IDS 575: Assignment 3

- Turn in solutions as a single notebook (ipynb) *and* as a pdf on Blackboard. No need to turn in datasets/word-docs.
- Answer the following questions concisely, in complete sentences and with full clarity. Across group collaboration is *strictly* not allowed. Always cite all your sources.
- Make appropriate assumptions necessary in order to answer the questions, and mention them explicitly at the beginning of each answer.

1. Probabilistic Perspective of Linear Regression (1pt)

In addition to the least square objective, we also learned the probabilistic perspective for linear regression, where each observation is assumed to have a Gaussian noise. (i.e., Noise of each example is an independent and identically distributed sample from a normal distribution). Lets assume that we are given the following regression model that includes two linear features and one quadratic feature.

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2)$$

Your goal is to relate the likelihood maximization objective to the least squares objective.

1. Is y linear with respect to θ? Is y linear with respect to $\{x_1, x_2\}$?
2. Given the definition of noise, derive the corresponding mean and variance parameters of the normal distribution for $y|x_1, x_2; \theta$. Also write down its probability density function.
3. You are provided with a training observations $D = \{x_1^{(i)}, x_2^{(i)}, y^{(i)}\}$ where $i = 1, \ldots, m$. Derive the conditional log-likelihood that will be later maximized to make the dataset D most likley.
4. If you omit all the constants that do not relate to our parameters θ, what will the objective function $J(\{\theta_0, \theta_1, \theta_2, \theta_3\})$ (to perform Maximum Likelihood Estimation) look like? Does J look similar to the Least Square objective for this problem? What are the differences, if any?

2. Logistic Regression Toy Problem (1pt)

Recall the grading problem to predict pass/fail in the class. Suppose you collect data for a group of students in the class that consist of two input features X_1 = hours studied and X_2 = undergrad GPA. Your goal is to predict the output $Y \in \{pass, fail\}$. Suppose that you fit a logistic regression, learning its parameter $(\theta_0, \theta_1, \theta_2) = (-6, 0.05, 1)$.

1. Define a python function, which given inputs X_1, X_2, θ will return the probability of passing the class.
2. What will be the probability of passing the class for a student who studies for 40 hours and has a GPA of 3.5?
3. How many hours would the aforementioned student need to study in order to have at least 50% chance of passing the class?

3. Logistic Regression (6pt)

The following questions must be answered using the Weekly dataset (originally found in the ISLR package in R). This dataset contains 1,089 weekly returns for 21 years from the beginning of 1990 to the end of 2010. You will use its 1990-2008 as a training data and 2009-2010 as a test data to perform logistic regression.

1. In the data, the input features are five of *Lag* variables and *Volume*, and the binary output is *Direction*. Read the csv into a pandas dataframe and describe the relevant features and target variables of the dataset.

2. We plan to reuse the SGD procedure from Assignment 2 to fit a logistic regression model. Do we need to write a new gradient function tailored to binary logistic regression, or can we use the one written in Assignment 2?
3. Extend the gradient function such that a regularized logistic loss is being minimized. In particular, consider an ℓ_2 penalty on the parameters (except the intercept). For this, the gradient function will need: (a) a penalty input that can take two values: "none" and "l2", and (b) an optional regularization coefficient C input. Summarize the changes from the previously written gradient function that were necessary to accomplish this.
4. Write a python function that takes the predicted labels and the true label arrays, and computes the confusion matrix.
5. Use the above function to report the confusion matrix and the accuracy on both training and test data for two models: