CS436/536: Introduction to Machine Learning
Homework 3
(Due Sunday 3/3 **before 11.59pm**). No late days.

Instructions: To solve these problems, you are allowed to consult your classmates, as well as the class textbook (*Learning from Data* by Abu-Mostafa, Magdon-Ismail, and Lin, which we will call LFD) and the slides posted on Brightspace, but no other sources. You are encouraged to collaborate with other students, while respecting the collaboration policy (please see the module on Academic Honesty on Brightspace). Please write the names of all the other students you collaborated with on the homework. Everyone must write up their assignments separately.

Please write clearly and concisely, and use rigorous, formal arguments. Homework is due at the beginning of lecture, and homework turned in later will be considered late and will use up one of your late days. You must use Brightspace to submit the homework as a single neatly typed pdf file. Hand-drawn formulas or figures are okay and may be included as images within the pdf. If a programming assignment calls for plotting the results, axes must be clearly labeled, and its meaning must be obvious to anyone with only a rudimentary knowledge of machine learning and computer science. Emailed copies will not be accepted.

(1) [200 points] Gradient Descent on a Simple Function.
Consider the function $f(x, y) = 2x^2 + y^2 + 3\sin(2\pi x)\cos(2\pi y)$.

(a) Implement gradient descent to minimize this function. Run gradient descent starting from the point $(x = 0.1, y = 0.1)$. Set the learning rate to $\eta = 0.01$ and the number of iterations to 50. Give a plot that displays how the function value drops through successive iterations of gradient descent. Repeat this with a learning rate of $\eta = 0.1$ and provide a plot of the function value with each iteration. What do you observe?

(b) Obtain the "minimum" value and location of the minimum value of the function you get using gradient descent with the same learning rate $\eta = 0.01$ and number of iterations (50) as part (a), from the following starting points: (i) $(0.1, 0.1)$, (ii) $(1, 1)$, (iii) $(0.5, 0.5)$, (iv) $(0.0, 0.5)$, (v) $(-0.5, -0.5)$, (vi) $(-1, 1)$. Write down the minimum value obtained using gradient descent and the location of the minimum value for each of these starting points. As you may appreciate, finding the "true" *global* minimum value of an arbitrary function is a hard problem.

(2) [600 points] Classifying Handwritten Digits: 1 vs. 5.

Implement logistic regression for classification using gradient descent. Find the best separator *using the training data only* (using `ZipDigits.train`). Use the data and features you generated for solving HW2. For each example, use the two features you computed in HW2 to as the inputs; and the output is $+1$ if the handwritten digit is 1 and -1 if the handwritten digit is 5. Once you have found a separator using your classification algorithm:

(a) Give separate plots of the training data (`ZipDigits.train`) and test data (`ZipDigits.test`) which display the data points using the two features you computed in HW2, together with the separator.

(b) Compute E_{in} on your training data (`ZipDigits.train`) and E_{test}, the error of your separator on the test data (`ZipDigits.test`).

(c) Obtain a bound on the true out-of-sample error (E_{out}). You should get two bounds, one based on E_{in} and another based on E_{test}. Use a tolerance of $\delta = 0.05$. *Which is the better bound?*

(d) Repeat parts (a)-(c) using a 3-rd order polynomial transform.

(e) Which model would you deliver to the USPS, the linear model with the 3-rd order polynomial transform or the one without? *Explain.*