IE613: Online Machine Learning **Jan-Apr 2018**

Assignment 2: April 1

Instructions: *You are free to code in Python/Matlab/C/R. Discussion among the class participants is highly encouraged. But please make sure that you understand the algorithms and write your own code. If you share any code with any other student then you will be penalized and can be given 0 mark for that question.*

Submit the code and report by 11:59PM, 1^{st} April on Moodle. Late submission will not be evaluated and given 0 mark.

Question 1 (FoReL and WM Equivalence) *Proof that Follow-the-Regularized-Leader (FoReL) with linear loss functions and negative entropy as the regularizer (i.e.* $R_w = \frac{1}{\eta}\sum_{j=1}^{d} w_j \log w_j$*) is equivalent to Weighted Majority algorithm. Let, in round* $t = 1, 2, \cdots, T$*, the linear function is* $f_t(w) = \langle w, v_t \rangle = \sum_{j=1} w_j v_{j,t}$ *where* $\sum_{j=1}^{d} w_j = 1$ *and* $v_t \in \{0,1\}^d$ *is loss vector generated in round* t*. What is the optimal* $\eta^\star$*?*

Hint: *Let* w_j^f *is the* j^{th} *component of* w *in FoReL and* w_j^{wm} *is the* j^{th} *component of* w *in WM algorithm. For equivalence, show that* $w_j^f = w_j^{wm}$ *and then find* $\eta^\star$ *by using this equivalence.*

Question 2 (Online Convex Optimization) *Consider the problem of prediction with expert advice with* $d = 10$*. Assume that the losses assigned to each expert are generated according to independent Bernoulli distributions. The adversary/environment generates loss for experts* 1 *to* 8 *according to* $Ber(.5)$ *in each round. For the* 9*th expert, loss is generated according to* $Ber(.5 - \Delta)$ *in each round. The losses for the* 10*th expert are generated according to different Bernoulli random variable in each round for the first* $T/2$ *rounds, they are generated according to* $Ber(0.5 + \Delta)$ *and the remaining* $T/2$ *rounds they are generated according to Bernoulli random variable* $Ber(0.5 - 2\Delta)$*.* $\Delta = 0.1$ *and* $T = 10^5$*.*

Generate plots for (pseudo) cumulative regret vs T *for each of the following algorithms. The averages should be taken over at-least 20 sample paths (more is better). Display 95 confidence intervals for each plot.*

- *Follow-The-Leader (FTL)*
- *Follow-the-Regularized-Leader (FoReL) with optimal value be* η^* *from* ***Question 1***

Question 3 *Assume that the losses are generated as in* ***Question 2****. Generate (pseudo) regret values for Follow-the-Regularized-Leader (FoReL) with optimal value be* η^* *from* ***Question 1*** *for different learning rates* ($\eta = c\eta^*$)*. Vary* c *in the interval* $[0.1\ 2.1]$ *in steps of size* 0.2 *to get different learning rates. The averages should be taken over at-least* 20 *sample paths (more is better) and also display* 95% *confidence intervals for plot.*

Question 4 (Online Classifier) *Train a classifier using winnow* ($\eta = 0.25$) *and perceptron algorithm for following datasets:*

1. *Skin Segmentation Data Set*

2. *Online News Popularity Data Set*
3. *Wine Quality Data Set*

Plot $\| \frac{w_{t+1}}{\|w_{t+1}\|_2} - \frac{w_t}{\|w_t\|_2} \|_2$ *vs number of data points for each datasets. For perceptron algorithm, start plotting after smallest* t *for which* $\| w_t \|_2 \neq 0$.

Note: *Please visit above links for more information about datasets. The processed datasets for this assignment is available on* **course web-page**.

Question 5 (Online-to-batch Conversions) *Consider a PAC learning problem for binary classification parameterized by an instance domain,* $\mathcal{X}$ *, and a hypothesis class,* H*. Suppose that there exists an online learning algorithm,* A*, which enjoys a mistake bound* $M_A(H) < \infty$*. Consider running this algorithm on a sequence of* T *examples which are sampled i.i.d. from a distribution* $\mathcal{D}$ *over the instance space* $\mathcal{X}$*, and are labeled by some* $h \in \mathcal{H}$*. Suppose that for every round* t*, the prediction of the algorithm is based on a hypothesis* $h_t : \mathcal{X} \to \{0, 1\}$*. Show that*

$$\mathbb{E}[L_{\mathcal{D}}(h_r)] \leq \frac{M_A(H)}{T}$$

where the expectation is over the random choice of the instances as well as a random choice of r *according to the uniform distribution over* $[T]$.

Question 6 *Construct a linearly separable dataset in 2 dimensions as follows: sample 1000 points from a normal distribution with mean* $[-2 \ -2]^\top$ *and variance 0.5I, where I is the* 2×2 *identity matrix; similarly sample from a different normal distribution of mean* $[-10 \ -10]^\top$ *with variance 0.25I. Label the points from* $N([-2, -2]^\top, 0.5I)$ *as -1 and the remaining points as +1. Compute an estimate of the margin.*

- *Run the perceptron algorithm and* ***report*** *the number of times the perceptron algorithm makes mistakes. Verify the mistake bound.*
- *Run the winnow algorithm on the above dataset and use at-least 10 different value of the learning rate (*$\eta <= 0.5$*). Verify the mistake bound and* ***report*** *the numbers of mistakes for given learning rate.*

Question 7 *Using the dataset construction procedure illustrated in the previous question, construct another linearly separable dataset of 100000 points and 500 dimensions (sample 50000 points from a 500-dimensional normal distribution with mean* $[-2 - 2 \ldots - 2]^\top$ *and variance 0.5I, where I is an identity matrix of size* 500×500 *and sample 50000 points from another 500 dimensional normal distribution with mean* $[-10 - 10 \ldots - 10]^\top$ *and variance 0.25I). Use the cost function* $c_t = \max\{0, 1 - y_t w^\top x_t\}$ *and*

- *Run the online gradient descent algorithm.*
- *Use the regularizer* $R(w) = w^\top \log(w) = \sum_{j=1}^{500} w_j \log w_j$ *and run the online mirror descent algorithm.*

Plot the regret vs round t *for both algorithms in one figure. Compare and contrast the online gradient descent regret and online mirror descent regret.*

Submission Format and Evaluation: Your should submit a report along with your code. Please zip all your files and upload via moodle. The zipped folder should named as YourRegistrationNo.zip e.g. '154290002.zip'. The report should contain six figures: one figure should have two plots corresponding to each algorithm in Q.2, one figure for Q.3, one figure for Q.7 and the other figures should have 2 plots one corresponding to each algorithm in Q.4 for each dataset. For each figure, write a brief summary of your observations. We may also call you to a face-to-face session to explain your code.