and PW (1 or 2 for two full-contests, respectively).

2 Transfer Learning

You are required to do the transfer learning on (at least) one dataset below. The following procedures is for your reference.

- Feature extraction by pre-trained deep neural networks, e.g. VGG19, and resnet18, etc.;
- Visualize these features using classical unsupervised learning methods, e.g. PCA, clustering, etc.;
- Image classifications using traditional supervised learning methods based on the features extracted, e.g. LDA, logistic regression, SVM, random forests, etc.;
- (Optional) Train the last layer or fine-tune the deep neural networks in your choice, that may need GPUs to speed up;
- Compare the results you obtained and give your own analysis on explaining the phenomena.

Below are some candidate datasets.

2.1 MNIST dataset – a Warmup

Yann LeCun's website contains original MNIST dataset of 60,000 training images and 10,000 test images.

`http://yann.lecun.com/exdb/mnist/`

There are various ways to download and parse MNIST files. For example, Python users may refer to the following website:

`https://github.com/datapythonista/mnist`

or MXNET tutorial on mnist

`https://mxnet.incubator.apache.org/tutorials/python/mnist.html`

2.2 Fashion-MNIST dataset

Zalando's Fashion-MNIST dataset of 60,000 training images and 10,000 test images, of size 28-by-28 in grayscale.

`https://github.com/zalandoresearch/fashion-mnist`

2.3 Identification of Raphael's paintings from the forgeries

The following data, provided by Prof. Yang WANG from HKUST,

`https://drive.google.com/folderview?id=0B-yDtwSjhaSCZ2FqN3AxQ3NJNTA&usp=sharing`

contains a 28 digital paintings of Raphael or forgeries. Note that there are both jpeg and tiff files, so be careful with the bit depth in digitization. The following file

`https://docs.google.com/document/d/1tMaaSIrYwNFZZ2cEJdx1DfFscIfERd5Dp2U7K1ekjTI/edit`

contains the labels of such paintings, which are

1 Maybe Raphael - Disputed

2 Raphael

3 Raphael

4 Raphael

5 Raphael

6 Raphael

7 Maybe Raphael - Disputed

8 Raphael

9 Raphael

10 Maybe Raphael - Disputed

11 Not Raphael

12 Not Raphael

13 Not Raphael

14 Not Raphael

15 Not Raphael

16 Not Raphael

17 Not Raphael

18 Not Raphael

19 Not Raphael

20 My Drawing (Raphael?)

21 Raphael

22 Raphael

23 Maybe Raphael - Disputed

24 Raphael

25 Maybe Raphael - Disputed

26 Maybe Raphael - Disputed

27 Raphael

28 Raphael

There are some pictures whose names are ended with alphabet like A's, which are irrelevant for the project.

The challenge of Raphael dataset is: can you exploit the known Raphael vs. Not Raphael data to predict the identity of those 6 disputed paintings (maybe Raphael)? Textures in these drawings may disclose the behaviour movements of artist in his work.

One preliminary study in this project can be: take all the known Raphael and Non-Raphael drawings and use leave-one-out test to predict the identity of the left out image; you may break the images into many small patches and use the known identity as its class.

Remind: You should be really careful when reporting test error evaluation. For example, You cannot directly tune parameters (shallow learning or fine tuning) to make your leave one out error least and report it as the test error estimation. In this problem, it's easy to find some hyperparameters to overfit due to the small size of data (even if you augment training dataset, batch effect make the augmented crop images from the same paint similar in feature space, then finding such hyperparamters to overfit is basically as easy as before).

The following student poster report seems a good exploration

`https://github.com/yuany-pku/2017_CSIC5011/blob/master/project3/05.GuHuangSun_poster.pdf`

The following paper by Haixia Liu, Raymond Chan, and me studies Van Gogh's paintings which might be a reference for you:

`http://dx.doi.org/10.1016/j.acha.2015.11.005`

3 From Project 2: Kaggle contest classification: Predict survival on the Titanic

The following website contains the Kaggle contest on predicting survival (binary classification) on the Titanic:

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

Register the Kaggle and join the contest by submitting your predictions. Report your methods and the corresponding scores (accuracy) on the leaderboard (your registered name and ranking results).

4 From Project 2: Kaggle contest regression: Predict house sales prices

The following website contains a Kaggle contest on predicting house sales prices (regression) using the Ames Housing dataset:

`https://www.kaggle.com/c/house-prices-advanced-regression-techniques/`

It is aimed for practicing feature engineering, RFs, and gradient boosting etc. Register the Kaggle and join the contest by submitting your predictions. Report your methods and the corresponding scores (RMSE) on the leaderboard (your registered name and ranking results).