ECE 5403: Machine Learning **(Due: 04/30/2020)**

Submission Assignment #3

Instructor: Chun-Shu Wei *Name:* Student name, *Student Id:* Student Id

Course Policy: Read all the instructions below carefully before you start working on the assignment, and before you make a submission. For this assignment, please hand in the following two things: a pdf file and a ipynb file.

- PDF file: contains both your results and explanations. Please name this pdf file as **HW3_StudentID_Name.pdf** and **remember to type your Student ID and Name in pdf (e.g. HW3_9400000_chunshuwei)**.
- Ipynb file: write the comment to explain your code. Please name this ipynb file as **HW3_StudentID_Name.ipynb**
- Please name your assignment as **HW3_StudentID_Name.zip**. The archive file contains source code(ipynb file) and report (pdf file).
- Implementation will be graded by completeness, algorithm correctness, model description, and discussion.
- PLAGIARISM IS STRICTLY PROHIBITED.
- Please submit your assignment as ONE single zip file on the E3 system. Paper submission is not allowed. Inserting clear scanned image of handwritten derivations is accepted. Denote date and time on the first page.
- Submission deadline: **2020.04.30 11:55:00 PM**.

Problem 1: Cross Entropy, KL divergence and Logistic Regression (15 points)

Let the true probability $p(x)$ be the true label, and the given distribution $q(x)$ be the predicted value of the current model. Now we apply logistic regression to classify observations into two classes (0 and 1). The output of the model for a given observation, given a vector of input features x, can be interpreted as a probability, which serves as the basis for classifying the observation. The probability is modeled using the logistic function $g(z) = \frac{1}{1+e^{-z}}$ where $z = w \cdot x$ is a linear function of x. The probability of the output $y = 1$ is given by:

$$q_{y=1} = \hat{y} = g(w \cdot x) = \frac{1}{1 + e^{-w \cdot x}} \tag{0.1}$$

where the vector of weights w is optimized through some appropriate algorithm such as gradient descent. Similarly, the complementary probability of finding the output $y = 0$ is simply given by

$$q_{y=0} = 1 - \hat{y} \tag{0.2}$$

Having set up our notation $p \in \{y, 1 - y\}$ and $q \in \{\hat{y}, 1 - \hat{y}\}$, we can use cross entropy to get a measure of dissimilarity between p and q:

$$H(p(x), q(x)) = -E_p\{\log q(x)\} = \sum_x p(x) \log \frac{1}{q(x)} = -y \log \hat{y} - (1 - y) \log 1 - \hat{y} \tag{0.3}$$

That is, cross entropy can be rewriten as loss of logistic regression. The definition may be formulated using the Kullback–Leibler divergence (KL divergence) $D_{KL}(p||q)$ where KL divergence is defined as:

$$D_{KL}(p(x) \| q(x)) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Please use above definition to prove that:

$$H(p(x), q(x)) = D_{KL}(p(x) \| q(x)) + H(p(x))$$

Problem 2: Python code Exercise (10+10+10+10+15+10=65 points)

Please use make_moon function in HW3_sample_code.ipynb to generate raw data and remember to standarize raw data. Notice that you need to show decision regions, number of misclassified samples and accuracy of in the following sub-problems.

(a) KNN classification
Please implement a KNN classifier in scikit-learn using a Euclidean distance metric where $K = 11$.

(b) SVM classifier
Please implement a SVM classifier in scikit-learn using 'rbf' kernel where $random\_state = 0$, $\gamma = 0.2$, and $C = 10.0$.

(c) SVM classifier
Please implement a SVM classifier in scikit-learn using 'linear' kernel where $C = 1000.0$ and $random\_state = 0$.

(d) SVM classifier
Please implement a SVM classifier in scikit-learn using 'sigmoid' kernel.

(e) Given $C \in \{0.1, 1.0, 10.0, 100.0, 1000.0, 10000.0\}$ and $\gamma \in \{0.00001, 0.0001, 0.001, 0.01, 0.1, 1.0\}$, please find the best combination of (C, γ) for default SVM classifier with $random\_state = 0$ in scikit-learn.

(f) Logistic Regression
Please implement Logistic Regression in scikit-learn where $C = 1000.0$, $random\_state = 0$, and $solver =$ "$liblinear$".

Problem 3: Loss functions comparison (5+5+5+5=20 points points)

Please write down the pros and cons of the following loss functions:

(a) Mean Squared Error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n} ||y_i - \hat{y}_i||_2^2$$

(b) Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n} ||y_i - \hat{y}_i||$$

(c) Hinge Loss for SVM

$$l(y) = \max\{0, 1 + \max_{y \neq t}\{w_y x - w_t x\}\}$$

Where t the target label, w_t and w_y the model parameters.
Note: In sub-problem(c), just explain the meaning of Hinge Loss by its definition.

(d) Cross Entropy Loss

$$\sum_{i=1}^{n} H(p(x_i), q(x_i)) = -\sum_{i=1}^{n} y_i \log \hat{y}_i + (1 - y_i) \log 1 - \hat{y}_i$$

Problem 4: Kernel Ridge Regression and Soft Margin SVM (bonus) (10 points)

Please explain the similarities of Soft Margin SVM and Kernel Ridge Regression.