**IEE 598: Data Science for System Informatics** **Spring 2018**

# Assignment 3 — April 9

*Instructor: Hao Yan* *Scribe: Arun Bala Subramaniyan*

**Submission Guideline** For submission of the homework, please submit the report and the code. The code should be in either the notebook file (.ipynb) or python file (.py) if you prefer to directly write the python outside the notebook. Even though you provide the notebook, it is required to put the necessary plots/tables inside the report to answer the homework questions. Also, please make sure your code can run through before the submission.

## Problem 1: Kernel regression

Kernel method can be a powerful tool for spatial dataset or images. For this example, we would like to explore the use of Kernel Ridge Regression for estimating a 2D surface.

The true function is given in the Jupyter notebook as $z = x\exp(-x^2 - y^2)$. The goal is to use $(x, y)$ to predict $z$.

1. Please start with the linear regression method and report the training and testing accuracy. Please visualize the final prediction in an image.

2. Please use the Kernel Ridge Regression and report the training and testing accuracy. Please visualize the final prediction in an image. Please use cross-validation to select the best tuning parameters.

3. (Bonus 10 points) Please use the Random Forest Regressor and report the training and testing accuracy. Please visualize the final prediction in an image. Please use cross-validation to select the best tuning parameters. From the three models that you have tried, which one works the best and why?

4. (Bonus 20 points) For Ridge regression problem, the primal problem is given as:

$$\min_w l(w) = \frac{1}{2}\|y - Xw\|^2 + \frac{\lambda}{2}\|w\|_2^2$$

   The dual of the problem is given as:

$$\max_\alpha -\lambda^2\|\alpha\|^2 + 2\lambda\alpha^T y - \lambda\alpha^T XX^T\alpha$$

   Please solve the analytical solution and gradient descent algorithm for the dual problem and compare its complexity against the primal problem in the following table. When would you prefer to solve the problem in the primal or dual?

| | Primal | | Dual | |
|---|---|---|---|---|
| | Time | Space | Time | Space |
| Gradient Descent | $O(np)$ | $O(np)$ | | |
| Analytical Solution | $O(np^2 + p^3)$ | $O(np + p^2)$ | | |

# Problem 2: Click Through Rate Prediction

The following problem is to predict the click through rate. The dataset description is given as follow:

- Data fields
  - id: ad identifier
  - click: 0/1 for non-click/click
  - hour: format is YYMMDDHH, so 14091123 means 23:00 on Sept. 11, 2014 UTC.
  - C1 -- anonymized categorical variable
  - banner_pos
  - site_id
  - site_domain
  - site_category
  - app_id
  - app_domain
  - app_category
  - device_id
  - device_ip
  - device_model
  - device_type
  - device_conn_type
  - C14-C21 -- anonymized categorical variables

1. Let's Start with SVM. Please use svm.LinearSVC, Let's try to add balanced weight to handle the class-imbalance issue.
   (a) please compute the precision/recall, f1-score, and confusion matrix.
   (b) Please run the algorithm for multiple times and observe the result.
2. Regularized SVM

   (a) Let's try to add penalty, please exlpore the use of the 'l1' and'l2' penalty in Scikit-learn, Please also use cross validation to select the best tuning parameters C.

   (b) please compute the precision/recall, f1-score, and confusion matrix for 'l1' and 'l2' model with the best tuning parameter C.

3. Please also explore using Logistic Regression on this problem and report the result.

   (a) Please plot the ROC curve and compute the area under the ROC curve. (You don't need to explore the use of penalty since the cross validation can be very slow)

   (b) Please plot the precision recall curve and compute the average precision

   (c) Please compute the F1-score and confusion matrix.

4. Please also explore using Random Forest on this problem and report the result.

   (a) Please use cross-validation to select the best tuning parameters

   (b) Please plot the ROC curve and compute the area under the ROC curve.

   (c) Please plot the precision recall curve and compute the average precision

   (d) Please compute the F1-score and confusion matrix

5. (Bonus 15 points) Please try to implement the xgboost library to this dataset.

   (a) Please use cross-validation to select the best tuning parameters

   (b) Please plot the ROC curve and compute the area under the ROC curve.

   (c) Please plot the precision recall curve and compute the average precision

   (d) Please compute the F1-score and confusion matrix