THE UNIVERSITY OF HONGKONG
DEPARTMENT OF PHYSICS

Exercise sheet #3

Course: *Machine Learning in Physics (PHYS3151)* – Professor: *Dr. Ziyang Meng*
Due date: *Mar.17th, 2022*

1. **Properties of logistic function** $g(z)$

(a) We have the logistic function $g(z) = \frac{1}{1+e^{-z}}$. Please prove by hand that
(1) $g(z) + g(-z) = 1$
(2) $g'(-z) = g'(z)$
(3) $g'(z) = g(z)(1 - g(z)) = g(z)g(-z)$

(b) We can add a parameter T to the logistic function, then it becomes

$$g(z) = \frac{1}{1 + e^{-z/T}}$$

Draw $g(z)$ at $T = 10, 5, 2, 1, 0.5, 0.1$ on one graph, label them separately, and discuss the shape of $g(z)$ when $T \to 0$.

2. **Logistic Regression**

In high energy physics experiments, such as the ATLAS (A Toroidal LHC ApparatuS) and CMS (Compact Muon Solenoid) detectors at the CERN (The European Organization for Nuclear Research, Conseil européen pour la recherche nucléaire) LHC (Large Hadron Collider), one major hope is the discovery of new particles. To accomplish this task, physicists attempt to sift through data events and classify them as either a **signal** of some new physics process or particle, or instead a **background** event from understood Standard Model processes. Unfortunately we will never know for sure what underlying physical process happened. However, we can attempt to define parts of phase space that will have a high percentage of signal events.

Here we have a simulated sample of "signal" events and a simulated sample of "background" events. Please follow the example 3 of *logistic-regression.ipynb*, train a logistic regression based on two variables to determine whether an unknown event is "signal" or "background", and plot the decision boundary.

3. **Support vector machine**

(a) Given a straight line $(2x + 3y = 1)$ in the 2D space, what is the normal vector of this line, what is the distance of the line to the origin $(0, 0)$, what is the distance of an arbitrary point (x_0, y_0) to the line ?

(b) Please read the example 4 of *support-vector-machine.ipynb* , check the codes step by step to see if there is anything you can improve (like adjusting the learning rate

or regularization parameter), and verify that the margin could really separate the two classes of data.

(c) Please find *SVMIris.ipynb* and Iris.csv, we have discussed using two features *SepalLengthCm* and *SepalWidthCm* to separate the Iris-setosa with Iris-versicolor. Now please use the other two features *PetalLengthCm* and *PetalWidthCm* to separate the two species Iris-setosa with Iris-virginica using SVM and logistic regression separately. For SVM please draw the hyperplane and margins; for logistic regression please draw the decision boundary.