ISyE 6416: Computational Statistics
Homework 2
(20 points for each question. 100 points total.)

- Please write your team member's name is you collaborate.

1. **Medical image estimation.**

   Suppose x_i, $i = 1, \ldots, n$ are i.i.d. Poisson with

   $$P(x_i = k) = \frac{e^{-\mu_i}\mu_i^k}{k!}$$

   with unknown mean μ_i. The variables x_i represent the number of times that one of n possible independent events occurs during a certain period. In emission tomography, they may represent the number of photons emitted by n sources.

   We consider an experiment designed to determine the means μ_i. The experiment involves m detectors. If event i occurs, it is detected by detector j with probability p_{ji}. We assume the probabilities p_{ji} are given (with $p_{ji} > 0$ and $\sum_{j=1}^m p_{ji} \leq 1$. The total number of events recorded by detector j is denoted by y_j,

   $$y_j = \sum_{i=1}^n y_{ji}, \quad j = 1, \ldots, m.$$

   Formulate the maximum likelihood estimation problem of estimating the means μ_i, based on observed values of y_j, $j = 1, \ldots, m$. Will the maximum likelihood function returns a unique maximizer? (Hint: the variables y_{ji} have Poisson distribution with means $p_{ji}\mu_i$. The sum of n independent Poisson variables with means $\lambda_1, \ldots, \lambda_n$ has a Poisson distribution with mean $\lambda_1 + \cdots + \lambda_n$.

2. **Logistic regression.**

   Given n observations (x_i, y_i), $i = 1, \ldots, n$, $x_i \in \mathbb{R}^p$, $y_i \in \{0, 1\}$, parameters $a \in \mathbb{R}^p$ and $b \in \mathbb{R}$. Consider the log-likelihood function for logistic regression:

   $$\ell(a, b) = \sum_{i=1}^n \{y_i \log h(x_i; a, b) + (1 - y_i) \log(1 - h(x_i; a, b))\}$$

   (a) Derive the Hessian H of this function and show that H is negative semi-definite (this implies that ℓ is concave and has no local maxima other than the global one.)

   (b) Use data logit-x.dat and logit-y.dat, which contain the predictors $x_i \in \mathbb{R}^2$ and response $y_i \in \{0, 1\}$ respectively for logistic regression problem. Implement Newton's method for optimizing $\ell(a, b)$ and apply it to fit a logistic regression model to the data. Initialize Newton's method with $a = 0$, $b = 0$. Plot the value of the log likelihood function versus iterations. What are the coefficients a and b from your fit?

   (c) Find a value of step-size that gives you convergence, and another value of step-size (larger) where your algorithm diverges.

3. **Locally weighted linear regression.**

Consider a linear regression problem in which we want to weight different training examples differently. Specifically, suppose we want to minimize

$$J(\theta) = \frac{1}{2}\sum_{i=1}^{n} w_i(\theta^T x_i - y_i)^2.$$

In class, we have worked out what happens for the case where all the weights are the same. In this problem, we will generalize some of those ideas to the weighted setting, and also implement the locally weighted linear regression algorithm.

(a) Show that $J(\theta)$ can also be written as

$$J(\theta) = (X\theta - y)^T W(X\theta - y)$$

for an appropriate diagonal matrix W, matrix X and vector y. State clearly what these matrices and vectors are.

(b) Suppose we have samples (x_i, y_i), $i = 1, \ldots, n$ of n independent examples, but in which the y_i's were observed with different variances, and

$$p(y_i|x_i, \theta) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp(-\frac{(y_i - \theta^T x_i)^2}{2\sigma_i^2})$$

i.e. y_i has mean $\theta^T x_i$ and variance σ_i^2 (where σ_i^2 are fixed, known, constants). Show that finding the maximum likelihood estimate of θ reduces to solving a weighted linear regression problem. State clearly what the w_is are in terms of σ_i^2's.

(c) Use data rx.dat and ry.dat, which contain the predictors x_i and response y_i respectively for our problem. Implement gradient descent for (unweighted) linear regression that we derived in class on this dataset, and plot on the same figure the data and the straight line resulting from your fit. (Remember to include the intercept term.)

(d) Implement locally weighted linear regression on this dataset, using gradient descent, and plot on the same figure the data and the line resulting from your fit. Using the following weights

$$w_i = \exp(-x_i^2/(20)).$$

Plot the $J(\theta)$ versus iterations.

4. **Exponential family and Fisher information.** A PDF $f(x|\theta)$ of a random variable is called to be from an exponential family if we can write

$$f(x|\theta) = g(x)e^{\beta(\theta)+h(x)^\top \gamma(\theta)}$$

for some $g(x)$, $\beta(\theta)$, $h(x)$ and $\gamma(\theta)$.

(a) Show that Bernoulli, Binomial, Poisson, Exponential and Gaussian distributions all belong to exponential family. Here the PDF for them are given by

$$\text{Bernoulli: } f(x|p) = p^x(1-p)^{1-x}, \quad x = \{0,1\}$$

$$\text{Binomial: } f(x|n,p) = \binom{n}{x} p^x(1-p)^{n-x}, \quad x = \{0,1,\ldots,n\}$$

$$\text{Poisson: } f(x|\lambda) = e^{-\lambda}\lambda^x/x!, \quad x = \{0,1,\ldots\}$$

$$\text{Exponential: } f(x|\lambda) = e^{-\lambda x}\lambda, \quad x \geq 0$$

$$\text{Gaussian: } f(x|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^p|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)}, \quad x \in \mathbb{R}^p$$

(b) Find the Fisher information for Bernoulli distribution.

5. **House price dataset.**

The HOUSES dataset contains a collection of recent real estate listings in San Luis Obispo county and around it. The dataset is provided in RealEstate.csv.

The dataset contains the following fields:

- MLS: Multiple listing service number for the house (unique ID).
- Location: city/town where the house is located. Most locations are in San Luis Obispo county and northern Santa Barbara county (Santa Maria-Orcutt, Lompoc, Guadelupe, Los Alamos), but there some out of area locations as well.
- Price: the most recent listing price of the house (in dollars).
- Bedrooms: number of bedrooms.
- Bathrooms: number of bathrooms.
- Size: size of the house in square feet.
- Price/SQ.ft: price of the house per square foot.
- Status: type of sale. Thee types are represented in the dataset: Short Sale, Foreclosure and Regular.

Fit linear regression model to predict Price using remaining factors (except Status), for each of the three types of sales: Short Sale, Foreclosure and Regular, respectively.